

IntechOpen

IntechOpen Series
Artificial Intelligence, Volume 42

Transfer Learning

Unlocking the Power of Pretrained Models

Edited by Pier Luigi Mazzeo and Alessandro Bruno



Transfer Learning - Unlocking the Power of Pretrained Models

*Edited by Pier Luigi Mazzeo
and Alessandro Bruno*

Published in London, United Kingdom

Transfer Learning - Unlocking the Power of Pretrained Models

<http://dx.doi.org/10.5772/intechopen.1008615>

Edited by Pier Luigi Mazzeo and Alessandro Bruno

Contributors

Ahmed Siala, Betty Campos Segales, David He, Feten Slimeni, Heejin Jeong, Jaspreet Kaur, John Kelleher, Lenin Huayta Flores, Lenin Samir Huayta Poma, Lucio Ticona Carrizales, Muhammad Imran, Muhammad Waqas Nawaz, Olaoluwa Popoola, Qammer H. Abbasi, Raghavendran Ramakrishnan, Safanah Abbas, Steven Davy, Tehseen Ullah, Tijeni Delleji, Wanquan Zhang

© The Editor(s) and the Author(s) 2026

The rights of the editor(s) and the author(s) have been asserted in accordance with the Copyright, Designs and Patents Act 1988. All rights to the book as a whole are reserved by INTECHOPEN LIMITED. The book as a whole (compilation) cannot be reproduced, distributed or used for commercial or non-commercial purposes without INTECHOPEN LIMITED's written permission. Enquiries concerning the use of the book should be directed to INTECHOPEN LIMITED rights and permissions department (permissions@intechopen.com).

Violations are liable to prosecution under the governing Copyright Law.



Individual chapters of this publication are distributed under the terms of the Creative Commons Attribution 4.0 License which permits commercial use, distribution and reproduction of the individual chapters, provided the original author(s) and source publication are appropriately acknowledged. If so indicated, certain images may not be included under the Creative Commons license. In such cases users will need to obtain permission from the license holder to reproduce the material. More details and guidelines concerning content reuse and adaptation can be found at <http://www.intechopen.com/copyright-policy.html>.

Notice

Statements and opinions expressed in the chapters are those of the individual contributors and not necessarily those of the editors or publisher. No responsibility is accepted for the accuracy of information contained in the published chapters. The publisher assumes no responsibility for any damage or injury to persons or property arising out of the use of any materials, instructions, methods or ideas contained in the book.

First published in London, United Kingdom, 2026 by IntechOpen

IntechOpen is the global imprint of INTECHOPEN LIMITED, registered in England and Wales, registration number: 11086078, 167-169 Great Portland Street, London, W1W 5PF, United Kingdom

For EU product safety concerns: IN TECH d.o.o., Prolaz Marije Krucifikse Kozulić 3, 51000 Rijeka, Croatia, info@intechopen.com or visit our website at intechopen.com.

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library

Transfer Learning - Unlocking the Power of Pretrained Models

Edited by Pier Luigi Mazzeo and Alessandro Bruno

p. cm.

This title is part of the Artificial Intelligence Book Series, Volume 42

Topic: Machine Learning and Data Mining

Series Editor: Andries Engelbrecht

Topic Editor: Marco Antonio Aceves-Fernández

Print ISBN 978-1-83635-352-2

Online ISBN 978-1-83635-351-5

eBook (PDF) ISBN 978-1-83635-353-9

ISSN 2633-1403

If disposing of this product, please recycle the paper responsibly.

IntechOpen

intechopen.com

Built by scientists, for scientists



[Explore all IntechOpen books](#)

IntechOpen Book Series

Artificial Intelligence

Volume 42

Aims and Scope of the Series

Artificial Intelligence (AI) is a rapidly developing multidisciplinary research area that aims to solve increasingly complex problems. In today's highly integrated world, AI promises to become a robust and powerful means for obtaining solutions to previously unsolvable problems. This Series is intended for researchers and students alike interested in this fascinating field and its many applications.

Meet the Series Editor



Andries Engelbrecht received the Masters and Ph.D. degrees in Computer Science from the University of Stellenbosch, South Africa, in 1994 and 1999 respectively. He is currently appointed as the Voigt Chair in Data Science in the Department of Industrial Engineering, with a joint appointment as Professor in the Computer Science Division, Stellenbosch University. Prior to his appointment at Stellenbosch University, he has been at the University of Pretoria, Department of Computer Science (1998-2018), where he was appointed as South Africa Research Chair in Artificial Intelligence (2007-2018), the head of the Department of Computer Science (2008-2017), and Director of the Institute for Big Data and Data Science (2017-2018). In addition to a number of research articles, he has written two books, *Computational Intelligence: An Introduction* and *Fundamentals of Computational Swarm Intelligence*.

Meet the Volume Editors



Pier Luigi Mazzeo obtained an MSc in Computer Science from the University of Salento, Lecce, Italy, in 2001. Since then, he has been working on several research topics regarding artificial intelligence and computer vision. Dr. Mazzeo joined the Italian National Research Council (CNR) as a researcher in 2002. He is currently involved in projects related to algorithms for video object tracking, face detection and recognition, facial expression recognition, deep neural networks, and machine learning. He has authored and co-authored 150 publications, including more than 25 papers published in international journals and book chapters. He has also co-authored five national and international patents. Dr. Mazzeo serves as a reviewer for several international journals and book publishers. He has been regularly invited to take part in the scientific committees of national and international conferences.



Alessandro Bruno is an Associate Professor of Computer Science at IULM University (Milan, Italy) and Research Associate at CNR ISASI (Lecce, Italy). He was an Assistant Professor at Humanitas University, a Lecturer in Computing at Bournemouth University, and a Research Associate at NCCA (National Centre for Computer Animation) in the UK. He was a research visitor at the University College London (UCL). He held postdoctoral positions at the Italian National Institute for Astrophysics (INAF) and at the University of Palermo, where he earned his Ph.D. in Computer Engineering in 2012. He served as PI for a European research project, namely, HeReFaNMi, funded by NGI Search.

Contents

Preface	XV
Section 1	
Multiple Domains Transfer Learning	1
Chapter 1	3
Bridging Domains with Transfer Learning: Strategies for Multi-Task and Multi-Dataset Finetuning <i>by Tehseen Ullah, Steven Davy and John Kelleher</i>	
Chapter 2	21
Method for Developing Object and Image Recognition Systems Using Transfer Learning in Computer Vision <i>by Lenin Huayta Flores, Lucio Ticona Carrizales, Lenin Samir Huayta Poma and Betty Campos Segales</i>	
Chapter 3	39
Transfer Learning with CLIP Model for Multitasking Workload Prediction in Mixed Reality Environment <i>by Safanah Abbas, Heejin Jeong and David He</i>	
Chapter 4	59
Domain Adaptation in Multimodal Models <i>by Raghavendran Ramakrishnan</i>	
Section 2	
Knowledge and Data Representation	79
Chapter 5	81
Good Data Beats More Data: Building a Thermal Air-Image Dataset for Ground-to-Air Surveillance <i>by Tijeni Delleji, Feten Slimeni and Ahmed Siala</i>	
Chapter 6	99
Knowledge Aggregation in Autonomous Systems via Pre-Trained Model <i>by Muhammad Waqas Nawaz, Wanquan Zhang, Jaspreet Kaur, Qammer H. Abbasi, Muhammad Imran and Olaoluwa Popoola</i>	

Preface

A paradigm change is currently occurring in the field of artificial intelligence. Initially, a straightforward yet expensive formula, involving more data and more computation, described the course of machine learning. We started from scratch and trained ever-larger models from random initialization on enormous, task-specific datasets. Although effective, this strategy consumed a significant amount of resources, made it challenging to enter new markets, and was fragile in conditions with limited data.

A revolution is changing this traditional paradigm and is the subject of the book *Transfer Learning – Unlocking the Power of Pretrained Models*.

Developing AI systems in a more effective, clever, and approachable manner rather than creating standalone models on isolated datasets. For handling particular, frequently unique problems with exceptional speed, we utilize extensive, general-purpose “foundation models” that have already been trained on internet-scale data and modify their rich, learned representations. This is how transfer learning works. The journey we will take you on in this book moves from the theoretical foundations to the cutting-edge practical applications that are defining the next generation of intelligent systems.

This book compiles chapters that address the issue of transfer learning, helping you build your understanding in a progressive manner. The chapters are divided into two sections: Section I, *Multiple Domains Transfer Learning*, and Section II, *Knowledge and Data Representation*.

Section I contains four chapters:

- *Bridging Domains with Transfer Learning: Strategies for Multi-Task and Multi-Dataset Finetuning* provides you with an advanced discussion of the strategies and techniques to become a master practitioner, capable of designing sophisticated learning regimens that allow a single model to excel at multiple tasks by intelligently leveraging diverse datasets.
- *Method for Developing Object and Image Recognition Systems Using Transfer Learning in Computer Vision* proposes a comprehensive guide and methodology for building state-of-the-art vision systems without the need for massive datasets, demonstrating how transfer learning has become the default starting point for virtually all practical computer vision projects.
- *Transfer Learning with CLIP Model for Multitasking Workload Prediction in Mixed Reality Environment*. Here, the reader sees how visionary models like CLIP can be repurposed to solve unique challenges at the intersection of the physical and digital worlds, such as predicting system load to ensure seamless user experiences.

- *Domain Adaptation in Multimodal Models*, laying the groundwork for how knowledge can be transferred not just across tasks, but across entire data domains, even when dealing with complex, multi-modal data like text, images, and sensor feeds.

Section II has two chapters:

- *Good Data Beats More Data: Building a Thermal Air-Image Dataset for Ground-to-Air Surveillance* serves as a crucial case study. It underscores that while pre-trained models provide a powerful engine, thoughtfully curated, domain-specific data is the high-octane fuel that drives breakthrough performance in specialized fields, such as security and surveillance.
- *Knowledge Aggregation in Autonomous Systems via Pre-Trained Model* illustrates how autonomous agents, from self-driving cars to robotic manipulators, can use transfer learning to build a consolidated understanding of the world, combining lessons from numerous pre-trained subsystems to operate safely and effectively in unstructured environments.

This book is designed for engineers, researchers, students, and technology leaders who want not only to understand this shift but also to actively participate in it. Our goal is to provide you with both the conceptual framework and the practical expertise to harness the power of pre-trained models, saving you time and computational resources, and ultimately enabling you to solve problems that were previously out of reach.

Welcome to the future of machine learning. Welcome to the era of transfer learning.

Pier Luigi Mazzeo
National Research Council of Italy (CNR),
Institute of Applied Sciences and Intelligent Systems (ISASI),
Lecce, Italy

Alessandro Bruno
IUM University,
Milano, Italy

Section 1

Multiple Domains Transfer Learning

Bridging Domains with Transfer Learning: Strategies for Multi-Task and Multi-Dataset Finetuning

Tehseen Ullah, Steven Davy and John Kelleher

Abstract

The chapter explores the innovative integration of transfer learning with multi-dataset and multi-task fine-tuning to build versatile, resource-efficient models for computer vision-related tasks. It delves into the paradigm shift from traditional single-task training, where models are optimized for isolated datasets to a unified multi-task framework where a shared backbone supports diverse, domain-specific tasks. By leveraging pretrained models and implementing different multi-datasets and multi-tasks fine-tuning techniques such as mixture of experts, model merging, adaptive multi-task learning and modular architectures for multi-dataset training techniques, the discussed approaches enhance performance across varied datasets as well as significantly reduce computational cost and training time. This chapter addresses challenges such as data imbalance, conflicting gradients and sensitivity to merging coefficients, outlining how modern transfer learning techniques can reconcile these issues to produce scalable models that are both efficient and adaptable to real-world scenarios.

Keywords: finetuning, transfer learning, image classification, computer vision, multi-dataset finetuning, multi-task learning, resource efficient finetuning, fine-tuning techniques, mixture of experts

1. Introduction

The field of artificial intelligence (AI) has witnessed unprecedented growth over the past decade, driven largely by advances in deep learning. From diagnosing medical conditions to enabling autonomous vehicles, machine learning models have become indispensable tools for solving complex real-world problems. Yet, as these applications grow more sophisticated, a critical challenge emerges: how can we develop models that adapt efficiently to diverse tasks and datasets without requiring exorbitant computational resources or massive amounts of labeled data? This question lies at the heart of modern AI research and serves as the foundation for this chapter's exploration of transfer learning, a paradigm that re-imagines how models acquire, retain, and reuse knowledge across domains.

1.1 The evolution of model training: From isolation to integration

Traditional approaches to training machine learning models, particularly in computer vision, have long relied on a single-task and single-dataset framework. In this paradigm, a model is meticulously trained from scratch for a specific objective, such as classifying images in the CIFAR-100 [1] dataset or detecting objects in COCO [2]. While this method has yielded remarkable results, it suffers from significant limitations. For example, training models in isolation demands vast amounts of task-specific labeled data, a luxury unavailable in many practical scenarios, such as medical imaging or rare defect detection in manufacturing. Furthermore, these models often lack versatility; a network trained exclusively for facial recognition cannot pivot to segmenting tumors in MRI scans without starting over and training it from scratch for that specific application. This rigidity is compounded by the computational costs of training separate models for each task, which strains resources and hinders scalability [3].

To counter these limitations, researchers and engineers introduced the so-called transfer learning technique, a technique that has reshaped the AI landscape. At its core, transfer learning allows a model trained on one task or dataset to repurpose its knowledge for a related but distinct task. For instance, a neural network pretrained on millions of generic images can be “fine-tuned” to recognize specific types of plants with only a small botanical dataset [4]. This approach mimics how humans learn, applying foundational knowledge from one domain to accelerate mastery in another. Early applications of transfer learning focused on single-task adaptation, but the research community soon recognized its potential for broader integration and came up with the idea of what if a model could simultaneously tackle multiple tasks or generalize across multiple datasets, sharing insights between them? This question catalyzed a paradigm shift toward multi-task and multi-dataset fine-tuning, which is the focus of this chapter.

1.2 The promise of multi-task and multi-dataset learning

Imagine a single model that can analyze satellite imagery to predict weather patterns, detect deforestation, and monitor urban sprawl, all while drawing from datasets as varied as aerial photographs, climate logs, and social media feeds. Such a model would help save computational resources as well as uncover latent connections between seemingly unrelated tasks. This vision underpins the strategies discussed in this chapter. By unifying tasks and datasets under a shared framework, we can create models that are both resource-efficient and adaptable [5].

The key to this unification is the concept of a shared backbone, a common neural network architecture that processes input data and extracts features usable across multiple tasks. For example, the early layers of a convolutional neural network (CNN) might learn to detect edges and textures relevant to both object detection and semantic segmentation [6], while task-specific “heads” branch off from the backbone will learn to generate specialized outputs. This design drastically reduces redundancy. This means that instead of training separate models for each task, a model’s generalized features serve as a foundation, minimizing computational overhead. However, integrating multiple tasks and datasets is far from straightforward. Challenges such as data imbalance, conflicting gradients, and sensitivity to hyperparameters threaten to destabilize the training process. These obstacles demand innovative solutions such as: mixture of experts (MoE), model merging, adaptive multi-task learning, and other modular architectures. All of these innovative techniques will be discussed in more details in the later sections.

1.3 Bridging domains in real-world applications

The implications of these approaches extend far beyond academic benchmarks. Consider health care, where a model trained on diverse datasets such as X-rays from one hospital and pathology slides from another could improve diagnostic accuracy while preserving patient privacy through federated learning [7]. Similarly, in autonomous driving, a vehicle's perception system must simultaneously detect pedestrians, read traffic signs, and predict lane boundaries, often under varying weather conditions. A multi-task model trained on datasets spanning sunny, rainy, and snowy environments would inherently generalize better than a collection of single-task models [8].

Moreover, multi-dataset training mitigates the “domain gap” problem, where a model's performance plummets when applied to data that deviates from its training distribution. By exposing the model to varied datasets during fine-tuning, it learns robust features that transcend individual domains [9]. This is particularly vital in global applications, where data may reflect cultural, geographic, or demographic diversity.

1.4 Navigating challenges: Toward efficient and scalable AI

While the benefits of multi-task learning and multi-dataset fine-tuning are compelling, they raise ethical and practical considerations. For example, combining datasets risks amplifying biases if not carefully audited. A model trained on facial recognition data from predominantly one demographic group may perpetuate inequities when deployed in multi-ethnic regions [10]. Similarly, the computational savings of shared backbones must be weighed against the energy costs of training large models, a tension underscoring the need for efficiency-focused architectures [11].

This chapter does not merely catalog these challenges but provides a roadmap for addressing them. By reconciling conflicting gradients through adaptive loss weighting or designing modular systems that compartmentalize tasks, modern transfer learning techniques pave the way for scalable and sustainable AI systems.

2. Foundations of transfer learning in computer vision

Transfer learning (TL) has revolutionized computer vision by enabling models to repurpose knowledge from one task or dataset to solve new, related problems. This section builds the conceptual bedrock for the rest of the chapter, clarifying how TL evolved from its single-task origins to today's multi-task and multi-dataset paradigms. By anchoring technical strategies in their historical and theoretical context, we equip experts and non-experts alike to grasp the innovations discussed in later sections.

At its core, TL relies on the idea that models trained on large, diverse datasets learn reusable features such as textures, object parts that generalize across domains. For example, convolutional layers of a network pretrained on natural images can extract meaningful patterns from medical scans, reducing the need for extensive labeled healthcare data. **Figure 1** illustrates this process, where the pretrained convolutional layers are reused as a frozen feature extractor, while task-specific dense layers are replaced with a new classifier tailored to the target domain.

While early TL focused on adapting models to single tasks (e.g., fine-tuning ResNet for flower classification), real-world applications demand systems that handle multiple tasks (e.g., autonomous vehicles detecting pedestrians, traffic signs, and lanes simultaneously) and diverse datasets (e.g., merging satellite imagery and

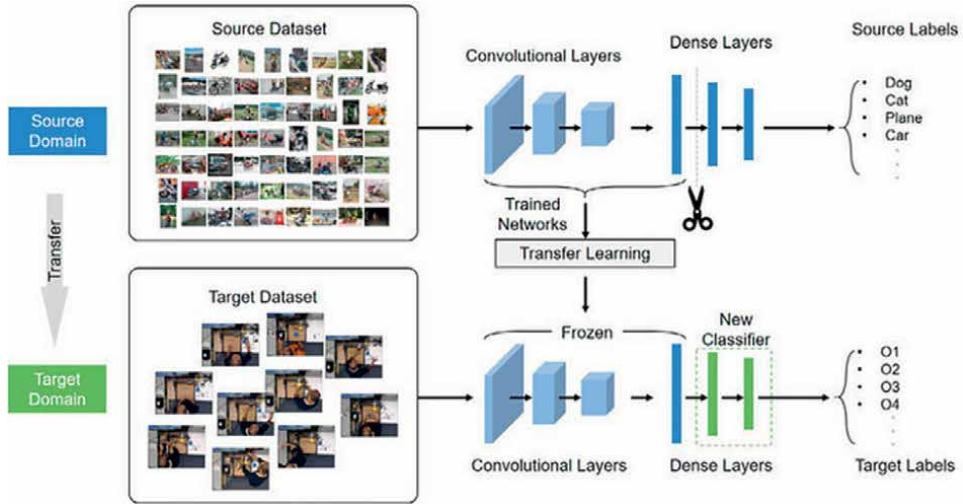


Figure 1. Illustration of the transfer learning process from a source to a target domain by reusing pretrained convolutional layers. The dense layers are replaced with a new classifier tailored to the target labels, while the feature extractor remains frozen.

street-view photos). This shift necessitated architectures and training strategies that balance efficiency, scalability, and performance.

2.1 From single-task adaptation to multi-domain generalization

The limitations of single-task TL became apparent as AI applications grew more complex. Training separate models for each task is computationally expensive and waste of resources. Moreover, features learned in isolation often fail to generalize across domains. The following breakthroughs addressed these issues:

2.1.1 Multi-task learning (MTL)

Multi-task learning (MTL) trains a single model to perform multiple tasks by sharing the same backbone network while using task-specific output heads. For instance, a shared CNN backbone can extract features for both object detection (bounding boxes) and semantic segmentation (pixel masks). Key innovations include:

- *Adaptive loss balancing:* Dynamically weighting task losses to prevent one task from dominating the training [12].
- *Gradient conflict mitigation:* Techniques like PCGrad [13] adjust gradients during backpropagation to avoid interference between tasks.

2.1.2 Multi-dataset training

Models trained on diverse datasets (e.g., combining COCO, ADE20K, and DomainNet) learn robust, domain-invariant features. For example, a model exposed to sketches, paintings, and photos can better recognize objects in abstract art [14]. However, merging datasets introduces the following challenges:

- *Label incompatibility*: Resolving mismatched annotation formats (e.g., bounding boxes vs. segmentation masks).
- *Domain gaps*: Aligning feature distributions across datasets with differing styles (e.g., medical scans vs. natural images).

Recent methods like dataset condensation and model soups [15, 16] address these issues by synthesizing representative data subsets or averaging weights of models fine-tuned on individual datasets.

2.2 The rise of generalist models

The pursuit of adaptable, multi-purpose systems have driven the emergence of generalist models capable of handling diverse tasks and modalities with minimal task-specific tuning. Central to this shift are advances in self-supervised learning (SSL) and vision-language models (VLMs), which learn rich, transferable representations without relying on labeled data. For instance, masked autoencoders like MAE [17] reconstruct images from partial inputs to capture spatial hierarchies, while VLMs such as CLIP [18] align text and image embeddings for zero-shot inference such as classifying medical scans using descriptive prompts. Complementing these innovations, parameter-efficient fine-tuning (PEFT) techniques like LoRA and adapters [19, 20] enable lightweight updates to massive pretrained models, making them viable for edge deployment. Together, these approaches reduce reliance on task-specific data and expensive computations, paving the way for models that fluidly adapt to multi-domain challenges, ranging from analyzing satellite imagery with text-guided prompts to diagnosing diseases across varied medical datasets while maintaining efficiency and scalability.

3. Strategies for multi-task and multi-domain fine-tuning

The shift from isolated, single-task models to unified systems capable of learning from multiple datasets and tasks represents a pivotal advancement in modern deep learning paradigm. However, this integration introduces unique challenges such as: How can models reconcile conflicting objectives across tasks? How can heterogeneous datasets, with mismatched labels, domains, or scales could be harmonized without catastrophic interference? In the following sections, we will explore some cutting-edge research and methodologies that address these challenges, balancing computational efficiency with robust performance. Central to these approaches is the principle of shared representation learning, where a single deep learning model extracts features relevant to diverse tasks (e.g., object detection and segmentation) and datasets (e.g., natural images and medical scans). Techniques like parameter-efficient fine-tuning (PEFT) minimize training costs by updating only small subsets of pre-trained weights, while adaptive multi-task learning (MTL) dynamically balances task losses to prevent dominance by data-rich objectives. Meanwhile, modular approaches like mixture of experts (MoE) and model merging enable multi-task learning without sacrificing scalability, routing inputs to domain-specific subnetworks or combining pretrained models into cohesive systems. **Figure 2** illustrates a unified view of prominent parameter-efficient fine-tuning (PEFT) approaches, including adapter tuning, prompt tuning, prefix tuning, and side tuning, each offering different trade-offs

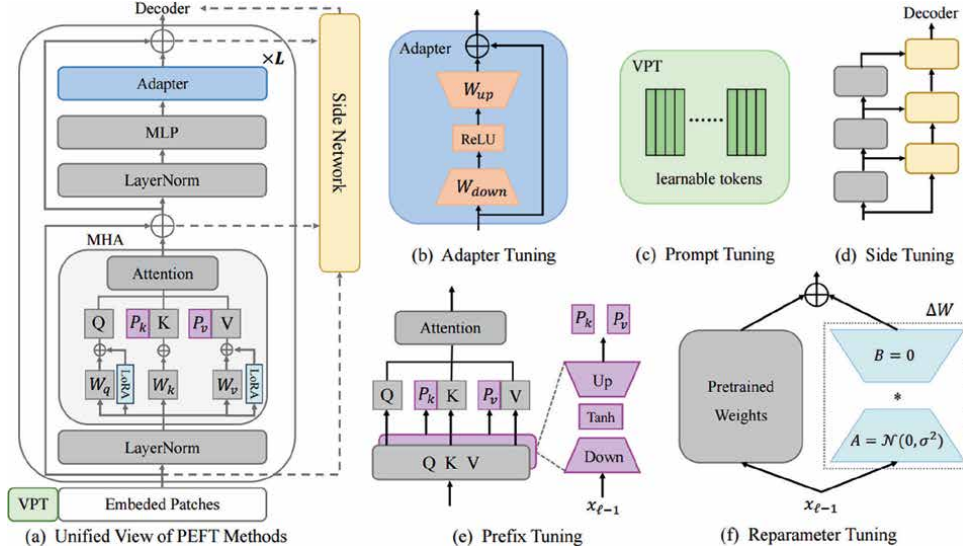


Figure 2.
Detailed architecture of various PEFT methods.

between trainable parameters, architectural modification, and performance efficiency. These innovative approaches collectively aim to overcome data imbalance, domain shifts, and gradient conflicts. The following subsections dissect these methodologies, highlighting their theoretical underpinnings, implementation trade-offs, and applications across different industries.

3.1 Parameter-efficient fine-tuning (PEFT)

Parameter-efficient fine-tuning (PEFT) has emerged as a pivotal method for adapting large-scale pretrained vision models (PVMs) to downstream tasks without the computational and storage burdens of full fine-tuning method. Unlike conventional approaches that update all of the model’s parameters, PEFT updates only a small subset of weights or introduces lightweight modules, preserving the pretrained knowledge while achieving competitive performance. This section synthesizes the latest advancements in PEFT, focusing on methodologies, applications, and challenges, with insights drawn from foundational and cutting-edge research.

3.1.1 Methodologies and taxonomies

PEFT methods are broadly categorized into three paradigms: addition-based, partial-based, and unified-based fine-tuning [21]. Addition-based methods insert trainable modules into the PVM architecture, such as adapters or prompts, while keeping the original weights frozen. For instance, adapter-based modules employ bottleneck structures with down-projection and up-projection layers to transform intermediate features. These modules, integrated into transformer layers, enable task-specific adaptation with minimal parameter overhead, typically 0.05–10% of the full model size. Vision prompt tuning (VPT) [22] extends this idea by prepending learnable prompts to input embeddings, aligning downstream data distributions with pretraining objectives.

On the other hand, partial-based methods update only specific subsets of existing parameters. For example, BitFit [23] fine-tunes bias terms exclusively, achieving over 95% of full fine-tuning performance on benchmarks, while LoRA [19] injects low-rank matrices into attention layers to approximate weight updates. These methods excel in efficiency, often training less than 1% of total parameters. Unified-based approaches combine multiple strategies into cohesive frameworks. For example, the VMT-Adapter [24], a state-of-the-art method, unifies shared projections for cross-task interaction and task-specific modules for dense scene understanding. By decomposing parameters via Kronecker products, VMT-Adapter-Lite further reduces trainable weights to 0.36% of the original model while maintaining competitive performance.

3.1.2 Applications and performance

PEFT has been widely applied to tasks ranging from image classification to dense scene understanding. As shown in **Table 1**, VMT-Adapter significantly outperforms full fine-tuning on PASCAL-Context dataset with only 1% of trainable parameters, achieving +3.96% performance gain, while its lightweight variant (VMT-Adapter-Lite) also provides strong results with minimal computational footprint, highlighting the practical advantage of unified PEFT methods in multi-task settings. Similarly, adapters like AdaptFormer [25] demonstrate robust performance on fine-grained visual classification (FGVC) and video action recognition, highlighting their versatility. A critical advantage of PEFT lies in its computational efficiency. Methods like LST [26] which decouple gradient backpropagation from the frozen backbone, reducing GPU memory usage by up to 70%. For trillion-parameter models, such efficiencies are indispensable. However, challenges persist, including gradient conflicts in multi-task settings and interpretability of learned prompts or adapters.

3.1.3 Challenges and future directions

Despite its promise, PEFT faces unresolved issues. Such as negative transfer, where irrelevant pretrained knowledge harms downstream performance, particularly in cross-domain scenarios [27]. Secondly, scaling PEFT to generative models (e.g., diffusion models) and multimodal systems (e.g., vision-language models) requires novel architectures. Recent work on DiffFit [28] fine-tunes diffusion models via bias and scaling terms, but broader exploration is needed. Finally, the lack of standardized libraries for visual PEFT hinders reproducibility. While NLP benefits from tools like Hugging Face’s PEFT, the vision community lacks equivalent resources, necessitating frameworks that unify diverse methods.

Method	Params (M)	Perf. Gain	Tasks
Full Fine-Tuning	112.62	0% (baseline)	Single
VMT-Adapter	1.13	+3.96%	Multi
VMT-Adapter-Lite	0.40	+1.34%	Multi
Shared Adapter	4.71	−0.64%	Multi

Table 1.
Performance comparison of parameter-efficient fine-tuning methods on dense scene understanding tasks.

3.2 Multi-task and multi-domain learning

Multi-task learning (MTL) enhances model generalization by training a single model to perform multiple related tasks simultaneously, leveraging shared representations to exploit task commonalities while minimizing interference. This approach reduces computational costs and improves data efficiency compared to training separate task-specific models. Multi-domain learning (MDL), conversely, focuses on adapting models to perform robustly across diverse domains such as datasets, environments, or distributions, addressing challenges like domain shifts and label incompatibility. Both paradigms are critical for scalable transfer learning, enabling models to generalize across tasks and domains with minimal retraining.

In order to achieve this, modern approaches toward multi-task and multi-domain learning integrate mixture of experts (MoE), which dynamically routes inputs to specialized subnetworks for task- or domain-specific processing, and multi-dataset training, which harmonizes heterogeneous data through techniques like domain-invariant feature alignment. Additionally, model merging techniques combine pretrained models into unified architectures, preserving their individual strengths while reducing redundancy. These methods are particularly valuable in vision tasks where labeled data is scarce or fragmented across domains.

Recent advancements in transfer learning increasingly rely on mixture of experts (MoE) architectures to address the dual challenges of multi-task learning (MTL) and multi-domain learning (MDL). By dynamically routing inputs to specialized subnetworks, MoE frameworks enable models to balance shared knowledge with task- or domain-specific adaptations, minimizing interference and enhancing generalization. Two pioneering implementations: MoE-FFD for face forgery detection [29] and AdaMV-MoE for multi-task vision recognition [30] exemplify how MoE principles advance scalable and efficient learning across tasks and domains.

3.2.1 Adaptive feature extraction with MoE-FFD

MoE-FFD tackles the challenge of detecting deepfakes across diverse datasets such as CelebDF-v2 and DFDC-P, by integrating the global context of vision transformers (ViTs) with the local sensitivity of CNNs. The architecture, as shown in **Figure 3**, preserves the pretrained ViT backbone while injecting domain-specific adaptations through two modular components:

- **Low-rank adaptation (LoRA):** Lightweight low-rank adaptation (LoRA) layers adjust the ViT's attention mechanisms to capture subtle forgery clues, such as identity inconsistencies or unnatural textures, without overwriting the foundational ImageNet knowledge.
- **Convolutional adapters:** Five specialized convolutional experts, including angular difference and central difference convolutions, localize artifacts like blending boundaries or noise patterns. Moreover, a gating network dynamically selects the optimal combination of experts for each input, ensuring domain-specific processing.

Their results show that MoE-FFD achieves an 86.78% average AUC on unseen datasets with only 15.51 M activated parameters, outperforming full ViT fine-tuning by a significant margin. Notably, it maintains robustness under perturbations like

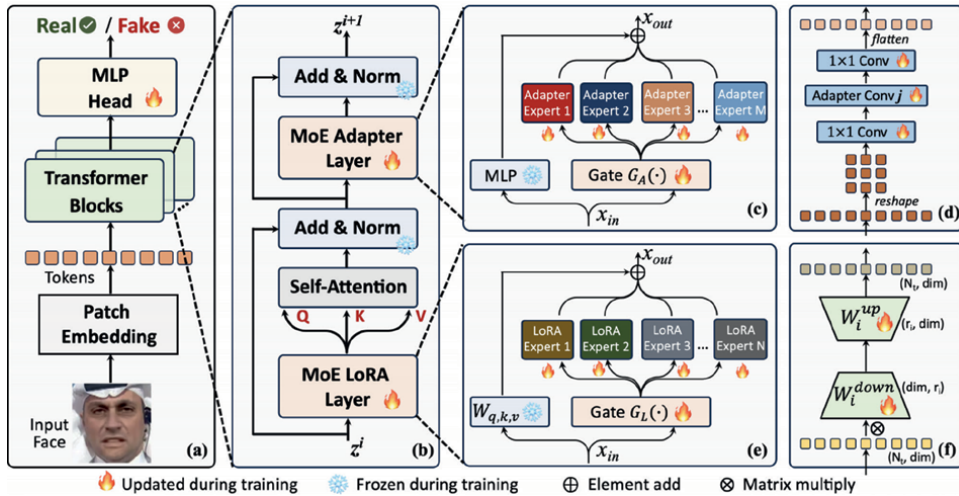


Figure 3.
 Overview of the designed MoE-FFD framework. (a) Overall model architecture; (b) MoE-FFD transformer block; (c) design of MoE Adapter layer; (d) Details of each Adapter expert; (e) Details of the designed MoE LoRA layer; (f) Details of each LoRA expert.

Gaussian blur, retaining 72.3% AUC at high severity levels compared to ViT-B's 54.1%. This efficiency stems from sparse activation where only relevant experts are engaged per input, making it ideal for deployment in resource-constrained environments.

3.2.2 Dynamic task specialization with AdaMV-MoE

AdaMV-MoE addresses the complexity of multi-task vision recognition through adaptive expert allocation. As illustrated in **Figure 4**, unlike conventional MTL models that rigidly share a backbone, AdaMV-MoE employs task-dependent routers and an adaptive expert selection (AES) mechanism. The AES monitors validation loss to dynamically adjust the number of experts per task, expanding capacity for complex tasks like detection while contracting it for simpler ones like classification. For example, detection tasks activate 50.94 M parameters compared to 42.95 M for classification, reflecting their higher computational demands. This approach mitigates gradient conflicts inherent in MTL by isolating task-specific computations. On the UViT-Base backbone [31], AdaMV-MoE achieves 79.65% classification accuracy and 44.14% detection AP, surpassing shared-backbone ViT models by 6.66% and 1.13%, respectively.

3.2.3 Model merging for multi-task learning

In scenarios where collecting joint training data for all tasks or domains is impractical due to privacy constraints or computational costs, model merging presents an efficient alternative for enabling multi-task learning (MTL). This means that instead of retraining a single model from scratch or maintaining multiple task-specific models, recent approaches aim to merge independently fine-tuned models into one unified architecture capable of performing multiple tasks. A particularly promising example of such approach is AdaMerging, a data-free and unsupervised method for adaptive model merging [32]. The core idea is to autonomously learn how to combine

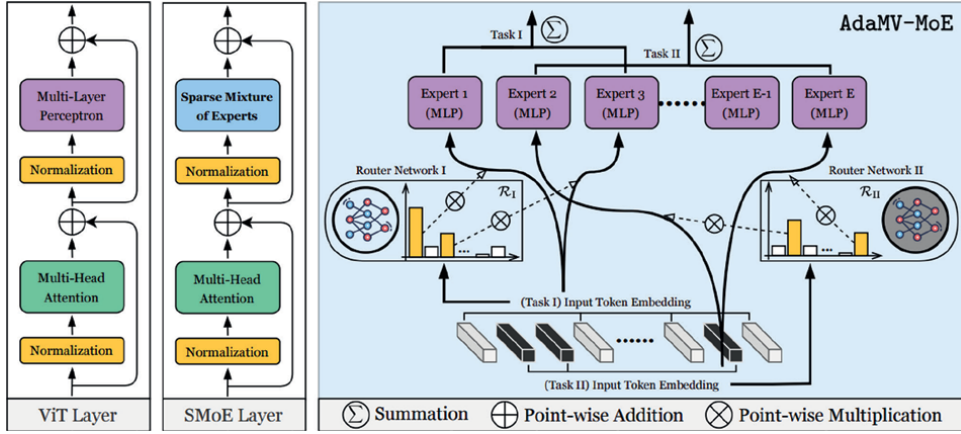


Figure 4.

High-level overview of AdaMV-MoE consists of both ViT (left) and SMoE (middle) layer, where the SMoE layer is built by replacing the original multi-layer perceptron (MLP) with a sparsely activated mixture of experts (MLPs). It enables multi-task learning by leveraging task-specific router networks (right). Each router network determines how many (adaptive) and which (specialized) experts are appropriate to activate for the given task.

multiple task-specific models, each fine-tuned on a shared pretrained backbone into a single multi-task model, without requiring access to the original training data.

AdaMerging builds upon the concept of task arithmetic, where each task is represented by a task vector, defined as the difference between the task-specific model parameters and the base pretrained model parameters. By summing these task vectors and merging them into the pretrained model, a new model can theoretically support multiple tasks. However, previous methods such as task arithmetic and TIES-Merging suffered from performance drops due to their reliance on fixed or manually-tuned merging coefficients, which often fail to resolve conflicts between tasks [33, 34]. To address this, AdaMerging introduces two key innovations:

- *Task-wise AdaMerging*: Assigns a unique merging coefficient to each task vector, allowing the model to weigh tasks differently based on their contribution to overall performance.
- *Layer-wise AdaMerging*: Further refines this approach by assigning distinct coefficients to each layer of each task vector, accounting for the observation that shallow layers typically encode generalizable features, while deeper layers capture task-specific nuances.

The innovation lies in how these coefficients are optimized. Rather than relying on grid search or labeled data, AdaMerging uses entropy minimization on unlabeled test samples as a surrogate objective. This is grounded in the finding that prediction entropy correlates strongly with model loss; lower entropy often indicates more confident and accurate predictions. By minimizing entropy during the merging process, AdaMerging effectively enhances model performance across tasks, even in the absence of supervision.

The experiments in the AdaMerging paper, conducted on diverse image classification tasks such as SUN397, SVHN, DTD, and EuroSAT using ViT-B/32 and ViT-L/14 models, demonstrate that layer-wise AdaMerging++ outperforms existing methods by up to 11% in average accuracy over task arithmetic baselines. Moreover, it shows:

- Superior generalization to unseen tasks, with improvements of up to 9.1%.
- Robustness under distribution shifts, outperforming competitors under corrupted data scenarios such as Gaussian noise and JPEG compression by up to 11.2%.

Importantly, this method is more accurate as well as more scalable because traditional techniques require manually tuning one coefficient per task or layer, which becomes infeasible at scale. AdaMerging eliminates this bottleneck through automated optimization using gradient descent and backpropagation on entropy. In the context of scalable transfer learning, AdaMerging aligns well with trends in model modularity and reusability. It preserves the benefits of individual models while unifying them into a single robust architecture. This has particular relevance for vision tasks with limited labeled data, where pretrained models are often fine-tuned independently for each domain or task.

In summary, AdaMerging represents a significant step forward in model merging for MTL, offering a practical, unsupervised, and data-efficient pathway to fuse multiple specialized models into one cohesive system. As such, it complements emerging paradigms like mixture of experts by providing a post-hoc merging strategy when dynamic routing or joint training is not feasible.

3.2.4 Scalable multi-task learning with disjoint datasets

Recent research in multi-task learning (MTL) increasingly explores scenarios where datasets are disjoint, meaning each task is associated with its own exclusive dataset, without overlapping samples or shared labels. This reflects many real-world applications, such as facial analysis or autonomous driving, where datasets for distinct tasks such as expression recognition and pose estimation come from different distributions or label spaces. Two key papers address this challenge using distinct and complementary strategies: MTL-SA [35] and heterogeneous task learning [36].

The paper “Beyond Without Forgetting: Multi-task Learning for Classification with Disjoint Datasets” introduces MTL-SA, a framework that builds upon the classic “Learning without Forgetting” (LwF) [37] paradigm, adapting it for disjoint datasets in classification tasks. In this setup, each task has a unique labeled dataset. In this scenario, conventional joint training fails because unlabeled datasets from other tasks are not directly utilized to improve learning. MTL-SA introduces a two-fold strategy to bridge this gap:

1. Augment each task using pseudo-labeled data from the other task’s dataset.
2. Selectively weight and integrate those samples based on pseudo-label confidence and domain similarity.

The method introduces label vector interpolation, where each unlabeled sample’s training label is a weighted combination of:

- *A soft label*: output from the previous model epoch, preserving learned knowledge.
- *A pseudo label*: converted from soft label by choosing the most confident class, adding new knowledge.

Their experimental results show that MTL-SA significantly outperformed both joint and alternating training baselines. For example, MTL-SA improves the expression recognition accuracy on SFEW dataset from 50.11% to 53.50% and age-stage classification accuracy on PETA dataset from 79.83% to 81.78%.

Moreover, MTL-SA excels in cases where training data is kept across tasks or domains and cannot be merged directly. It demonstrates how data selection and label interpolation can enhance MTL performance even in disjoint settings, making it highly relevant to transfer learning scenarios.

Another innovative method that tackle the challenge of jointly training on tasks with different data modalities, label spaces, and image properties is introduced in the paper “Multitask Learning with Heterogeneous Tasks”. These tasks use disjoint datasets which vary in complexity, resolution, and output types. In order to handle such heterogeneity, the authors propose an extension to the classic hard-parameter sharing strategy, by introducing task-specific stems before the shared encoder, where each task’s input passes through a stem unit that adapts its format such as resolution alignment, before reaching the common feature extractor. The overall structure involves a shared backbone with task-specific stems to preprocess inputs and separate prediction heads per task.

The proposed architecture is optimized using a weighted sum of individual task losses, similar to other MTL formulations. Unlike prior works like UberNet, which operates on a single dataset with multiple tasks, this study evaluates performance across four diverse datasets: CIFAR-10 and STL-10 for classification, MINI-COCO for object detection, and VOC0712 for semantic segmentation task. Furthermore, this work underscores the scalability benefits of MTL in resource-constrained settings, even across semantically and structurally diverse tasks. While performance on complex tasks like detection might slightly lag, the overall parameter reduction and performance gains on simpler tasks make this approach practical for real-world applications such as edge computing.

4. Conclusions

Transfer learning represents a significant paradigm shift within artificial intelligence field, particularly within computer vision area, by fostering more efficient, scalable, and versatile model development. Throughout this chapter, we have explored how modern AI practices have evolved from isolated single-task learning approaches toward unified multi-task and multi-dataset frameworks. By leveraging shared neural network architectures, transfer learning facilitates the reuse of learned representations, thereby significantly reducing computational overhead and training times.

Various advanced strategies for fine-tuning, including parameter-efficient fine-tuning (PEFT), mixture of experts (MoE), adaptive multi-task learning (MTL), and model merging, have been comprehensively reviewed. Techniques like LoRA, adapters, and vision prompt tuning (VPT) demonstrate that substantial improvements in efficiency and effectiveness can be achieved by updating only small subsets of pretrained weights or adding lightweight trainable modules. Moreover, integrating MoE architectures have notably enhanced the capabilities of models in handling multi-task and multi-domain challenges. These architectures dynamically route inputs to specialized subnetworks, thus effectively mitigating issues such as conflicting gradients, domain discrepancies, and data imbalances. Techniques like MoE-FFD and AdaMV-MoE further underscore the potential of these approaches by demonstrating substantial performance gains in complex, real-world scenarios.

Model merging strategies, such as adaptive merging methods like AdaMerging, illustrate another critical advancement, allowing the fusion of independently fine-tuned models into cohesive multi-task models without requiring additional labeled data. These methods have successfully addressed previous limitations such as sensitivity to merging coefficients and have showcased considerable improvements in scalability and generalization capabilities.

However, while these strategies offer promising avenues for advancing AI efficiency and adaptability, they also bring forth critical ethical and practical considerations. Issues such as potential biases in multi-dataset training and the energy costs associated with large-scale models underscore the importance of ongoing research into ethical and sustainable practices in model training and deployment.

The techniques and insights presented in this chapter broaden the applicability of transfer learning across diverse tasks and domains as well as significantly advance the pursuit of creating adaptable, robust, and resource-conscious AI systems. Continued innovation in these areas will undoubtedly shape the future landscape of artificial intelligence, aligning technological advancements with real-world demands and ethical imperative.

Acknowledgements

This publication has emanated from research supported in part by grants from Research Ireland under Grant numbers SFI/21/FFP-A/9174, SFI/12/RC/2289_P2. For the purpose of open access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

Abbreviations

AI	artificial intelligence
MTL	multi-task learning
MDL	multi-domain learning
TL	transfer learning
PEFT	parameter-efficient fine-tuning
MoE	mixture of experts
VPT	vision prompt tuning
LoRA	low-rank adaptation
FGVC	fine-grained visual classification
MLP	multi-layer perceptron
AES	adaptive expert selection
AUC	area under curve
AdaMV-MoE	adaptive multi-task vision mixture-of-experts
FFD	face forgery detection
ViT-B	vision transformer base
ViT-L	vision transformer large
MTL-SA	multi-task learning with selective augmentation
LwF	learning without forgetting
PCGrad	projected conflicting gradient
SSL	self-supervised learning
VLM	vision-language model

Author details

Tehseen Ullah^{1*†}, Steven Davy^{1†} and John Kelleher²

¹ Technological University Dublin, Dublin City, Ireland

² Trinity College, Dublin, Ireland

*Address all correspondence to: d23125133@mytudublin.ie

† These authors contributed equally.

IntechOpen

© 2025 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

References

- [1] Krizhevsky A, Hinton GE. Learning multiple layers in deep convolutional neural networks for image classification. In: Pereira FC, Burges CJ, editors. *Advances in Neural Information Processing Systems*. Vol. 22. Red Hook, NY: Curran Associates, Inc.; 2009. pp. 1097-1105
- [2] Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft COCO: Common objects in context. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T, editors. *Computer Vision – ECCV 2014*. Cham: Springer International Publishing; 2014. pp. 740-755
- [3] Zhang W, Shen L, Foo CS. Rethinking the role of pre-trained networks in source-free domain adaptation. In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. Los Alamitos, CA: IEEE Computer Society; 2022. pp. 18795-18805. Available from: <https://api.semanticscholar.org/CorpusID:254686034>
- [4] Kornblith S, Shlens J, Le QV. Do better ImageNet models transfer better?. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 2019. Long Beach, CA: IEEE; 2019. pp. 2656-2666. DOI: 10.1109/CVPR.2019.00277
- [5] Zhang Y, Yang Q. A Survey on Multi-Task Learning. *IEEE Transactions on Knowledge and Data Engineering*. 2017. DOI: 10.1109/TKDE.2021.3070203
- [6] Zamir A, Sax A, Shen W, Guibas L, Malik J, Savarese S. Taskonomy: Disentangling task transfer learning. *arXiv*. 2018;1804:08328. DOI: 10.48550/arXiv.1804.08328
- [7] Guan H, Yap PT, Bozoki A, Liu M. Federated learning for medical image analysis: A survey. *Pattern Recognition*. 2023;151:110424
- [8] Zhang Y, Carballo A, Yang H, Takeda K. Perception and sensing for autonomous vehicles under adverse weather conditions: A survey. *ISPRS Journal of Photogrammetry and Remote Sensing*. 2023;196:146-177. Available from: <https://www.sciencedirect.com/science/article/pii/S0924271622003367>
- [9] Yeo T, Kar OF, Sax A, Zamir A. Robustness via cross-domain ensembles. *arXiv*. 2021;2103:1-12. Available from: <https://arxiv.org/abs/2103.10919>
- [10] Buolamwini J, Gebru T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Friedler SA, Wilson C, editors. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. Vol. 81 of *Proceedings of Machine Learning Research*. [Internet]. Brookline, MA: PMLR; 2018. pp. 77-91. Available from: <https://proceedings.mlr.press/v81/buolamwini18a.html>
- [11] Strubell E, Ganesh A, McCallum A. Energy and policy considerations for deep learning in NLP. In: Korhonen A, Traum D, Màrquez L, editors. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics; 2019. pp. 3645-3650. Available from: <https://aclanthology.org/P19-1355/>
- [12] Kendall A, Gal Y, Cipolla R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. *arXiv*. 2018;1705:1-13. Available from: <https://arxiv.org/abs/1705.07115>

- [13] Yu T, Kumar S, Gupta A, Levine S, Hausman K, Finn C. Gradient surgery for multi-task learning. arXiv. 2020;**2001**:1-15. Available from: <https://arxiv.org/abs/2001.06782>
- [14] Wang X, Li H, Fang H, Peng Y, Xie H, Yang X, et al. LineArt: A knowledge-guided training-free high-quality appearance transfer for design drawing with diffusion model. arXiv. 2024;**2412**:1-14. Available from: <https://arxiv.org/abs/2412.11519>
- [15] Zhao B, Bilen H. Dataset condensation with distribution matching. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision 2023 (WACV). IEEE Winter Conference on Applications of Computer Vision. United States: Institute of Electrical and Electronics Engineers; 2023. pp. 6503-6512. Available from: <https://wacv2023.thecvf.com/>
- [16] Wortsman M, Ilharco G, Gadre SY, Roelofs R, Gontijo-Lopes R, Morcos AS, et al. Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. arXiv. 2022;**2203**:1-17. Available from: <https://arxiv.org/abs/2203.05482>
- [17] He K, Chen X, Xie S, Li Y, Dollar P, Girshick R. Masked autoencoders are scalable vision learners. In: Proceedings - 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Los Alamitos, CA: IEEE Computer Society; 2022. pp. 15979-15988
- [18] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. arXiv. 2021;**2103**:1-24. Available from: <https://arxiv.org/abs/2103.00020>
- [19] Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. LoRA: Low-rank adaptation of large language models. arXiv. 2021;**2106**:1-15. Available from: <https://arxiv.org/abs/2106.09685>
- [20] Houlsby N, Giurgiu A, Jastrzebski S, Morrone B, De Laroussilhe Q, Gesmundo A, et al. Parameter-efficient transfer learning for NLP. In: Chaudhuri K, Salakhutdinov R, editors. Proceedings of the 36th International Conference on Machine Learning. Vol. 97 of Proceedings of Machine Learning Research. Brookline, MA: PMLR; 2019. pp. 2790-2799. Available from: <https://proceedings.mlr.press/v97/houlsby19a.html>
- [21] Xin Y, Luo S, Zhou H, Du J, Liu X, Fan Y, et al. Parameter-efficient fine-tuning for pre-trained vision models: A survey. arXiv. 2024;**2401**:1-20. Available from: <https://api.semanticscholar.org/CorpusID:267412110>
- [22] Jia M, Tang L, Chen BC, Cardie C, Belongie S, Hariharan B, et al. Visual Prompt Tuning. In: Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXIII. Berlin, Heidelberg: Springer-Verlag; 2022. pp. 709-727. DOI: 10.1007/978-3-031-19827-4_41
- [23] Ben Zaken E, Goldberg Y, Ravfogel S. BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In: Muresan S, Nakov P, Villavicencio A, editors. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Dublin, Ireland: Association for Computational Linguistics; 2022. pp. 1-9. Available from: <https://aclanthology.org/2022.acl-short.1/>
- [24] Xin Y, Du J, Wang Q, Lin Z, Yan K. VMT-adapter: Parameter-efficient

transfer learning for multi-task dense scene understanding. arXiv. 2023;**2312**:1-14. Available from: <https://arxiv.org/abs/2312.08733>

[25] Chen S, Ge C, Tong Z, Wang J, Song Y, Wang J, et al. AdaptFormer: Adapting vision transformers for scalable visual recognition. arXiv. 2022;**2205**:1-10. Available from: <https://arxiv.org/abs/2205.13535>

[26] Sung YL, Cho J, Bansal M. LST: Ladder side-tuning for parameter and memory efficient transfer learning. In: Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K, Oh A, editors. Advances in Neural Information Processing Systems. Vol. 35. Red Hook, NY: Curran Associates, Inc.; 2022. pp. 12991-13005. Available from: https://proceedings.neurips.cc/paper_files/paper/2022/file/54801e196796134a2b0ae5e8adef502f-Paper-Conference.pdf

[27] Raghu M, Unterthiner T, Kornblith S, Zhang C, Dosovitskiy A. Do vision transformers see like convolutional neural networks? In: Ranzato M, Beygelzimer A, Dauphin Y, Liang PS, Vaughan JW, editors. Advances in Neural Information Processing Systems. Vol. 34. Red Hook, NY: Curran Associates, Inc.; 2021. pp. 12116-12128. Available from: https://proceedings.neurips.cc/paper_files/paper/2021/file/652cf38361a209088302ba2b8b7f51e0-Paper.pdf

[28] Xie E, Yao L, Shi H, Liu Z, Zhou D, Liu Z, et al. DiffFit: Unlocking transferability of large diffusion models via simple parameter-efficient fine-tuning. arXiv. 2023;**2304**:1-15. Available from: <https://arxiv.org/abs/2304.06648>

[29] Kong C, Luo A, Bao P, Yu Y, Li H, Zheng Z, et al. MoE-FFD: Mixture of experts for generalized and parameter-efficient face forgery detection. arXiv.

2024;**2404**:1-12. Available from: <https://arxiv.org/abs/2404.08452>

[30] Chen T, Chen X, Du X, Rashwan A, Yang F, Chen H, et al. AdaMV-MoE: Adaptive multi-task vision mixture-of-experts. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Los Alamitos, CA: IEEE Computer Society; 2023. pp. 17300-17311

[31] Chen W, Du X, Yang F, Beyer L, Zhai X, Lin TY, et al. A simple single-scale vision transformer for object localization and instance segmentation. arXiv. 2022;**2112**:1-17. Available from: <https://arxiv.org/abs/2112.09747>

[32] Yang E, Wang Z, Shen L, Liu S, Guo G, Wang X, Tao D. AdaMerging: Adaptive model merging for multi-task learning. In: Proceedings of the Twelfth International Conference on Learning Representations (ICLR). Vol. 2024. [Internet]. OpenReview; 2024. pp. 1-16. Available from: <https://openreview.net/forum?id=nZP6NgD3QY>

[33] Ilharco G, Ribeiro MT, Wortsman M, Schmidt L, Hajishirzi H, Farhadi A. Editing models with task arithmetic. In: Proceedings of the Eleventh International Conference on Learning Representations (ICLR). arXiv preprint arXiv. 2023;**2212**:1-18. Available from: <https://openreview.net/forum?id=6t0Kwf8-jrj>

[34] Yadav P, Tam D, Choshen L, Raffel C, Bansal M. TIES-Merging: Resolving interference when merging models. arXiv. 2023;**2306**:1-12. Available from: <https://arxiv.org/abs/2306.01708>

[35] Hong Y, Niu L, Zhang J, Zhang L. Beyond without forgetting: multi-task learning for classification with disjoint datasets. arXiv. 2020;**2003**:1-10. Available from: <https://arxiv.org/abs/2003.06746>

[36] Kim C, Kim E. Multitask learning with heterogeneous tasks. In: 2022 13th International Conference on Information and Communication Technology Convergence (ICTC). Piscataway, NJ: IEEE; 2022. pp. 1024-1026

[37] Li Z, Hoiem D. Learning without forgetting. arXiv. 2017;**1606**:1-11. Available from: <https://arxiv.org/abs/1606.09282>

Method for Developing Object and Image Recognition Systems Using Transfer Learning in Computer Vision

*Lenin Huayta Flores, Lucio Ticona Carrizales,
Lenin Samir Huayta Poma and Betty Campos Segales*

Abstract

Object and image recognition through transfer learning is a key area in computer vision and deep learning. This process encompasses several fundamental stages for developing efficient models. It begins with data collection, ensuring a representative and high-quality dataset with proper annotation and regulatory compliance. Next, preprocessing enhances image quality using techniques, such as normalization and segmentation. Subsequently, the use of pretrained deep neural networks leverages prior knowledge to optimize learning. During the training phase, the model fine-tunes its parameters through iterative optimization to identify relevant patterns. In the validation phase, the model's performance is evaluated, allowing for hyperparameter tuning to improve generalization and robustness across different scenarios. Finally, in the recognition phase, the trained model makes predictions on new data, enabling automated decision-making or actions. This structured approach enhances recognition accuracy and model generalization, ensuring robust performance across various computer vision applications.

Keywords: computer vision, convolutional neural networks, image preprocessing, image recognition, transfer learning

1. Introduction

In the current context of artificial intelligence (AI), image recognition systems have gained increasing relevance in various fields, such as medicine, security, agriculture, and industry, thanks to advances in deep learning. Transfer learning (TL) reuses a pretrained architecture (e.g., Classification: AlexNet, ConvNeXt, MaxVit, RegNet, ShuffleNet V2, SwinTransformer, Wide ResNet. Semantic Segmentation: DeepLabV3, FCN (fully convolutional network), and LRASPP (Lite Reduced Atrous Spatial Pyramid Pooling). Object Detection: Faster R-CNN (faster region-based

convolutional neural network), FCOS (fully convolutional one-stage object detection), RetinaNet, SSD (single shot detector), and SSDlite. Video Classification: Video MViT (multiscale vision transformer), Video ResNet, Video S3D, and Video SwinTransformer. Optical Flow: RAFT (Recurrent All-Pairs Field Transforms). Instance Segmentation: Mask R-CNN (mask region-based convolutional neural network). Keypoint Detection: Keypoint R-CNN. Quantized Models: Quantized GoogLeNet, Quantized MobileNet V2, Quantized MobileNet V3, Quantized ShuffleNet V2, Quantized InceptionV3, Quantized ResNet, and Quantized ResNeXt) for a new problem [1, 2] or for a similar task. This is particularly useful when data are scarce, private, or costly to obtain or label, as it enables leveraging knowledge from a broader domain [2, 3]. There are multiple approaches to transfer learning: reusing only the network structure, using both the architecture and the pretrained weights—either partially adjusting (e.g., final layers) or fully fine-tuning all layers—or adding additional layers to adapt the model to the new task. This technique optimizes computational resources and improves performance on problems with limited or small-scale data [4]. This chapter presents a comprehensive methodological approach for the development of an image recognition system aimed at computer vision (CV) tasks, based on modern deep learning techniques.

The proposed methodology consists of six sequential phases: data collection, image preprocessing, pretraining, training, validation, and recognition. Each of these phases fulfills a specific and complementary role within the overall process. The data collection phase includes the precise definition of the task, acquisition and labeling of datasets, as well as verifying their quality and compliance with legal and ethical requirements. Image preprocessing aims to optimize the model's input conditions through techniques, such as resizing, normalization, contrast adjustment, segmentation, and data augmentation. Next, the pretraining phase incorporates previously trained models by leveraging transfer learning or feature extraction, allowing for reduced training time and improved performance. The training phase focuses on pattern learning by DCNNs (deep convolutional neural networks), using optimization algorithms such as gradient descent. The model validation phase evaluates performance with metrics like accuracy, precision, and recall, using methods like cross-validation and hyperparameter tuning to prevent overfitting.

Finally, the recognition phase involves applying the trained model to real-world contexts, enabling high-accuracy image identification and classification. This chapter details each stage, offering a practical and conceptual guide for those developing intelligent computer vision systems.

2. Phases

The six phases comprise a sequential methodological process that guides the development of image recognition systems based on deep learning and transfer learning. These phases are data collection, image preprocessing, pretraining, training, validation, and recognition, as shown in **Figure 1**. It is worth mentioning that, although they are presented in an orderly manner, the completion of these phases is not necessarily rigid or strict. In practice, the process may require iterations between phases, on-the-fly adjustments, or even the omission of some stages depending on the context, the nature of the problem, data availability, technical resources, or the specific objectives of the project.

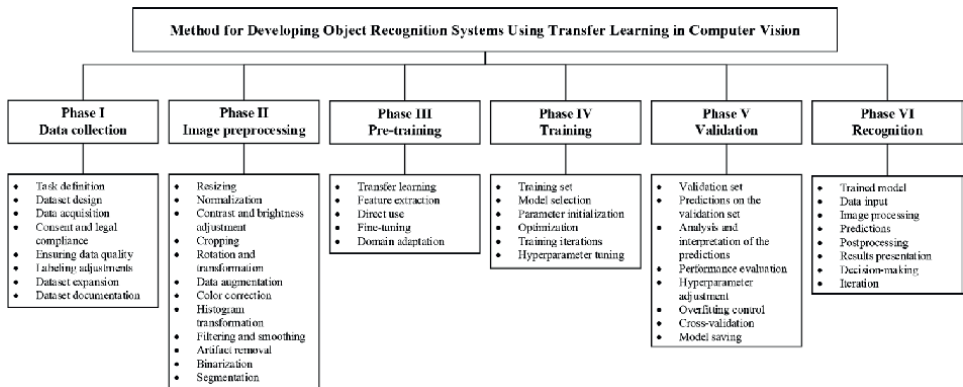


Figure 1.
Method for developing object recognition systems using transfer learning in computer vision.

2.1 Data collection

A dataset of images is a structured collection of visual representations, categorized into one or more classes, stored locally or in the cloud, and designed to be processed and used for training machine learning (ML) models. In this phase, the image dataset [5–9] containing examples or data related to the task at hand is obtained. The framework begins with defining the task and designing the dataset to align it with existing resources, followed by data acquisition and immediate legal compliance verification. Subsequently, data quality is evaluated, and labeling adjustments are made, applying dataset augmentation if the information is insufficient. Finally, each stage is continuously documented and iterated as necessary, ensuring efficient, ethical, and effective use of available data and models while minimizing redundancy in the process. Data collection can be carried out through a logistical activity (e.g., setting up traffic cameras), a technical activity (software connected to a digital catalog database), or a commercial activity (purchasing an image archive) [10]. Data sharing and reuse [11]; discovery and exploration [12].

Task definition involves understanding the image recognition task to be performed; whether it is object recognition [13–16] (locating multiple instances with bounding boxes), image classification [17, 18] (assigning a global label), or face detection [19–25] (identifying human faces, possibly extending to facial recognition). Each variant requires different technical approaches, from neural network architectures (YOLO (You Only Look Once), ResNet (Residual Network), MTCNN (multi-task cascaded convolutional neural network)) to specialized datasets (COCO (Common Objects in Context), ImageNet, WIDER Face), determining the choice of metrics and computational resources needed for effective implementation.

The design of the image dataset, or acquiring a representative dataset for the task, is a fundamental stage in developing computer vision systems. This process involves the careful collection of images from diverse sources, such as digital cameras, scanners, mobile devices, specialized sensors, or public databases [26] available in academic repositories and specialized platforms. Additionally, when existing data are insufficient or do not meet specific project requirements, it may be necessary to manually generate images [27] using techniques like controlled photography, three-dimensional (3D) rendering, or image synthesis via generative algorithms. It is crucial that the acquired dataset is diverse, balanced (SMOTE (Synthetic Minority

Oversampling Technique)) [28], and representative [29, 30] of the real-world conditions where the system will operate, considering factors such as lighting variations, angles, resolutions, and possible noise or artifacts to ensure that trained models can generalize properly in practical scenarios.

Data acquisition is the process of finding datasets to train models [31]. Three approaches are identified [32]: Data discovery involves searching and indexing datasets in repositories such as corporate data lakes or the web using tools like DataHub, Google Fusion Tables, or Kaggle, which combine collaborative analysis, public management, and access to massive repositories. Search specializes in locating data either in corporate lakes or via HTML (Hypertext Markup Language) table extraction. Data augmentation complements existing datasets by integrating external sources or generating new data through crowdsourcing or synthetic data [31, 33]. Finally, data generation creates datasets from scratch using crowdsourcing (with procedural or declarative tasks) or synthetic techniques, such as generative adversarial networks (GANs) [34], reinforcement learning policies, or specific methods (Semantic Clustering and Partitioning (SCPN)/Semantic Extraction and Relation (SEAR) in natural language processing (NLP)) [32, 33].

Consent and legal compliance entail that collecting visual data involving people or objects in public or private spaces must be conducted under strict adherence to legal frameworks related to information protection and privacy [35, 36]. When collected data allow direct or indirect identification of individuals, obtaining informed, free, and explicit consent is essential, following principles of legality, transparency, and accountability. Complementary measures, such as data minimization [37, 38], anonymization [39, 40], and access controls, must be considered to ensure image processing is proportional, secure, and consistent with predefined purposes.

Ensuring the quality of collected data involves cleaning the data by removing defective images (artifacts, low resolution, distortions), correcting labeling errors by verifying that each label exactly corresponds to the visual content and fixing ambiguities or inconsistencies, and normalizing data [41] by standardizing formats, sizes, color spaces, and statistical distributions to guarantee consistency during training. Data quality [42–44] and labeling accuracy are critical, as errors in this phase can propagate during training and negatively impact the model's final performance.

Labeling adjustments are made to effectively train a deep learning model, usually requiring prior data labeling [4, 45, 46]. This process assigns specific labels or categories [47] to each image in the dataset, enabling the model to learn patterns through supervised learning. Labeling data, when building knowledge bases from the web, initially assumes extracted facts are correct. Techniques include using existing labels [31]; crowd-based methods such as active learning and collective collaboration [32]; weak supervision, which generates imperfect labels at scale (e.g., data programming using multiple labeling functions instead of individual ones) [31, 48]; and data extraction leveraging existing knowledge bases to obtain entity attributes. These strategies balance volume, accuracy, and efficiency, adapting to contexts with initially scarce labels.

Dataset augmentation is performed if the existing image dataset is small and it is necessary to increase its size to improve model performance; this includes collecting more data, generating synthetic data, or increasing diversity within the dataset [2, 49–51].

Dataset documentation is very important and should include information about its origin, label structure, and any other relevant details so others can understand and use the dataset [52, 53].

2.2 Image preprocessing

This second phase consists of the manipulation and enhancement of images before they are used in model training or other applications.

Resizing involves adjusting the size of images while maintaining the aspect ratio (width/height). We can resize an image with a defined aspect ratio [54], which is the width of the image divided by its height.

Normalization ensures that pixel values are within a specific range, such as [0, 1] or $[-1, 1]$ [55]. Normalization helps standardize data and speeds up the training process.

Adjusting the contrast and brightness of images improves the visibility of details [56]; this is useful when working with images of variable quality.

Cropping removes pixels from an image's edges [57]; sometimes, it is necessary to crop the region of interest to focus on a specific part of the image for recognition tasks.

Rotation and transformation involve aligning all images to the same orientation or applying geometric transformations to correct distortions. In image processing, the rotation anchor point is the top-left corner at (0, 0) [58]. An image is rotated by an angle θ by transforming the (x, y) coordinates of each pixel relative to the origin [59, 60]. Morphological transformations, focused on the shape and structure of image features [61], include basic operations like erosion, which shrinks objects by removing edge pixels, and dilation, which expands objects by adding pixels [62]. Opening smooths contours and removes protrusions; closing fills holes and connects regions [63]. Advanced techniques such as hit-or-miss transformation (specific shape detection), morphological gradient (difference between dilation and erosion to highlight edges), and top-hat transforms (extracting small elements: white for bright, black for dark) optimize structure analysis [62, 64]. Algorithms like boundary extraction, hole filling, skeletonization (structural representation via iterative erosions), and convex hull (convexity approximation) allow describing shapes and patterns [65]. Operations such as thinning (thinning without losing connectivity) and pruning (removal of spurious pixels) refine results, while morphing (interpolation between images) gradually transforms structures [66]. Applicable to binary and grayscale images, these techniques are essential in pattern recognition, medical image processing, and computer vision [67]. Geometric transformations rearrange image pixels through basic operations, such as translation (displacement along x and y axes using matrices) [54, 57, 68], rotation (coordinate adjustment with trigonometric formulas for angle θ) [58, 60], scaling (resizing while preserving aspect ratio) [54, 69], flipping (reflection on axes) [58, 70], and cropping (extracting subregions using window and plop operators) [66]. Techniques like interpolation—nearest neighbor, bilinear, and bicubic—adjust resolution while preserving visual quality [71, 72], whereas advanced transformations like affine (combination of rotation, scaling, shear, and translation in homogeneous coordinates) [61] and projective (mapping quadrilaterals for perspective correction) [73] allow complex manipulation. Methods like deformation [61, 72], polar mapping [66], and barrel/pincushion transformations (non-linear radial distortion) [66] adapt images to specific geometric systems. Image arithmetic, consisting of addition, subtraction, and multiplication with values clipped to [0, 255] [54, 68], and masking (focusing on specific areas) [54, 74] optimize data for analysis, while techniques like Procrustes alignment (translation, scaling, and rotation to preserve shape) [61] ensure accuracy in computer vision applications. These operations, based

on matrix algebra and coordinate systems [60, 66], are essential in image processing to improve quality, adaptability, and computational efficiency.

Data augmentation creates modified images (rotations, shifts, brightness) to boost diversity and improve model generalization. Models trained with data augmentation on minority classes and transfer learning provide the highest classification accuracies [49]. There are many methods for data augmentation, among which Kumar [75] mentions deriving latent semantics, entity augmentation, and data integration. Advanced data augmentation techniques include adversarial training [76], synthetic data generation: GANs [34] or variational autoencoders (VAEs) [77], advanced geometric transformations, semantic-based augmentations, Mixup [78], CutMix [79], and controlled noise-based data augmentation.

Color correction involves adjusting the color balance of images to correct lighting and color representation problems [80, 81].

Image histogram transformation is an abstraction of an image where the frequency of each image value (contrast/brightness/intensity) is determined [59, 71, 74, 82]. A histogram shows the distribution of pixel intensities in an image, visualized as a graph that summarizes intensity levels [74]. Types of histograms include one-dimensional (1D) histograms or grayscale image histograms that graphically represent the distribution of pixel values, where the x-axis indicates intensities (0–255) and the y-axis their frequencies [74, 82]. In color images, histograms can analyze individual channels (RGB (Red, Green, Blue), HSL (Hue, Saturation, Lightness)) or two-dimensional (2D)/3D combinations, though the latter require reduced quantization to avoid complexity [64, 71]. Masks allow focusing on specific regions by calculating partial histograms to adjust contrast or segment objects [74]. Histogram equalization improves contrast by redistributing values through the cumulative distribution function (CDF), idealizing a uniform histogram [64, 83]. Techniques like histogram matching transform values to align distributions with a desired target, useful in color correction [83]. Additionally, histogram comparison and backprojection facilitate retrieving similar images or detecting objects through probability analysis [64, 71]. These tools are essential in digital processing to optimize quality, segmentation, and comparative analysis. Histogram transformation consists of aligning the histogram of an image to redistribute intensity levels to highlight certain features.

Regarding filtering and smoothing, noise can be caused by factors, such as poor lighting or distortion during image acquisition; filtering techniques are used to remove unwanted details, improve, smooth, or enhance features in images. Filters modify the appearance of an image through operations in the spatial domain (direct manipulation of pixels) or frequency domain (transforms like Fourier). In the spatial domain, linear filters (weighted combination of neighboring pixels) include the mean filter (smoothing but blurs edges) [61] and Gaussian filter (noise reduction with kernel based on σ) [67], while non-linear filters like the median filter (removes “salt and pepper” noise) [61] and order filters (max/min to enhance bright/dark points) [67] preserve edges. Edge detection filters use derivatives: Sobel and Prewitt (horizontal/vertical gradients) [84], Canny (non-maximum suppression and hysteresis thresholding) [85], and Laplacian (second derivative for zero crossings) [61]. In the frequency domain, low-pass filters (smoothing: ideal, Butterworth, Gaussian) [60] and high-pass filters (edge enhancement: Butterworth, Gaussian) [86] adjust spectral components, while techniques like CLAHE (contrast-limited adaptive histogram equalization) [87] and homomorphic filtering (illumination/reflectance correction via logarithm and Fourier) [88] enhance contrast in complex images. Advanced methods like Frangi filter (detection of tubular structures via Hessian eigenvalues)

[89] and blur masking (sharpening by subtracting a smoothed version) [61] optimize applications in computer vision, medical processing, and image restoration.

Artifact removal involves eliminating unnecessary or irrelevant data from the image, such as watermarks, text, or unwanted artifacts.

Binarization consists of converting an image into a binary image where pixels are classified as black or white. This is useful for object segmentation applications.

Segmentation refers to separating objects of interest from the background or other objects in the image. Segmentation is useful when specific objects need to be identified and analyzed. It divides an image into homogeneous regions based on attributes such as discontinuities (edges) or similarities (intensity, texture) to facilitate interpretation [61, 62, 90]. Edge-based methods use operators like Sobel, Prewitt, Canny, and Scharr to identify abrupt intensity transitions [62, 64], while histogram-based approaches apply global thresholding (Otsu, variance maximization; Kapur, maximum entropy) or generalized entropies (Tsallis, Rényi) to separate objects and background [90–92]. In environments with variable lighting, adaptive thresholds (Niblack, Sauvola, Bernsen) dynamically adjust thresholds based on local statistics (mean, standard deviation, range), optimizing segmentation in subregions [82, 93, 94]. These techniques, fundamental in computer vision, enable applications such as Optical Character Recognition (OCR), medical diagnosis, and texture analysis, balancing accuracy and computational efficiency.

2.3 Pretraining

A pretrained model is a deep learning model trained on large datasets for general tasks like image or text recognition, ready to be adapted for various uses.

Regarding transfer learning, large pretrained models exhibit a remarkable ability to generalize and transfer knowledge across different datasets and tasks, making them highly effective tools for a wide range of applications. Transfer learning fine-tunes a pretrained model for a specific task or smaller dataset [76, 95]. Instead of training the model from scratch, the weights and parameters of the pretrained model are adjusted to adapt it to the new task. This is especially useful when the new dataset is small or when there is insufficient time to train a model from scratch, since the pretrained model has already captured useful general features. Transfer learning techniques can be implemented in various ways. One option is to use only the base architecture as a template, without incorporating the previously learned weights, which corresponds to the feature extraction approach. Another option is to leverage both the structure and pretrained weights of the original model, known as direct utilization. In this latter case, different strategies are possible: adjusting only certain layers, such as the output layer, to adapt them to the new domain; performing full fine-tuning by modifying all the model parameters during retraining; or expanding the architecture by adding new specialized layers at the end, which is associated with the multitask learning approach. Transfer learning methodologies are categorized into four approaches: feature extraction, fine-tuning, domain adaptation, and multitask learning [76], being a successful strategy in domains like computer vision and natural language processing [95].

Feature extraction uses a pretrained convolutional neural network's (CNN's) learned features as input for a new dataset or model without fine-tuning the entire model [16, 96].

In some cases, pretrained models are also used directly without fine-tuning, especially if the tasks are general ones for which the model was originally trained, such

as object recognition in images or natural language processing. We can potentially improve the estimation of a local model by relying on the global model without using it directly; this idea can be called soft transfer learning, as opposed to hard transfer learning, where the global model is used directly without adjustments [97].

Fine-tuning focuses on adapting the upper layers while preserving the generic features learned in the initial layers from the source task, thereby reducing overfitting [76]. Fine-tuning a model trained on a source domain helps reduce the amount of data and training time required in a new target domain by avoiding training from scratch and decreasing data collection costs [95], significantly improving classification accuracy [96]. Fine-tuning adjusts a pretrained model's weights for a task, improving performance but needing more computation. Feature extraction uses pretrained weights without changes, saving resources for small data [76, 96]; ablation studies help determine which layers should be fine-tuned depending on the case.

Domain adaptation reduces differences between source and target domains so models work well across them [95, 98]. Rule adaptation reuses general knowledge from previous tasks to facilitate new ones, while domain adaptation aligns datasets with similar distributions to enable the use of a single classifier across different contexts [96].

Multitask learning trains a model on multiple related tasks simultaneously, leading to better generalization and performance by sharing representations—particularly useful when data per task are limited [96]. However, it faces challenges, such as overfitting, catastrophic forgetting, and task interference; therefore, in NLP, separate models are often preferred for each task [99]. Multitask transfer learning improves performance and efficiency by sharing representations across related tasks in real-world applications.

2.4 Training

In this phase, a model is taught to recognize patterns, features, or classes in images.

The training dataset consists of labeled images, where each label indicates the class or category to which the image belongs. When training with reduced training datasets, Lu and Li [100] proposes a method combining data augmentation with transfer learning, allowing convolutional neural networks (CNNs) to avoid the overfitting problem and achieve excellent classification accuracy.

The choice of an appropriate model, such as a DCNN, is important. The model has layers of neurons that will be trained to learn patterns in images. Model selection involves balancing approximation and estimation errors, adding complexity due to the transfer distance between source and target distributions, which varies by hypothesis class [101].

Parameter initialization of the model with random values [102], which can be adjusted during training so that the model can make accurate predictions.

In optimization, a loss function measures the gap between model predictions and true labels in training.

The goal is to minimize loss by iteratively adjusting model parameters using optimization algorithms like gradient descent. In transfer learning, when pre-trained models are reused and adapted to a new classification task, the cross-entropy loss function remains one of the most commonly used, especially in multi-class classification tasks. This function combines LogSoftmax (logarithm of the softmax function) and NLLLoss (negative-log-likelihood loss), just like in training from scratch [21].

The training process involves iterations through the training set. In each iteration, the model processes one or several images, makes predictions, and adjusts its parameters based on the calculated loss. This process repeats many times (epochs) until the model converges or a predefined stopping criterion is met.

When necessary, hyperparameters—learning rate, batch size, or architecture—can be adjusted [101].

2.5 Validation

It is performed after the training of the model and aims to evaluate its performance and generalization capability on unseen data.

The validation dataset is an independent set containing images not used during training; it adjusts hyperparameters to prevent overfitting. These images are also labeled with their corresponding categories, and the test set evaluates the final model performance.

Predictions on the validation set are made using the trained model. Each image is passed through the model, which generates class predictions or continuous values, depending on the specific task.

The analysis and interpretation of predictions aims to explain the model's decisions, detect failures, and validate that it uses relevant features through interpretability tools. Some interpretability techniques include: Grad-CAM (Gradient-weighted Class Activation Mapping) [103], saliency maps, LIME (Local Interpretable Model-Agnostic Explanations) [104], SHAP (SHapley Additive Explanations) [105], feature importance in models, such as Random Forest or XGBoost, PDP (Partial Dependence Plot), and integrated gradients.

Performance evaluation consists of comparing the model's output with the true labels in the validation set. The model's performance is assessed using metrics appropriate for the task, such as accuracy, precision, recall, F1-score, among others [8], and these reflect the quality of the model's predictions on unseen data.

Model hyperparameters may need adjustment if the performance on the validation set is unsatisfactory. Such adjustments can include changes in learning rate, model architecture, batch size, among others.

Validation helps detect and prevent overfitting, where the model fits training data too closely and fails on new data. Ahmed [1], Dawson et al. [7], and Alhanaee et al. [22] use image augmentation to expand training data, improving generalization and reducing overfitting. Selectively fine-tuning the top layers while sharing the initial layers with the source task reduces overfitting and favors better feature reuse [76].

In some cases, cross-validation is used, a technique that involves splitting the dataset into multiple partitions (folds) [1, 6, 7, 97] and performing multiple rounds of training and validation. This provides a more robust estimate of the model's performance. Cross-validation allows evaluating the statistical significance between models, eliminating dependence on chance [6].

The model is saved once it achieves satisfactory performance, for later use in the recognition stage.

2.6 Recognition

It involves using a pretrained model to identify objects, patterns, or features in images, prerecorded videos, or real-time videos.

First, there must be a pretrained model; this model has learned to identify patterns and features in images, thanks to the training phase. The trained model can be used directly for the target task or a new task [98]. The model's performance may degrade due to changes in the network, traffic patterns, usage behavior, or the incorporation of new devices [95].

The input data provide images for the model to recognize. These images can be captured in real-time through a camera or loaded from a file.

Before feeding the images to the model, preprocessing may be necessary, which includes normalization [55], resizing [54] and in some cases segmentation [61, 62, 90] or noise removal [60, 61, 67, 84–88].

The model processes the images and generates predictions; in most cases, it provides a class label or a probability associated with each class, depending on whether you are performing image classification or an object detection task.

Predictions may require post-processing depending on the application; this could include removing low-confidence predictions, grouping similar predictions, or generating additional information from the predictions, such as bounding boxes in the case of object detection.

The results are presented through the predictions in a meaningful way for the end user; this consists of displaying class labels, locations of detected objects in the image, or any other relevant information.

Automated decisions can be made or information provided based on the model's predictions to a human user for informed decision-making.

Regarding iteration, the image recognition system can operate continuously, providing real-time predictions and improving over time as more data are collected and the model is recalibrated in real-world applications.

3. Conclusion

The comprehensive methodology for the development of image/object recognition systems based on deep learning is structured into six sequential phases: data collection, preprocessing, pretraining, training, validation, and recognition. By leveraging transfer learning with pretrained architectures, computational resources are optimized and performance is improved in scenarios with limited data, as is the case in specialized domains, while prioritizing ethical and legal aspects in the acquisition, labeling, and anonymization of datasets. The preprocessing phase integrates advanced visual manipulation techniques to ensure quality and diversity in the model inputs, while training and validation employ optimization algorithms, evaluation metrics, and strategies against overfitting. Finally, the recognition phase applies the model in real-world contexts, demonstrating its capability to classify, detect objects, or segment images with high accuracy. This systemic approach, supported by technical examples and practical considerations, offers a robust guide for researchers and professionals, facilitating the implementation of efficient and scalable solutions in computer vision, with cross-disciplinary applications in medicine, security, industry, and more, thereby contributing to the advancement of adaptive and ethical artificial intelligence systems.

Conflict of interest

The authors declare that they have no conflict of interest.

Acronyms

1D	one-dimensional
2D	two-dimensional
3D	three-dimensional
AI	artificial intelligence
CNN	convolutional neural network
COCO	common objects in context
CV	computer vision
DCNN	deep convolutional neural network
GAN	generative adversarial networks
ML	machine learning
MTCNN	multi-task cascaded convolutional neural networks
NLP	natural language processing
TL	transfer learning
YOLO	you only look once

Author details

Lenin Huayta Flores^{1*}, Lucio Ticona Carrizales², Lenin Samir Huayta Poma¹
and Betty Campos Segales¹

¹ Universidad Nacional del Altiplano, Puno, Peru

² Universidad Nacional de Juliaca, Juliaca, Peru

*Address all correspondence to: lhuayta@unap.edu.pe

IntechOpen

© 2025 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

References

- [1] Ahmed MR. Leveraging convolutional neural network and transfer learning for cotton plant and leaf disease recognition. *International Journal of Image, Graphics and Signal*. 2021;**10**:47
- [2] Fayaz S, Ahmad Shah SZ, Ud din NM, et al. Advancements in data augmentation and transfer learning: A comprehensive survey to address data scarcity challenges. *Recent Advances in Computer Science and Communications (Formerly Recent Patents Comput Sci)*. 2024;**17**:14-35
- [3] Teixeira AC, Ribeiro J, Morais R, et al. A systematic review on automatic insect detection using deep learning. *Agriculture*. 2023;**13**:713
- [4] Elsen I, Ferrein A, Schiffer S. Anomaly Detection in Metal-Textile Industries. Epub ahead of print 2025. DOI: 10.5772/intechopen.1008411
- [5] Mutawa AM, Alnajdi S, Sruthi S. Transfer learning for diabetic retinopathy detection: A study of dataset combination and model performance. *Applied Sciences*. 2023;**13**:5685
- [6] Luna-Jiménez C, Griol D, Callejas Z, et al. Multimodal emotion recognition on RAVDESS dataset using transfer learning. *Sensors*. 2021;**21**:7665
- [7] Dawson HL, Dubrule O, John CM. Impact of dataset size and convolutional neural network architecture on transfer learning for carbonate rock classification. *Computational Geosciences*. 2023;**171**:105284
- [8] Khan A, Hassan B, Khan S, et al. DeepFire: A novel dataset and deep transfer learning benchmark for forest fire detection. *Mobile Information Systems*. 2022;**2022**:5358359
- [9] Mei X, Liu Z, Robson PM, et al. RadImageNet: An open radiologic deep learning research dataset for effective transfer learning. *Radiology: Artificial Intelligence*. 2022;**4**:e210315
- [10] Lakshmanan V, Görner M, Gillard R. *Practical Machine Learning for Computer Vision*. Sebastopol, CA: O'Reilly Media, Inc.; 2021
- [11] Chapman A, Simperl E, Koesten L, et al. Dataset search: A survey. *VLDB Journal*. 2020;**29**:251-272
- [12] Paton NW, Chen J, Wu Z. Dataset discovery and exploration: A survey. *ACM Computing Surveys*. 2023;**56**:1-37
- [13] Funabashi S, Yan G, Hongyi F, et al. Tactile transfer learning and object recognition with a multifingered hand using morphology specific convolutional neural networks. *IEEE Transactions on Neural Networks and Learning Systems*. 2022;**35**(6):7587-7601. DOI: 10.1109/TNNLS.2022.3215723
- [14] Moghimi MK, Mohanna F. Reliable object recognition using deep transfer learning for marine transportation systems with underwater surveillance. *IEEE Transactions on Intelligent Transportation Systems*. 2022;**24**:2515-2524
- [15] Fan Z, Shi L, Liu Q, et al. Discriminative fisher embedding dictionary transfer learning for object recognition. *IEEE Transactions on Neural Networks and Learning Systems*. 2021;**34**:64-78
- [16] Murali PK, Wang C, Lee D, et al. Deep active cross-modal visuo-tactile

transfer learning for robotic object recognition. *IEEE Robotics and Automation Letters*. 2022;7:9557-9564

[17] Kabakus AT, Erdogmus P. An experimental comparison of the widely used pre-trained deep neural networks for image classification tasks towards revealing the promise of transfer-learning. *Concurrency and Computation: Practice and Experience*. 2022;34:e7216

[18] Bai J. Research and application of deep learning based on transfer learning in image classification tasks. In: 2024 IEEE 4th International Conference on Power, Electronics and Computer Applications (ICPECA). Piscataway, NJ: IEEE; 2024. pp. 1292-1297

[19] Narayan K, Nair NG, Xu J, et al. Petalface: Parameter efficient transfer learning for low-resolution face recognition. *arXiv Prepr arXiv241207771*. 2025:804-814. DOI: 10.1109/WACV61041.2025.00088. Epub ahead of print 2024

[20] Kute R, Vyas V, Anuse A. Transfer learning for face recognition using fingerprint biometrics. *Journal of King Saud University, Science*. 2021;33(5):1-6. DOI: 10.1016/j.jksues.2021.07.011

[21] Chambino LL, Silva JS, Bernardino A. Multispectral face recognition using transfer learning with adaptation of domain specific units. *Sensors*. 2021;21:4520

[22] Alhanaee K, Alhammadi M, Almenhali N, et al. Face recognition smart attendance system using deep transfer learning. *Procedia Computer Science*. 2021;192:4093-4102

[23] Sreekala K, Cyril CPD, Neelakandan S, et al. Capsule network-based deep transfer learning model for face recognition. *Wireless*

Communications and Mobile Computing. 2022;2022:2086613

[24] Elaggoune H, Belahcene M, Bourennane S. Hybrid descriptor and optimized CNN with transfer learning for face recognition. *Multimedia Tools and Applications*. 2022;81:9403-9427

[25] Kwak N, Kim D. Study on masked face detection and recognition using transfer learning. *International Journal of Advanced Culture Technology*. 2022;10:294-301

[26] Yi R, Tian H, Gu Z, et al. Towards artistic image aesthetics assessment: A large-scale dataset and a new method. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ, USA: IEEE – Institute of Electrical and Electronics Engineers; 2023. pp. 22388-22397

[27] Mondal T, Mendoza R, Drumetz L. Physics informed and data driven simulation of underwater images via residual learning. *arXiv Prepr arXiv240205281*. 2024. DOI: 10.48550/arXiv.2402.05281. Epub ahead of print 2024

[28] Mansourifar H, Shi W. Deep synthetic minority over-sampling technique. *arXiv Prepr arXiv200309788*. 2020. DOI: 10.48550/arXiv.2003.09788. Epub ahead of print 2020

[29] Mekonnen KA. Balanced face dataset: Guiding stylegan to generate labeled synthetic face image dataset for underrepresented group. *arXiv Prepr arXiv230803495*. 2023. DOI: 10.48550/arXiv.2308.03495. Epub ahead of print 2023

[30] Katare D, Noguero DS, Park S, et al. Analyzing and mitigating bias for vulnerable classes: Towards balanced

representation in dataset. arXiv Prepr arXiv:2401.10397. 2024. DOI: 10.48550/arXiv.2401.10397. Epub ahead of print 2024

[31] Whang SE, Lee J-G. Data collection and quality challenges for deep learning. Proceedings of the VLDB Endowment. 2020;**13**:3429-3432

[32] Roh Y, Heo G, Whang SE. A survey on data collection for machine learning: A big data-AI integration perspective. IEEE Transactions on Knowledge and Data Engineering. 2019;**33**:1328-1347

[33] Kumar V. Data Collection. 2020. Available from: <https://mgcub.ac.in/pdf/material/20200429021914369e0bded0.pdf>

[34] Esan DO, Owolawi PA, Tu C. Generative adversarial networks: Applications, challenges, and open issues. In: Deep Learning-Recent Findings and Research. London, UK: IntechOpen; 2023. Epub ahead of print 2023. DOI: 10.5772/intechopen.113098

[35] Acharya S, Paul P. The future of medical imaging: Ensuring ethical and legal compliance. In: Smart Medical Imaging for Diagnosis and Treatment Planning. Boca Raton, FL: Chapman and Hall/CRC; 2025. pp. 221-243

[36] Park H, Oh H, Choi JK. A consent-based privacy-compliant personal data-sharing system. IEEE Access. 2023;**11**:95912-95927

[37] Tiwari S, Raja R, Wadawadagi RS, et al. Emerging biometric modalities and integration challenges. In: Online Identity-an Essential Guide. London, UK: IntechOpen; 2024. Epub ahead of print 2024. DOI: 10.5772/intechopen.1003148

[38] Majeed A, Khan S, Hwang SO. Group privacy: An underrated but worth

studying research problem in the era of artificial intelligence and big data. Electronics. 2022;**11**:1449

[39] Siminiuc S. Advanced Methods of Data Anonymization in Medical Research. 2025. Available from: <http://repository.utm.md/handle/5014/29236>

[40] Gopireddy RR. Data anonymization techniques: Ensuring privacy in big data analytics. European Journal of Advances in Engineering and Technology. 2020;**7**:68-74

[41] Zhao M, Chen J. Sequential classification of hyperspectral images. In: Hyperspectral Imaging in Agriculture, Food and Environment. London, UK: IntechOpen; 2018. Epub ahead of print 2018. DOI: 10.5772/intechopen.73160

[42] Chernov Y. Data quality measurement based on domain-specific information. In: Data Integrity and Data Governance. London, UK: IntechOpen; 2022. Epub ahead of print 2022. DOI: 10.5772/intechopen.106939

[43] Vancauwenbergh S. Data quality management. Science Recent Advances. 2019;**1**:15

[44] Christiaanse R. Quality 4.0: Data quality and integrity—a computational approach. In: Six Sigma and Quality Management. London, UK: IntechOpen; 2022. Epub ahead of print 2022. DOI: 10.5772/intechopen.108213

[45] Baral P, Yang N, Weng N. IoT device identification using device fingerprint and deep learning. In: Deep Learning and Reinforcement Learning. London, UK: IntechOpen; 2023. Epub ahead of print 2023. DOI: 10.5772/intechopen.111554

[46] Rangel JC, Cruz E, Cazorla M. Automatic understanding and mapping of regions in cities using Google

street view images. *Applied Sciences*. 2022;12:2971

[47] Rajput AS, Rajput DAS, Shukla DS, et al. Cutting-Edge Artificial Intelligence in Agriculture: Deepfusionnet Model for Tomato Leaf Disease Classification. 2024. Available SSRN 4942525. DOI: 10.2139/ssrn.4942525

[48] Parameswaran AG, Garcia-Molina H, Park H, et al. Crowdscreen: Algorithms for filtering data with humans. In: *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*. 2012. pp. 361-372

[49] Saini M, Susan S. Data augmentation of minority class with transfer learning for classification of imbalanced breast cancer dataset using inception-V3. In: *Pattern Recognition and Image Analysis: 9th Iberian Conference, IbPRIA 2019, Madrid, Spain, July 1-4, 2019, Proceedings, Part I 9*. Cham: Springer; 2019. pp. 409-420

[50] Damaševičius R. Introductory chapter: Current state and achievements of data augmentation. *Deep Learn Find Research*. 2024;1-10. DOI: 10.5772/intechopen.112284. Epub ahead of print

[51] Jiang W, Zhang Y, Zheng S, et al. Data augmentation in human-centric vision. *Vicinagearth*. 2024;1:8

[52] Yang X, Liang W, Zou J. Navigating dataset documentation in ML: A large-scale analysis of dataset cards on hugging face. In: *NeurIPS 2023 Workshop on Regulatable ML*. 2023. Epub ahead of print 2023. DOI: 10.48550/arXiv.2401.13822

[53] Pushkarna M, Zaldivar A, Kjartansson O. Data cards: Purposeful and transparent dataset documentation for responsible AI. In: *Proceedings of the 2022 ACM Conference on Fairness,*

Accountability, and Transparency. New York, NY, USA: Association for Computing Machinery (ACM); 2022. pp. 1776-1826

[54] González-Acuña RG, Chaparro-Romo HA, Melendez-Montoya I. *Optics and Artificial Vision*. Bristol: IOP Publishing; 2021. Available from: <https://iopscience.iop.org/book/mono/978-0-7503-3707-6>

[55] Wankhede D, Athanikar A, Borude Y, et al. An experimental analysis of machine learning model in the context of sales forecasting:-A Walmart dataset as a case study. *GRENZE International Journal of Engineering and Technology*. 2024;10:1048-1057

[56] Rajesh PJ, Balambica V, Achudhan M. Automated gear inspection using image processing and machine learning techniques. In: *2024 4th International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*. Piscataway, NJ: IEEE; 2024. pp. 1643-1648

[57] Parkin A. *Computing Colour Image Processing*. Cham: Springer; 2018. Epub ahead of print 2018. DOI: 10.1007/978-3-319-74076-8

[58] Chung BWC. *Pro Processing for Images and Computer Vision with OpenCV*. 2017. DOI: 10.1007/978-14842-2775-6. Epub ahead of print 2017

[59] Demagad K, Oliver A, Oostendorp N, et al. *Practical Computer Vision with SimpleCV: The Simple Way to Make Technology See*. Sebastopol, CA: O'Reilly Media, Inc.; 2012

[60] Distant A, Distant C. *Handbook of Image Processing and Computer Vision: Volume 2: From Image to Pattern*. 2020. DOI: 10.1007/978-3-030-42374-2. Epub ahead of print 2020

- [61] Solomon C, Breckon T. Fundamentals of Digital Image Processing: A Practical Approach with Examples in Matlab. Chichester: John Wiley & Sons; 2011
- [62] Thanki RM, Kothari AM. Digital Image Processing Using SCILAB. Cham: Springer; 2019. Epub ahead of print 2019. DOI: 10.1007/978-3-319-89533-8
- [63] McAndrew A. A Computational Introduction to Digital Image Processing. Boca Raton, FL: CRC Press Taylor and Francis Group; 2016
- [64] Mordvintsev A, Abid K. Opencv-Python Tutorials Documentation. San José, CA, USA: OpenCV Project; 2016
- [65] Rafael CG, Richard EW. Digital Image Processing. 4th ed. New York, NY, USA: Pearson Education; 2017
- [66] Kinser JM. Image Operators: Image Processing in Python. Boca Raton, FL: CRC Press; 2018. Epub ahead of print 2018. DOI: 10.1201/9780429451188
- [67] Chityala R, Pudipeddi S. Image Processing and Acquisition Using Python. Boca Raton, FL: Chapman and Hall/CRC; 2020. Epub ahead of print 2020. DOI: 10.1201/9780429243370
- [68] Asad H, Shrimali VR, Singh N. The computer vision workshop. In: The Computer Vision Workshop: Develop the Skills you Need to Use Computer Vision Algorithms in your Own Artificial Intelligence Projects. Birmingham: Packt Publishing; 2020
- [69] Pajankar A. Python 3 Image Processing: Learn Image Processing with Python 3, NumPy, Matplotlib, and Scikit-Image. New Delhi: BPB Publications; 2019
- [70] Tyagi V. Understanding Digital Image Processing. Boca Raton, FL: CRC Press; 2018. Epub ahead of print 2018. DOI: 10.1201/9781315123905
- [71] Dawson-Howe K. A Practical Introduction to Computer Vision with Opencv. Chichester: John Wiley and Sons Ltd.; 2014
- [72] Szeliski R. Computer Vision: Algorithms and Applications. Cham: Springer Nature; 2022. Epub ahead of print 2022. DOI: 10.1007/978-3-030-34375-9
- [73] Glasbey CA, Mardia KV. A review of image-warping methods. Journal of Applied Statistics. 1998;25:155-171
- [74] Rosebrock A. Practical python and OpenCV. 2nd ed. Pyimagesearch; 2016. 160 p
- [75] Kumar DRV. Data Collection. Saarbrücken, Germany: LAP LAMBERT Academic Publishing; 2024
- [76] Majeed APPA, Engelbrecht A. Transfer learning: Leveraging the capability of pre-trained models across different domains. In: BoD–Books on Demand. Norderstedt, Germany: BoD – Books on Demand; 2025. DOI: 10.5772/intechopen.114788. Epub ahead of print 2025
- [77] Pinheiro Cinelli L, Araújo Marins M, Barros da Silva EA, et al. Variational autoencoder. In: Variational Methods for Machine Learning with Applications to Deep Networks. Cham: Springer; 2021. pp. 111-149
- [78] Zhang H, Cisse M, Dauphin YN, et al. Mixup: Beyond empirical risk minimization. arXiv Prepr arXiv171009412. 2018:1-13. DOI: 10.48550/arXiv.1710.09412. Epub ahead of print 2017
- [79] Yun S, Han D, Oh SJ, et al. Cutmix: Regularization strategy to train strong

- classifiers with localizable features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ, USA: IEEE – Institute of Electrical and Electronics Engineers; 2019. pp. 6023-6032
- [80] Lai YL, Ang TF, Bhatti UA, et al. Color correction methods for underwater image enhancement: A systematic literature review. *PLoS One*. 2025;20. DOI: 10.1371/journal.pone.0317306. Epub ahead of print 2025
- [81] Guerra JP, Cuevas F. Application of digital image processing techniques for agriculture: A review. *Digital Image Processing: Advanced Technologies and Applications*. 2024. DOI: 10.5772/intechopen.1004767. Epub ahead of print 2024
- [82] Burger W, Burge MJ. Principles of digital image processing: Fundamental techniques. In: Springer Science and Business Media. 2010. Epub ahead of print 2010. DOI: 10.1007/978-1-84800-191-6
- [83] Gavet Y, Debayle J. Image Processing Tutorials with Python®. Lyon, France: Spartacus-Idh; 2019. Available from: <https://hal-emse.ccsd.cnrs.fr/emse-02469242v1>
- [84] Sánchez FJ. Medición y Análisis de las Variaciones en el Nivel de un Modelo Físico Empleando Imágenes. Azcapotzalco, Mexico: UAM; 2009
- [85] Rebaza JV. Detección de bordes mediante el algoritmo de Canny. 2007;4:1-8
- [86] Zhang Y-J. A Selection of Image Processing Techniques: From Fundamentals to Research Front. Boca Raton, FL: CRC Press; 2022. Epub ahead of print 2022. DOI: 10.1201/9781003241416
- [87] Zuiderveld K. Contrast limited adaptive histogram equalization. *Graph Gems*. 1994;IV:474-485
- [88] Distant A, Distant C. Handbook of Image Processing and Computer Vision: Volume 1: From Energy to Image. Cham: Springer; 2020. Epub ahead of print 2020. DOI: 10.1007/978-3-030-38148-6
- [89] Frangi AF, Niessen WJ, Vincken KL, et al. Multiscale vessel enhancement filtering. In: Medical Image Computing and Computer-Assisted Intervention — MICCAI'98. Berlin, Heidelberg: Springer; 2006. pp. 130-137
- [90] Cervantes H, Miguel S. Algoritmos metaheurísticos para la segmentación de imágenes. Madrid: Universidad Complutense de Madrid; 2019. Available from: <https://hdl.handle.net/20.500.14352/17180>
- [91] De Albuquerque MP, Esquef IA, Mello ARG. Image thresholding using Tsallis entropy. *Pattern Recognition Letters*. 2004;25:1059-1065
- [92] Oliva D, Cuevas E. Advances and Applications of Optimised Algorithms in Image Processing. Cham: Springer; 2017. Epub ahead of print 2017. DOI: 10.1007/978-3-319-48550-8
- [93] Sauvola J, Pietikäinen M. Adaptive document image binarization. *Pattern Recognition*. 2000;33:225-236
- [94] Bernse J. Dynamic thresholding of grey-level images. In: Proceedings of the 8th International Conference on Pattern Recognition. Piscataway, NJ, USA: IEEE – Institute of Electrical and Electronics Engineers; 1986. pp. 1251-1255
- [95] Vandikas K, Moradi F, Larsson H, et al. Transfer Learning and Domain Adaptation in Telecommunications.

London, UK: IntechOpen; 2024.
DOI: 10.5772/intechopen.114932. Epub
ahead of print 2024

[96] Wei X. Transfer Learning for Non-Invasive BCI EEG Brainwave Decoding. London, UK: IntechOpen; 2024.
DOI: 10.5772/intechopen.115124. Epub
ahead of print 2024

[97] Hellum O, Pedersen LH, Rønn-Nielsen A. How global is predictability? The power of financial transfer learning. *Power Finance Transfer Learning*. 2023:66.
DOI: 10.2139/ssrn.4620157. Epub ahead
of print 2023

[98] Xu W, He J, Shu Y. Transfer learning and deep domain adaptation. *Advanced Applied Deep Learning*. 2020;**45**:600-620. DOI: 10.5772/intechopen.94072.
Epub ahead of print 2020

[99] Pilault J, Elhattami A, Pal C. Conditionally adaptive multi-task learning: Improving transfer learning in NLP using fewer parameters and less data. *arXiv Prepr arXiv200909139*. 2020.
DOI: 10.48550/arXiv.2009.09139. Epub
ahead of print 2020

[100] Lu C, Li W. Ship classification in high-resolution SAR images via transfer learning with small training dataset. *Sensors*. 2018;**19**:63

[101] Hanneke S, Kpotufe S, Mahdaviyeh Y. Limits of model selection under transfer learning. In: *The Thirty Sixth Annual Conference on Learning Theory*. Cambridge, MA, USA: PMLR; 2023. pp. 5781-5812

[102] Zhong X, Ban H. Pre-trained network-based transfer learning: A small-sample machine learning approach to nuclear power plant classification problem. *Annals of Nuclear Energy*. 2022;**175**:109201

[103] Selvaraju RR, Das A, Vedantam R, et al. Grad-CAM: Gradient-weighted class activation mapping. *Arxiv*. 2020;**128**:3364-3381. DOI: 10.1007/s11263-019-01228-7. Epub ahead of print 2016

[104] Ribeiro MT, Singh S, Guestrin C. ‘Why Should I Trust You?’ Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016. pp. 1135-1144

[105] Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*; 30

Transfer Learning with CLIP Model for Multitasking Workload Prediction in Mixed Reality Environment

Safanah Abbas, Heejin Jeong and David He

Abstract

Multitasking has become an important part of our modern life, especially when interacting with new technologies such as mixed reality (MR). Predicting human workload in MR environments is crucial for optimizing the user experience. This study adopts transfer learning for workload prediction and employs a generative model to expand the original collected dataset. A dataset has been collected from a multitasking experiment involving 36 participants, combining a real-world block-matching task and a virtual N-back task, with workload ratings recorded using the NASA task load index (NASA-TLX). The dataset was augmented by using the generative adversarial networks (GANs) model, resulting in 5000 synthesized observations. A hybrid workload prediction model integrating the contrastive language-image pre-training (CLIP) model, adapted from computer vision applications, achieved superior results with a root mean square error (RMSE) of 0.96, significantly outperforming traditional regression models. These findings highlight the potential of transfer learning in workload prediction, paving the way for more efficient human performance modeling in such hybrid environments.

Keywords: workload, prediction, CLIP, mixed reality, pre-trained models, transfer learning, NASA-TLX

1. Introduction

The nature of multitasking has increased in our daily lives because of the growing demands of work. Therefore, multitasking has proven significant in dealing with our complex environments [1]. With the increasing advancements in technology, more multitasking opportunities are opening up. For example, the new computing platforms, like virtual or augmented realities, will significantly shape our interactions for the next 50 years, similar to how personal computers did in the past [2]. This shift has transformed traditional work contexts, which allow humans to engage in a hybrid

world, in other words, mixed reality (MR). This mixed environment combines elements from the actual and digital realms [3, 4]. This rapid development of technology creates opportunities to further research the effect of multitasking between our real world and digital content.

This is very relevant given the increasing applications based on immersive technologies. For instance, a 2018 Toshiba survey revealed that 82% of enterprises foresaw integrating mixed reality into their business strategies [5]. Forrester Research's 2016 report estimated that by the current year 2025, more than 14 million workers will be using smart glasses [6]. The release of the Apple Vision Pro in February 2024 supports this trend. While multitasking by itself is not new, and has been implemented by humans for centuries to increase their efficiency, evolving technologies highly underscore the need to investigate its effect on human behavior. Therefore, there is an opportunity to bridge the gap between research and real-world applications in this matter.

Despite the advantages of multitasking in the hybrid or mixed world, such as providing maximum user interaction, potential challenges affecting users' well-being can occur [7]. This happens because of the various human limitations in physical or cognitive abilities, such as fatigue, reduced focus, or increased workload. Workload refers to the total mental, physical, or combined effort required from an individual to complete a specific task [8, 9]. Evaluating human workload, which is one of the most critical human performance measures, is crucial in human-computer interactions (HCI). A common approach to do that is to use the subjective NASA-TLX multi-dimensional tool that is used to assess six dimensions: mental demand (MD), physical demand (PD), temporal demand (TD), performance (P), effort (E), and frustration (F) [10]. However, collecting such large amounts of data about human behavior is very challenging due to time constraints, effort, cost, etc. Therefore, the urgent need to create innovative solutions to resolve this challenge has become more critical, as understanding such data can improve user experience, especially as technology advances and multitasking demands grow in parallel.

Machine learning modeling has earned more attention in the past few years in addressing challenges related to limited data availability or making predictions. To obtain more training samples, the generative adversarial network (GAN) model, for example, generates artificial samples that capture the same distribution of the original data samples. Instead of applying traditional models for predictions, transfer learning becomes useful. It uses pre-trained models to transfer knowledge from one domain to another similar domain; thus, it reduces training time and enables fast adaptation to new tasks [11]. Incorporating these advanced solutions into human workload prediction in a mixed reality environment can optimize the process, like reducing the need for manual data collection using NASA-TLX, which can be time-consuming and costly. The objective of this chapter is to propose advanced predictive models utilizing a transfer learning method to predict human workload in mixed-reality multitasking.

2. Literature review

2.1 Workload assessment in multitasking using the NASA-TLX

Describing the relationship between human workload and multitasking is complex. Besides the differences in cognitive capabilities between different individuals,

performance can also be affected by the characteristics of the tasks at hand [12]. Task difficulty as a task characteristic is one of the critical contributors impacting human performance [13]. Moreover, the context in which multitasking occurs can influence the results. For example, a quiet environment may result in better multitasking than a distracted environment. In a complex environment like multitasking between physical and digital tasks, balancing the imposed task demand and individual differences is essential. User characteristics, on the other hand, like age, familiarity, and experience levels in working with immersive realities, can also impact the resulting human workload [14].

Several studies in the literature evaluated the effect of multitasking on the human workload. For instance, see Ref. [15] for indicating a higher perceived workload when engaging in both a tracking task and a search task based on memory concurrently, compared to performing each task by itself, where workload, even the lowest reported multitasking load, surpassed the highest reported single-task load. In Ref. [16], the authors studied the effect of mental workload on the performance of an assembly task, which involved building a Lego car set and performing a secondary mental task at the same time. The results indicated impaired task performance and the resulting self-reported workload *via* the NASA-TLX was higher.

Regarding engagement in a digital environment, some studies have explored the effect of workload by comparing human performance in the real (physical) world with that in the virtual (digital) world or multitasking between both. For example, the authors in Ref. [17] compared the workload for logistics workers when using augmented reality to the traditional paper-based method when filling a pick list. The mental and temporal demand dimensions in the NASA-TLX were found to be significant, where more mental demand was found using the paper method than the augmented reality device. However, the other dimensions did not yield any significant difference. Similarly, Rebol et al. [18] evaluated the workload in a mixed reality communication system for remote assistance during CPR emergencies in comparison to the traditional method of video call. The resulting mental workload from the NASA-TLX yielded significant results, where experts experienced more workload using the video call method, and first responders experienced similar workload levels in both methods. On the contrary, Brunzini et al. [19] assessed the workload using the NASA-TLX in a mixed-reality medical simulation experiment using head-mounted displays like the HoloLens and the Vox Gear Plus for medical students, in comparison to the traditional method of not involving any immersive technology found a higher workload from when engaging in the MR simulation, especially in physical demand, effort, and frustration dimensions. However, the authors suggested that the mixed reality simulation promotes effective learning when used more.

Other studies found no significant difference in the same context. Nenna et al. [20] explored the effect of subjective workload in a virtual teleoperation multitasking task. The results reported by the NASA-TLX in this experiment did not yield a significant difference. However, it was observed that there was a tendency for higher mental and temporal demand in participants who had a lower sense of presence in the virtual environment. In the same way, Szczurek et al. [21] found no significant difference in workload from the reported NASA-TLX between operators using a new 3D mixed reality interface compared with a conventional 2D operational interface used previously in many maintenance tasks. The advantage was that the new MR interface enhanced the operational quality and safety.

The diversity of results observed in the literature indicates that the effect of multitasking on workload in digital environments like MR is not straightforward. Future

research should explore this dynamic interplay between individual differences, task characteristics, and immersive technologies like mixed reality to develop a more precise workload assessment method.

2.2 Data augmentation using GAN

This generative model has been used in many studies to generate fake data that resembles real samples. However, its main application is related to images rather than tabular data. For example, see Ref. [22] for the proposed stacked GAN model to synthesize image data from text descriptions on two existing datasets for images: CUB, which contains bird images from 200 species, and the Oxford-102 dataset, which includes images of flowers from 102 categories. The GAN was split into two stages because of the complexity of generating fake images from text descriptions. In the first stage, the GAN creates a basic outline and color of the object from the text description, which produces low-quality images in terms of resolution. In the second stage, the GAN model takes the results from the first stage in addition to the text description, thus it generates better-resolution fake images. The evaluation of the proposed GAN model in this study used the inception score metric, which is used for image evaluation based on image quality and image diversity. The StackGAN model achieved a 28.47% improvement in the inception score on the CUB dataset and 20.30% on Oxford-102. The authors recommend that their proposed StackGAN generates better quality images, unlike the current text-to-image models.

In another example, see Ref. [23] for the application of a variant from the GAN model that uses a different loss function based on the Wasserstein distance for image classification. The prediction model for image classification used transfer learning with a combination of two pre-trained models for images, UNet and ResNet, which was called the UResNet model. The GAN model was implemented on three existing image-based datasets, such as the Omniglot dataset, which contains handwritten characters from various alphabets; the EMNIST dataset, which includes handwritten letters and digits from the English language; and the VGG-face, which contains images of different faces. The results showed that using the GAN model improved accuracy across all the datasets, leading to better performance than using only the data without augmentation. The model with GAN achieved up to 87.42% accuracy in classifying images for the Omniglot dataset, compared to 81.97% without the augmented data. For the EMNIST dataset, the accuracy was 84.80% with GAN-augmented data and 83.78% without. Lastly, the VGG-face dataset had an accuracy of 58.46% with GAN-augmented data and 57.99% without. Similarly, Waheed et al. [24] applied the GAN model to generate fake X-ray images of human chests to be fed into a convolutional neural network model to predict the detection of the COVID-19 virus. The augmentation model in this study is named the CovidGAN, and the Auxiliary classifier GAN version is utilized to generate fake images while also learning the category to which each image belongs. They found that the accuracy of classification with the augmented data was 95%, compared to 85% without the augmentation.

However, when dealing with tabular data, the GAN model usually struggles to achieve good results. For example, the authors in Ref. [25] experimented with generating tabular data using a dataset from intrusion detection systems. They needed a large amount of data to be trained effectively, and accessing such data is limited because of privacy. The authors implemented multiple GAN models to augment the tabular data, such as conditional GAN and TableGAN. They found that all GAN models suffer when modeling continuous features. The authors mentioned that while it is

possible to replace real with artificial data to solve machine learning problems, they found that their generated data can be distinguishable from the actual data in this application, in which the use of GAN for tabular data can still be further improved.

Other authors in Ref. [26] also focused on using the GAN model to generate fake tabular data. They adopted the training-by-sampling method using conditional GAN by adding a conditional vector to represent the classes of the categorical columns, as the tabular data consists of both continuous and categorical samples. The proposed GAN model is named CTAB-GAN+ and implemented on an industrial dataset. The loss function of the generator in this GAN model incorporated information loss, which is used to reduce the differences between the actual data statistics and the fake data, downstream loss, which measures the differences between the actual data and the predicted values, and the generator loss which measures the difference between the given conditions and the output classes. The generated samples from this model were evaluated based on statistical similarity with the original samples. The proposed model provided statistically similar augmented data to the original data. However, there was no significant improvement. Therefore, they highlighted the difficulty in generating tabular data through GAN to capture complex statistical distributions within the domain of dealing with tabular data.

2.3 Human workload prediction using traditional machine learning

The literature on workload prediction has been limited in situations like mixed reality multitasking, with many studies focusing more on physiological measures, such as heart rate and eye gaze, rather than on subjective measures like the NASA-TLX. One example of a study using physiological measures for predicting human workload was conducted by Skaramagkas et al. [27]. The authors estimated the mental load between four different experimental conditions, with time pressure, without time pressure, monotasking, and multitasking, by extracting gaze features compared to the mean scores from the NASA-TLX evaluation. The single task was a visual search task represented by a puzzle on the computer screen. In the multitasking condition, another task involving working memory, counting backward from 1000 by subtracting 4, was added. Binary classification trials of low and high mental workload and multiclass classification trials of low, medium, and high mental workload for support vector machines, decision trees, random forests, and others were used for prediction. The random forest model outperformed the others. In the binary classification using eye-tracking data, the mental workload prediction accuracy was 88% compared to 69% in the multiclass classification. However, when using NASA-TLX data, the results were reversed, with the prediction accuracy for the multiclass classification being 84 and 81% for the binary classification.

In another example, see Ref. [28] for predicting the mental workload based on electroencephalogram signals and the NASA-TLX in 72 visualization evaluation tasks using scatter plots, bar charts, line charts, and maps, all presented on a computer screen. The data from the electroencephalogram signals were used as inputs, whereas the weighted workload from the NASA-TLX was the output. Support vector machines, convolutional neural networks, deep neural networks, and long short-term memory models were compared for prediction. The prediction accuracies for the models were 26.22%, 82.67%, 88.57%, and 80.76%, respectively. They concluded that a deep neural network is the best model since it achieved the highest accuracy.

In the context of prediction in immersive environments, some studies have put some effort into that manner. One study in Ref. [29] used a machine learning

algorithm to predict cognitive workload in a virtual reality driving environment when transferring the learning to actual driving using physiological measures like eye gaze and heart rate. They used easy routes and difficult routes as difficulty levels. The participants were requested to memorize the routes given to them in a video before experimenting with the VR device. Multiple machine learning models were used in workload prediction, such as discriminant analysis, k-nearest neighbor, decision trees, neural network, and support vector machine. All these models were considered in classifying workload into single or multimodal features, named feature-level fusion. For the single feature classifier, the best accuracy was achieved for the support vector machine when predicting workload based on pupil dilation with an accuracy of 94.72%. For the feature-level infusion, the k-nearest neighbor outperformed the others with an accuracy of 96.56%, suggesting that multiple features provide better prediction in workload. In another study, in Ref. [30], the authors predicted the mental workload based on the electroencephalogram brain signals in a wood frame assembly task using AR technology for construction students under time pressure. The authors only used the long short-term memory model for forecasting and evaluated the model using the RMSE. The model achieved a mean RMSE score of 0.038. They concluded that the model could effectively predict changes in mental workload based on brain signals in assembly tasks.

2.4 Human workload prediction using transfer learning

Multiple pre-trained vision models exist, such as the VGG [31] and ResNet [32]. The VGG uses stacked layers of convolutional neural networks with a small filter (3x3), followed by fully connected layers. Meanwhile, the ResNet is more advanced and easier to optimize, because it introduces residual connections to address the vanishing gradient issue that is seen in the VGG model [33]. Another example is the contrastive language-image pre-training (CLIP), a neural network trained on an extensive dataset of images and their corresponding textual descriptions [34]. Other pre-trained models are widely used for language processing, such as mT5 [35] and T5 [36], which were earlier works of LLMs using transfer learning. GPT-3, an OpenAI model, is another advanced example of LLM trained on extensive textual data, and it is a zero-shot transferable to other tasks without fine-tuning [37]. Bidirectional encoder representations from transformers (BERT) are another widely used LLM model that uses neural networks, and they were introduced by Google AI [38]. From its name, this model uses a bidirectional approach that understands context from both the left and right sides of a word. Its challenge lies in forming the architecture of its neural networks to be built for the specific machine learning problem, which requires intensive training.

While these models have been widely used in fields such as computer vision and language processing applications, their use in human workload prediction, particularly in mixed-reality multitasking environments, remains relatively underexplored. For example, the authors in Ref. [39] mention that language models generally result in slower training times and higher running costs when solving tasks requiring human-level performance. This gap shows there is still a lot to learn about applying transfer learning in workload prediction, especially in immersive environments like mixed reality.

Only a few studies have used transfer learning to improve the accuracy and efficiency of workload prediction models, mainly in applications derived from computer vision. For example, see reference [40] for predicting pilot workload based

on electrocardiogram (ECG) signals using pre-trained deep convolutional neural networks (CNNs). The researchers collected data by recording pilots while they performed helicopter maneuvers of different difficulty levels. They compared two types of ECG-based visual representations to see which worked better for predicting workload. They tested two ways of using pre-trained deep learning models, one involved the pre-trained CNNs to extract features and then training a traditional machine learning model like the support vector machines (SVMs), and the other involved fine-tuning the CNN using the ECG-derived images. They found that using a pre-trained CNN just for feature extraction gave better prediction accuracy than fine-tuning the model. The best accuracy achieved using transfer learning was 42.44%.

In another example, in Ref. [41], the authors predicted the cognitive workload using electroencephalogram brain signals collected during different versions of the N-back task in a virtual reality environment. They relied on traditional machine learning models for prediction, like logistic regression, Naïve Bayes, decision trees, k-nearest neighbor, random forest, and support vector machine. The model accuracies for prediction were 27%, 22%, 32%, 28%, 27%, and 33%, respectively. The support vector machine outperformed all the other models. They also tested deep neural networks, with and without neuron dropout, and long short-term memory (LSTM) networks for comparison. The prediction accuracies for these models were 34%, 29%, and 31%, respectively, suggesting that the support vector machine still achieved better overall. They also used the pre-trained VGG16 model to compare it to the other models. The accuracy of this model was 35%, which was higher than all the other models used in this study. Still, despite the improvement in the transfer learning approach, the authors noted that it did not yield a significant improvement.

3. Methodology

3.1 Experimental implementation

A total of 36 eligible participants with an average age of 23.9 and a standard deviation of 4.22, took part in a multitasking experiment involving both a real-world (RW) task and a virtual world (VW) task. The RW task was a block-matching task where participants used their hands to pair English letters (blocks) into their location on a wooden board. This task included a visual search, which has been shown in previous studies, such as in Ref. [42], to significantly impact cognitive workload. The N-back task was assigned as the VW task. This task was developed and augmented into the HoloLens 2 device. The N-back task is commonly used to evaluate working memory and its effect on workload, with the 'N' parameter indicating the number of steps needed to recall specific information from memory in response to a given stimulus [43]. In our experiment, the virtual N-back task involved matching the color of a virtual cube using hand gestures, all within a given 90-second time limit. Each participant was presented with 34 stimuli during the task.

To begin, participants practiced a trial version for each world task and completed a pairwise comparison survey of the NASA-TLX, which compared its dimensions 15 times. This survey lets participants weigh the relative importance of each dimension for each world task. After that, they moved on to the actual experiment with four multitasking conditions, varying in difficulty level from easy to hard. For the RW task, if participants were not given a specific order to pair the English letters on the board, the task was considered easy. However, when they were instructed to follow

a specific order for the search, it was classified as a hard level. For the VW task, the easy level involved an N-back value of 1, which means that participants needed to go one step back in memory to recall the color of the virtual cube. When the N-back level was set to 2, participants needed to remember the color of the virtual cube from two steps earlier, and this was classified as the hard level. After each multitasking condition, participants completed the post-task NASA-TLX survey, rating each dimension from 0 to 100% based on how they felt during the experiment. All participants were compensated with \$15 for their participation and time. The experiment lasted about 50 minutes for each participant, while the overall data collection process took around 3 months. All surveys were filled out using the Qualtrics survey tool. The study obtained approval from the Institutional Review Board (IRB) at the University of Illinois, Chicago (IRB# 2020-0466). **Figure 1** shows the experimental setup.

3.2 Data collection

In total, 144 observations were collected, with each of the 36 participants performing four multitasking conditions. There was no missing data. The responses recorded from the ratings of each NASA-TLX dimension were combined into a weighted average, which served as the output variable to be predicted. This weighted average was calculated by dividing the sum-product of the ratings for each NASA-TLX dimension (0–100%) from the post-task survey and their corresponding counts from the pairwise survey by 15. Eq. (1) represents the calculation of this overall workload:

$$\text{Weighted Rating} = \frac{\sum_{i=1}^6 (\text{Rating Response} \times \text{Count})_i}{15} \quad (1)$$

Four independent variables were selected for the study, including the task difficulty variable as categorical with four levels, gender is also categorical with two levels, and familiarity and experience with AR/VR handheld and wearable devices are numerical and expressed in percentages (0–100%). The age variable was excluded due to its low variation between participants. The dependent variable is the weighted

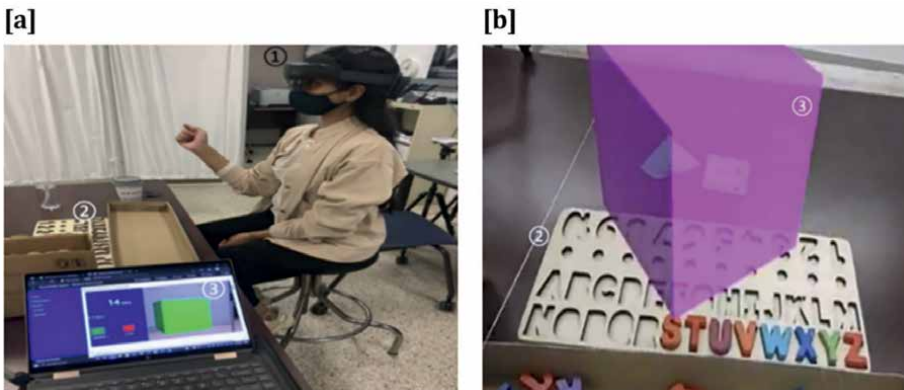


Figure 1. Experimental Setup. (a) The participants are engaging with the VW task. (b) The combination of RW and the VW tasks. ① Microsoft HoloLens 2, ② real-world task material, ③ virtual-world material (reproduced from Ref. [44], under a Creative Commons license).

workload averaged from the six NASA-TLX dimensions. However, the data collection process faced several challenges, such as time constraints, a limited number of participants, and disruptions caused by the COVID-19 pandemic. Therefore, machine learning for data augmentation was used to expand the original dataset for better prediction modeling.

3.3 Data augmentation

The original dataset was expanded *via* the generative adversarial networks (GANs) in a Google Colab notebook. The original dataset was divided based on each categorical variable to ensure equal representation in the generated data, matching that of the original dataset. This means that the categorical variables were reintroduced afterward. An 80/20 splitting rule was used after uploading the required libraries into the notebook. Three neural networks were constructed for the model: the generator network with four dense layers, the discriminator network with three dense layers, and the GAN model network, which combines the two networks. The 'LeakyReLU' activation function was applied for the hidden layers for both models, and the 'Sigmoid' was used for the output layers. The GAN model was trained as a binary classifier with a 'binary_crossentropy' loss function and optimized using the Adam optimizer. Eq. (2), derived from reference [45], represents how the GAN objective function is optimized by the generator G and the discriminator D through a min-max game:

$$\min_G \max_D V(D, G) = E_{x \sim P_{data}(x)} [\log D(x)] + E_{z \sim P_z} [1 - \log D(G(z))] \quad (2)$$

The optimal value function, $V(D, G)$, is used when the GAN model is trained. E_x is the expected value of the discriminator output for actual samples signed as x . $P_{data}(x)$ is the distribution of actual samples, and $D(x)$ gives a probability of an input x belonging to the actual data. $\log(D(x))$ represents the discriminator's confidence in correctly classifying a real sample as real. The higher this value, the better the discriminator identifies real data points. E_z is the expected value of the discriminator output for fake samples signed as z , where P_z is Gaussian noise, and $G(z)$ generates fake samples that struggle to deceive the discriminator by learning how to produce data that comes from the same distribution $P_{data}(x)$. $[1 - \log(D(G(z)))]$ is used to encourage the generator to maximize the probability of the discriminator classifying fake data as real, thus minimizing the difference between real and generated data distributions.

The generator and discriminator were trained individually, and the GAN was trained on a newly generated batch of fake samples that were labeled as 'real' for testing, with a total of 1000 epochs. Three losses were calculated after each epoch: generator loss, discriminator loss, and GAN loss. The goal is to minimize the GAN loss, which is the combination of both losses. The model assessed the current GAN loss and compared it with the 'best_loss' variable during each training epoch. The GAN loss is updated with the current loss value. If the current loss is better than the 'best_loss' by at least 'min_delta', which is assigned at 0.011. If there was no improvement in the GAN, the model stopped training. The early stopping criterion was triggered based on the 'patience' parameter, which was set to 20. For each dataset, the batch sizes used for comparison were 2, 3, 6, and 9, respectively. A batch size of 2 achieved the lowest losses. **Table 1** summarizes the best GAN losses achieved for batch size 2, categorized by task difficulty and gender.

Difficulty Level	Best GAN Loss (Male)	Best GAN Loss (Female)
Easy real task, easy digital task	0.63	0.59
Easy real task, hard digital task	0.66	0.6
Hard real task, easy digital task	0.62	0.53
Hard real task, hard digital task	0.52	0.60

Table 1.
The Best GAN Loss for Each Batch Size per Category.

The ‘latent_dim’ parameter was tested with different values of 10, 100, and 500 for experimentation. This parameter influences how diverse the generated output can be. However, despite the different trials, the different values of this parameter did not show a significant difference, so the parameter was set to 500, where a larger value indicates more diverse generated data. A total of 5000 artificial samples were generated using the ‘generate_samples’ function, after several trials to determine the appropriate number of synthetic samples. While more data is desirable, it is important to maintain a balance between data quality and diversity. After storing the resulting synthesized data in a CSV file, it was retrieved for evaluation and combined with the original data for use in model building, as shown in the next section. **Figure 2** illustrates the process of constructing the GAN model through a flow chart.

3.4 Prediction model construction

To enhance the workload prediction in our multitasking experiment, we applied a transfer learning approach using one of the pre-trained models from computer vision applications, called the CLIP model. Although this model has been used more in image-based applications in the previous literature, applying it to predict a numerical output variable from continuous tabular data presents some challenges. Thus, we propose a hybrid model that combines CLIP with a simple regression model, demonstrating the power of transfer learning even when tailoring pre-trained models to specific tasks.

The modeling was done in a Google Colab notebook. After uploading the necessary libraries and installing the CLIP model from its GitHub repository, the combined data file, including both the original and generated observations, was loaded into the code. The categorical variables were encoded as 0 and 1 s. Since the CLIP model is designed for images, a function called ‘rows_to_images’ was defined to convert the data into images. This function takes a set of arguments representing a range of rows (data_chunk) and an output directory (output_dir) to save the image data. A subplot with specified dimensions was created using the ‘matplotlib’ library. For each row in the given data chunk, relevant features were extracted, including the input variable columns. The chunk size of 10 was randomly selected to distribute the processing load among multiple processes for better optimization, as the combined data is large. A list of arguments named ‘args_list’ was created, where each argument corresponded to a specific chunk of data that included the start and end indices, the chunk of data itself, and the output directory for saving the images. The code then utilized the multiprocessing module to create a pool of processes, and the ‘rows_to_images’ function was applied to each chunk in parallel using the ‘map’ function. This helps employ multiple CPU cores to enhance efficiency. The resulting images were saved in the specified directory and named ‘image_dataset’.

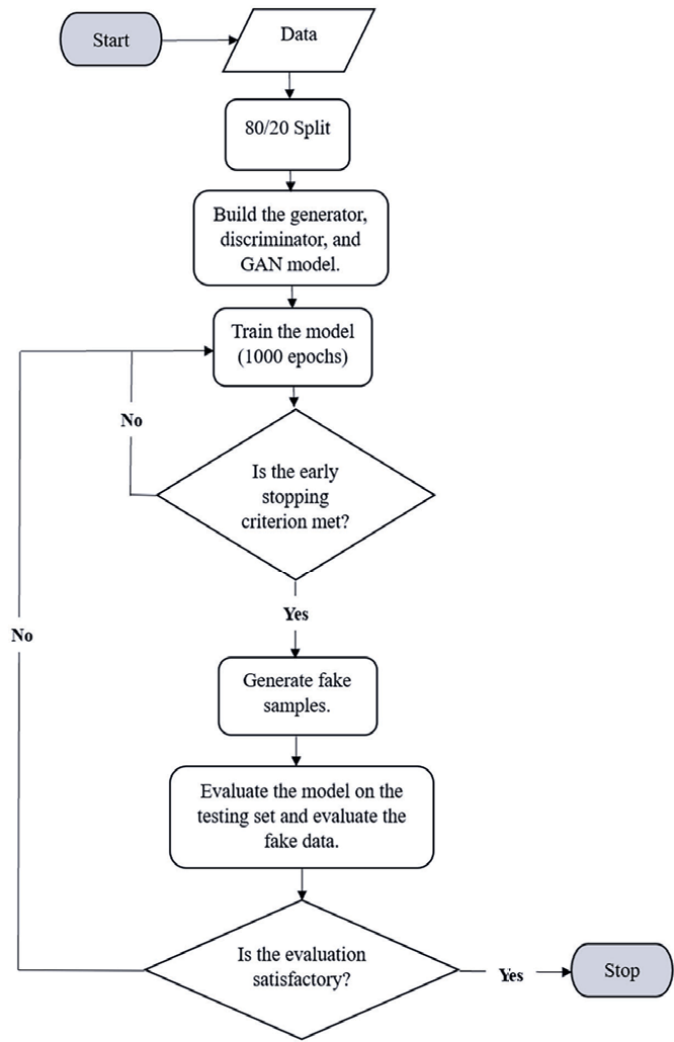


Figure 2.
GAN model construction process flowchart.

The variable 'data,' which originally stored the imported combined data, was then updated with the image paths and the corresponding actual values of the output variable, the weighted workload. Image pre-processing was performed using the PyTorch library, known for its dynamic computational graphs for deep learning tasks. The 'transform.Compose' function was then utilized to create a sequence of image transformations. These transformations included resizing each image to specified pixel dimensions using the 'transforms.Resize', converting the images to grayscale with three channels 'num_output_channels = 3', which is the default setting using 'transforms.Grayscale', and transforming the images into PyTorch tensors with the 'transforms.ToTensor'. For each path of images, the 'Image.open' function from the PIL library opened the images and applied the defined transformations. The resulting images were then stacked into a PyTorch tensor using the 'torch.stack'. This transformation prepared the images for further processing, like feeding them into a neural network.

The 'clip. Load' function loads the CLIP model with a specific architecture ('ViT-B/32') on the device after checking the availability of the (cuda) GPT that offloads the computation to the GPU to speed up the training. The ViT-B/32 is a vision-transformed neural network that applies transformer layers initially designed for natural language processing to computer vision applications where the input images are divided into patches of 32×32 pixels. The 'model.encode_image' function was then used to encode the preprocessed images into feature vectors converted to NumPy arrays. After that, the data were split into 80% for training and 20% for testing. The defined hybrid model, CLIP-regression, was first fine-tuned to improve its performance on our specific task of mixed-reality multitasking. In fine-tuning the model, a neural network is defined by the name CLIPRegression from the 'nn. Module', the base class for all neural networks in PyTorch. It maps 512-dimensional input features from the CLIP model to a single output value. The model layers were initiated using the '_init_' method and the 'ReLU' activation function. Then, the 'forward' method was used to define how the input data is moved through the network. An input tensor x , representing the features extracted from the CLIP model, was taken and passed through the fully connected layer 'self.fc', producing a single numerical value for the regression prediction. Then, the model was trained using the Adam optimizer, applying a learning rate of 0.0001. The choice of involving a regression model was to be able to predict the numerical weighted rating output. A total of 20 epochs was used based on multiple experimental trials to balance sufficient model training without overfitting. **Figure 3** shows an illustration of this model's construction.

The hybrid model's predictions were evaluated using the known RMSE metric, which involves predicting a numerical output. A simple regression model was applied to our combined dataset to compare the results of the hybrid model with those of a traditional machine learning model. After uploading the data file and the necessary libraries, the data was split using the 80/20 rule. The simple regression model was initiated using the 'LinearRegression()' function with the fit method. After training, the model predicted outcomes for the test dataset through the 'predict' method, and its performance was assessed using the RMSE value. Eq. (3) was used to calculate all the RMSE values:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (3)$$

Here, y_i refers to the actual values from the training data, while $\hat{y}_i$ represents the values predicted by the model. The difference between them is called the error or residual. Squaring the errors ensures that positives and negatives are counted equally

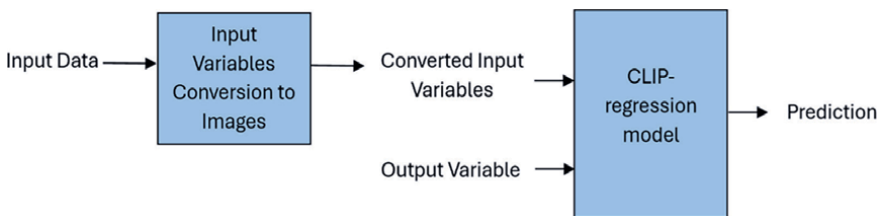


Figure 3.
CLIP-regression hybrid model architecture.

without canceling each other out. Then, summing up all the squared errors and taking the square root brings the error back to the same scale as the original data.

4. Results

The performance of the GAN model was evaluated first by comparing the distribution between the generated samples of each variable with its corresponding distribution in the testing dataset using the non-parametric Kolmogorov-Smirnov (KS) test [46]. The significance level was set at 0.05. The null hypothesis was set as the two datasets have different distributions. The overall quality score was also calculated by calling the 'evaluate_quality' function from the 'sdv.evaluation' module in case of a matching distribution. The synthesized samples were also evaluated using the KS statistical test of each variable with its corresponding variable in the original dataset. We found a matching distribution between the generated samples and the testing dataset, with p-values < 0.05 . The highest quality score achieved for the synthesized data was 65.85%. Additionally, from the KS test on the synthesized data, all p-values were < 0.05 , indicating a matching distribution.

For evaluating the proposed hybrid prediction model integrating the CLIP model, the best RMSE value achieved was 0.96, and the RMSE for the simple regression was 7.10. Thus, our results indicate that integrating the transfer learning approach to predict human workload in a multitasking mixed reality experiment yields better results than traditional machine learning models. **Figure 4** visually compares the performance between the two models on a random subset of data points, highlighting the variations between the actual and predicted values.

5. Discussion

This study aimed to construct a hybrid model that predicts human workload in mixed reality multitasking by incorporating the CLIP model that is widely used for computer vision applications. The GAN model was first employed to enlarge the

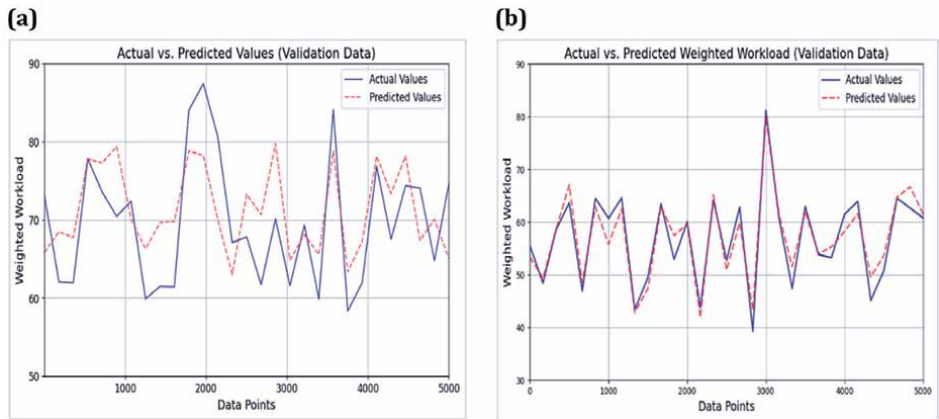


Figure 4.
Actual vs. Predicted Weighted Workload Values: (a) for the simple regression model. (b) for the hybrid CLIP-regression model.

original dataset, achieving results that closely resembled the original samples. This enlargement enhanced the prediction of the hybrid model, as the more data there is, the better the model performs. Our findings indicate that the proposed model outperformed the traditional regression model in machine learning, resulting in a much lower RMSE value.

This study emphasizes the possibility of engaging pre-trained models in particular tasks, such as mixed reality multitasking workload prediction. By leveraging the strengths of transfer learning, particularly with the CLIP model, the hybrid model successfully adapted to the unique characteristics of the task at hand. This suggests that pre-trained models, which have succeeded in computer vision, can be effectively repurposed for more complex, domain-specific applications. Nevertheless, the study has some limitations that can be addressed in future research. For instance, the age of the participants was excluded as most were in the same age group. Other variables could be considered to increase the number of inputs to the model, such as the dominant hand and overall health status, which may offer additional insights into individual differences in workload perception. Moreover, augmenting more than 5000 samples would be beneficial, but we had limitations in terms of data quality and distribution matching. Finally, while our hybrid model performed better than the traditional regression model, future studies could compare its performance with more advanced models, such as deep learning architecture, to fully assess its potential.

6. Conclusion

This study implemented a transfer learning approach to predict the weighted workload rating in multitasking between the real and the digital worlds, using the data collected from the widely used subjective workload measure, the NASA-TLX. This model aims to reduce the time, effort, and cost required to measure human workload in a new, similar experiment without the need to collect NASA-TLX ratings, instead using the model. Efforts to improve human-computer interactions have been increasing over time. Therefore, this study contributes to this effort by automating the process of evaluating human workload, making it more efficient and cost-effective, especially for industries utilizing mixed reality to help accommodate their workers' well-being. Participants found multitasking between tasks challenging due to the time pressure. Time pressure is a critical factor in digital environments [47]. However, their experience working with such digital technologies varied, contributing to the varied ratings of workload. Those with more experience tended to manage multitasking better, while participants with less experience struggled more, leading to higher ratings. This variation underscores the importance of considering individual differences, such as prior experience and familiarity with such technologies as mixed reality, when evaluating human workload in digital environments. Understanding these nuances will be crucial in optimizing human performance.

Acknowledgements

The authors would like to thank Bryan Havens and Yalda Ghasemi for their initial insights on data collection, and Edison Larco for his assistance in developing the virtual task using Unity Software.

Conflict of interest

The author(s) declared no potential conflicts of interest concerning this chapter's research, authorship, and/or publication.

Author details

Safanah Abbas¹, Heejin Jeong² and David He^{1*}

1 Department of Mechanical and Industrial Engineering,
University of Illinois at Chicago, Chicago, IL, United States

2 Human Systems Engineering, The Polytechnic School, Arizona State University,
Mesa, AZ, United States

*Address all correspondence to: davidhe@uic.edu

IntechOpen

© 2025 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

References

- [1] Spink A, Cole C, Waller M. Multitasking behavior. *Annual Review of Information Science and Technology*. 2008;**42**(1):93-118. DOI: 10.1002/aris.2008.1440420110
- [2] Abrash M. Creating the future: Augmented reality, the next human-machine interface. In: 2021 IEEE Int Electron Devices Meet (IEDM). Piscataway, NJ, USA: IEEE (Institute of Electrical and Electronics Engineers); 2021. DOI: 10.1109/iedm19574.2021.9720526
- [3] Speicher M, Hall BD, Nebeling M. What is mixed reality? In: Proc 2019 CHI Conf Hum Factors Comput Syst. New York, NY, USA: ACM (Association for Computing Machinery); 2019. DOI: 10.1145/3290605.3300767
- [4] Tang YM, Au KM, Lau HC, Ho GT, Wu CH. Evaluating the effectiveness of learning design with mixed reality (MR) in higher education. *Virtual Reality*. 2020;**24**(4):797-807. DOI: 10.1007/s10055-020-00427-9
- [5] Rogers S. What Is Mixed Reality, and What Does it Mean for Enterprise? *Forbes* [Internet]. New York City, NY, USA; 2018. Available from: <https://www.forbes.com/sites/solrogers/2018/12/04/what-is-mixed-reality-and-what-does-it-mean-for-enterprise/> [Accessed: April 16, 2025]
- [6] Gownder JP, Ask J, Hewitt A, Voce C, Birrell R. *How Enterprise Smart Glasses Will Drive Workforce Enablement*. Cambridge, MA, USA: Forrester; 2016. Available from: <https://www.forrester.com/report/How-Enterprise-Smart-Glasses-Will-Drive-Workforce-Enablement/RES133722>
- [7] Rokhsaritalemi S, Sadeghi-Niaraki A, Choi S-M. A review on mixed reality: Current trends, challenges and prospects. *Applied Sciences*. 2020;**10**(2):636. DOI: 10.3390/app10020636
- [8] Gawron VJ. *Human Performance, Workload, and Situational Awareness Measures Handbook, –2-Volume Set*. Boca Raton, FL, USA: CRC Press; 2019
- [9] Matthews G, Reinerman-Jones L, Wohleber R, Lin J, Mercado J, Abich J. Workload is multidimensional, not unitary: What now? In: *Foundations of Augmented Cognition*. Cham, Switzerland: Springer; 2015. pp. 44-55. DOI: 10.1007/978-3-319-20816-9_5
- [10] Hart SG, Staveland LE. Development of NASA-TLX (task load index): Results of empirical and theoretical research. *Advances in Psychology*. 1988;**52**:139-183. DOI: 10.1016/s0166-4115(08)62386-9
- [11] Zhuang F, Qi Z, Duan K, Xi D, Zhu Y, Zhu H, et al. A comprehensive survey on transfer learning. *Proceedings of the IEEE*. 2020;**109**(1):43-76. DOI: 10.1109/jproc.2020.3004555
- [12] Chen Y, Fang W, Guo B, Bao H. The moderation effects of task attributes and mental fatigue on post-interruption task performance in a concurrent multitasking environment. *Applied Ergonomics*. 2022;**102**:103764. DOI: 10.1016/j.apergo.2022.103764
- [13] Liu P, Li Z. Toward understanding the relationship between task complexity and task performance. In: *Lecture Notes in Computer Science*. Berlin, Heidelberg, Germany: Springer; 2011. pp. 192-200. DOI: 10.1007/978-3-642-21660-2_22

- [14] Jeffri NF, Awang Rambli DR. A review of augmented reality systems and their effects on mental workload and task performance. *Heliyon*. 2021;7(3):e06277. DOI: 10.1016/j.heliyon.2021.e06277
- [15] Rubio S, Díaz E, Martín J, Puente JM. Evaluation of subjective mental workload: A comparison of SWAT, NASA-TLX, and workload profile methods. *Applied Psychology*. 2004;53(1):61-86. DOI: 10.1111/j.1464-0597.2004.00161.x
- [16] Biondi FN, Cacanindin A, Douglas C, Cort J. Overloaded and at work: Investigating the effect of cognitive workload on assembly task performance. *Human Factors*. 2021;63(5):813-820. DOI: 10.1177/0018720820929928
- [17] Kretschmer V, Plewan T, Rinkenauer G, Maettig B. Smart palletisation: Cognitive ergonomics in augmented reality based palletising. In: *Intell Hum Syst Integr*. Cham, Switzerland: Springer; 2018. pp. 355-360. DOI: 10.1007/978-3-319-73888-8_55
- [18] Rebol M, Steinmaurer A, Gamillscheg F, Pietroszek K, Gütl C, Ranniger C, et al. CPR emergency assistance through mixed reality communication. In: *Lecture Notes in Computer Science*. Cham, Switzerland: Springer; 2023. pp. 415-429. DOI: 10.1007/978-3-031-32883-1_38
- [19] Brunzini A, Papetti A, Messi D, Germani M. A comprehensive method to design and assess mixed reality simulations. *Virtual Reality*. 2022;26(4):1257-1275. DOI: 10.1007/s10055-022-00632-8
- [20] Nenna F, Zanardi D, Gamberini L. Effects of presence on human performance and workload in simulated VR-based telerobotics. In: *Proc 16th Int Conf Pervasive Technol Relat Assist Environ*. New York, NY, USA: ACM (Association for Computing Machinery); 2023. DOI: 10.1145/3594806.3594856
- [21] Szczurek KA, Cittadini R, Prades RM, Matheson E, Di Castro M. Enhanced human-robot interface with operator physiological parameters monitoring and 3D mixed reality. *IEEE Access*. 2023;11:39555-39576. DOI: 10.1109/access.2023.3268986
- [22] Zhang H, Xu T, Li H, Zhang S, Wang X, Huang X, et al. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: *Proceedings IEEE International Conference on Computer Vision (ICCV)*. Piscataway, NJ, USA: IEEE; 2017. DOI: 10.1109/iccv.2017.629
- [23] Antoniou A, Storkey A, Edwards H. Augmenting image classifiers using data augmentation generative adversarial networks. In: *Artif Neural Networks Mach Learn – ICANN* 2018. Vol. 2018. Cham, Switzerland: Springer; 2018. pp. 594-603. DOI: 10.1007/978-3-030-01424-7_58
- [24] Waheed A, Goyal M, Gupta D, Khanna A, Al-Turjman F, Pinheiro PR. CovidGAN: Data augmentation using auxiliary classifier GAN for improved COVID-19 detection. *IEEE Access*. 2020;8:91916-91923. DOI: 10.1109/access.2020.2994762
- [25] Bourou S, El Saer A, Velivassaki T-H, Voulkidis A, Zahariadis T. A review of tabular data synthesis using GANs on an IDS dataset. *Information*. 2021;12(9):375. DOI: 10.3390/info12090375
- [26] Zhao Z, Kunar A, Birke R, Van der Scheer H, Chen LY. CTAB-GAN+: Enhancing tabular data synthesis. *Frontiers in Big Data*. 2024;6. DOI: 10.3389/fdata.2023.1296508
- [27] Skaramagkas V, Ktistakis E, Manousos D, Tachos NS, Kazantzaki E,

- Tripoliti EE, et al. Cognitive workload level estimation based on eye tracking: A machine learning approach. In: 2021 IEEE 21st International Conference on Bioinformatics and Bioengineering (BIBE). Piscataway, NJ, USA: IEEE; 2021. DOI: 10.1109/bibe52308.2021.9635166
- [28] Yim S, Yoon C, Yoo S, Jang Y. A mental workload estimation for visualization evaluation using EEG data and NASA-TLX. In: EuroVis (Posters). Chichester, UK: Eurographics Association; 2022. pp. 47-49
- [29] Abdurrahman UA, Yeh S-C, Wong Y, Wei L. Effects of neuro-cognitive load on learning transfer using a virtual reality-based driving system. *Big Data and Cognitive Computing*. 2021;5(4):54. DOI: 10.3390/bdcc5040054
- [30] Qin Y, Bulbul T. Electroencephalogram-based mental workload prediction for using augmented reality head mounted display in construction assembly: A deep learning approach. *Automation in Construction*. 2023;152:104892. DOI: 10.1016/j.autcon.2023.104892
- [31] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. (International Conference on Learning Representations) arXiv [Preprint]. 2014. Available from: <https://arxiv.org/abs/1409.1556> [Accessed: April 17, 2025]
- [32] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Las Vegas, NV, USA. Piscataway, NJ, USA: IEEE (Institute of Electrical and Electronics Engineers); 2016. pp. 770-778. DOI: 10.1109/CVPR.2016.90
- [33] Atliha V, Sesok D. Comparison of VGG and ResNet used as encoders for image captioning. In: 2020 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream). Piscataway, NJ, USA: IEEE; 2020. DOI: 10.1109/estream50540.2020.9108880
- [34] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. Cambridge, MA, USA: PMLR; 2021. pp. 8748-8763
- [35] Xue L, Constant N, Roberts A, Kale M, Al-Rfou R, Siddhant A, et al. MT5: A massively multilingual pre-trained text-to-text transformer. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA, USA: Association for Computational Linguistics (ACL); 2021. DOI: 10.18653/v1/2021.naacl-main.41
- [36] Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*. 2020;21(140):1-67
- [37] Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*. 2020;33:1877-1901
- [38] Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2019 Proceedings) arXiv [Preprint]. 2018. pp. 4171-4186. Available from: <https://arxiv.org/abs/1810.03702>

org/abs/1810.04805 [Accessed: April 17, 2025]

[39] Naveed H, Khan AU, Qiu S, Saqib M, Anwar S, Usman M, et al. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology (ACM TIST)*. arXiv [Preprint]. 2023;**16**. Available from: <https://arxiv.org/abs/2307.06435> [Accessed: April 17, 2025]

[40] Xi P, Law A, Goubran R, Shu C. Pilot workload prediction from ECG using deep convolutional neural networks. In: 2019 IEEE International Symposium on Medical Measurements and Applications (MeMeA). Pavia, Italy: IEEE (Institute of Electrical and Electronics Engineers); 2019. pp. 1-6. DOI: 10.1109/memea.2019.8802158

[41] Ved H, Yildirim C. Detecting mental workload in virtual reality using EEG spectral data: A deep learning approach. In: 2021 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR). Piscataway, NJ, USA: IEEE; 2021. DOI: 10.1109/aivr52153.2021.00039

[42] Pang L, Fan Y, Deng Y, Wang X, Wang T. Mental workload classification by eye movements in visual search tasks. In: 2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI). Piscataway, NJ, USA: IEEE; 2020. DOI: 10.1109/cisp-bmei51763.2020.9263668

[43] Chen Y-N, Mitra S, Schlaghecken F. Sub-processes of working memory in the N-back task: An investigation using ERPs. *Clinical Neurophysiology*. 2008;**119**(7):1546-1559. DOI: 10.1016/j.clinph.2008.03.003

[44] Abbas S, Jeong H. Unveiling gender differences: A mixed reality multitasking

exploration. *Frontiers in Virtual Reality*. 2024;**4**:1308133. DOI: 10.3389/frvir.2023.1308133

[45] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. In: *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Neural Information Processing Systems Foundation; 2014. pp. 2672-2680

[46] Kolmogorov AN, Smirnov NV. On the empirical determination of a distribution law. *Sankhyā: The Indian Journal of Statistics*. 1933;**4**(2):116-119

[47] Naumann J. The skilled, the knowledgeable, and the motivated: Investigating the strategic allocation of time on task in a computer-based assessment. *Frontiers in Psychology*. 2019;**10**:1429. DOI: 10.3389/fpsyg.2019.01429

Domain Adaptation in Multimodal Models

Raghavendran Ramakrishnan

Abstract

Multimodal AI systems have the ability to represent and relate information in multiple data modalities. This enables a wide range of new applications when deployed in real world. They also face major challenges particularly due to domain shift arising from differences in the data they are trained on and the environment they encounter. This chapter focuses on domain adaptation for these systems. It involves assisting models to adjust to new conditions across different input modalities, such as text, images, or sensor data. Unlike single-modality systems, multimodal models must deal with unique challenges. This includes mismatch in data types, data quality, and conflicts during training across multiple modalities. We introduce these unique challenges with the multimodal systems. We also provide theoretical foundations using risk minimization and measures of divergence under domain shift across different modalities. From there, we also explore practical approaches such as adversarial training to align features, contrastive learning, and flexible fusion techniques that adjust when some inputs are unreliable. We also explore advanced techniques such as deployment-stage adaptation and fine-tuning large foundation models without complete retraining. We provide a summary of different datasets developed for multimodal domain adaptation. The chapter ends with summarizing key insights discussed throughout the chapter along with highlighting emerging opportunities in the domain.

Keywords: multimodal domain adaptation, domain shift, modality gap, distribution alignment, adversarial alignment, contrastive learning

1. Introduction

Multimodal domain adaptation (MMDA) is at the junction of two active research areas in machine learning, multimodal learning, and domain adaptation [1]. Multimodal learning explores the integration of informations from heterogeneous data types (e.g., images, text, audio, and LiDAR). Domain adaptation studies the challenges of transferring models trained on a “source” environment to “target” environment where there are discrepancies between the two distributions [2]. In real-world, data usually occurs in various modalities and undergoes shifts in appearance, noise features, labeling standards, or environmental factors. It is essential for models to associate information from multiple modalities while being resilient to distribution shifts [3]. The techniques developed as part of multimodal domain adaptation find applications in wide range of areas ranging from autonomous driving, robotics to medical diagnostics, and multimedia exploration.

The techniques developed in traditional unimodal domain adaptation address challenges due to discrepancies between source and target distributions within a single modality. MMDA extends on this and addresses the discrepancies in domain shifts under different environmental conditions while leveraging associations between different data modalities. The complexity of MMDA arises from several related challenges that separate it from unimodal version. First, the domain shift impacts each modalities differently. Each modality displays diverse degrees of discrepancy between source and target domains [4]. It requires specialized alignment strategies that address the unique properties of each modality while preserving their associations. Second, the modality gap creates additional complexity since different modalities reside in unique feature spaces, making it difficult to compare and align them [3]. Third, the data modalities often vary in quality and might contain imbalance and incomplete information in real-world scenarios. This requires robust adaptation frameworks that are capable of handling missing modalities.

The theoretical frameworks developed in the field provide principled approaches to understanding several challenges in MMDA. The mathematical formulations of multimodal domain shift, modality-specific divergence metrics, and information-theoretic approaches help in understanding the tradeoffs between modality-association and domain-associations [2]. Novel techniques including adversarial alignment for multimodal settings and contrastive learning approaches that leverage associations between modalities have been developed to address the emerging challenges in the field.

This chapter offers a comprehensive outlook of multimodal domain adaptation, from both theoretical perspective, and offers practical insights for researchers and practitioners. First, we motivate the requirement for multimodal domain adaptation, by offering relevant real-world challenges. This includes discussion on domain discrepancy, data distribution shifts, modality incompleteness and imbalance issues, and challenges in integration with foundation models. Second, we lay out the foundational concepts involved in multimodal adaptation. We discuss the relationship between domain shift, modality gaps, and distribution alignment. Third, we describe the mathematical formulations for multimodal domain shift, modality-specific divergence metrics and offer perspectives from information theory. Fourth, we explore core techniques such as adversarial alignment, contrastive learning, and cross-modal interactions. Fifth, we discuss advanced techniques in the field including test-time adaptation frameworks, integration with foundation models, and emerging techniques in open-set and partial domain adaptation. Finally, we discuss standard benchmarks and datasets developed in the field for various adaptation tasks.

2. Multimodal domain adaptation

2.1 Real-world challenges in cross-domain deployment

Multimodal systems face unique challenges due to the interplay between different data types and changing real-world conditions.

2.1.1 Domain discrepancy and data distribution shift in multimodal systems

Multimodal models experience complex domain shifts when deployed in new environments [2]. The distribution mismatches simultaneously affect multiple input

modalities. One example is from action recognition systems, where the objective is to automatically identify and classify human actions into a set of action categories [5]. The systems can use multiple modalities such as RGB frames, optical flow patterns, and audio streams. Models trained on indoor kitchen environments (EPIC-Kitchens dataset [6]) show major performance degradation when applied to outdoor settings due to distribution shifts: variations in lighting settings, fixed cameras to handheld recording, and audio variations from indoor acoustics to ambient outdoor noise. This joint distribution shift can be quantified through multimodal Kullback-Leibler divergence:

$$D_{\text{KL}}(P_{\text{source}}, P_{\text{target}}) = \mathbb{E}_{\mathbf{x} \sim P_{\text{source}}} \left[\log \frac{P_{\text{source}}(\mathbf{x})}{P_{\text{target}}(\mathbf{x})} \right] \quad (1)$$

The results show accuracy drops in outdoor settings, which are addressed by authors using adaptive feature fusion that reweighs modality contributions based on the environment settings. The complexity of such systems increases exponentially compared to single-modality cases, as cross-modal interactions must be taken into account during domain alignment.

2.1.2 Modality incompleteness/imbalance

There can be cases where information from one modality is either partially available or degraded. This can result in alignment gaps in conventional fusion approaches. For example, noisy environments like construction sites could render audio streams unusable. The authors in [7] address this challenge of missing modality using relation-aware missing modal (RMM) generator which uses visual-text cues to reconstruct the missing audio features. Similarly, Bayesian feature reconstruction was used to improve the accuracy of action recognition when the audio was noisy, by preserving the temporal correlations between lip movements and phoneme sequences.

2.1.3 Data scarcity in specialized domains

In highly regulated domains such as medical imaging, data scarcity could pose a serious challenge in developing effective multimodal models. The scarcity could arise due to multiple factors, such as privacy restrictions, high annotation costs, and diverse data formats across imaging modalities and standards. Recently, multitask learning frameworks such as UMedPT address these data challenges by learning generalizable representations from fragmented datasets through gradient accumulation training [8]. UMedPT was trained on 17 tasks that include classification, segmentation, and detection, based on 15 datasets across diverse modalities (MRI, X-ray, and histopathology). Although sufficient data are available for some medical tasks, data scarcity could become an issue for a long tail of clinically relevant tasks [9]. Foundation models pretrained on medical data, such as BiomedCLIP, achieve superior few-shot learning performance by utilizing domain-specific visual-text correlations.

2.1.4 Integration with multimodal foundation models

The unique challenges with integration with multimodal foundation models arise from the architectural scale, cross-modal dependencies, and knowledge preservation requirements. We discuss five interconnected challenges that are unique from

traditional unimodal or small-scale models. These challenges are rooted in preserving pre-trained knowledge, maintaining cross-modal alignment and adapting to domain-specific requirements.

- *Modality-specific domain shift dynamics:* Each modality exhibit different levels of sensitivity to distribution changes. For instance, while text embeddings in CLIP are reliably stable across domains, visual features show higher variance under illumination or style changes [10]. It creates asymmetric adaptation requirements: visual components requires extensive retraining on target domains while textual components benefit from prompt engineering alone. Recent works address them through modality-specific adaptation layers that apply different regularization strengths [11].
- *Catastrophic forgetting in cross-modal spaces:* Full fine-tuning of MFMs (multimodal foundation models) introduce risk of overwriting carefully learnt cross-modal features impacting their zero-shot performance. Traditional fine-tuning is shown to reduce CLIP's out-of-distribution accuracy by 40–60% while improving target domain performance by only 15–20%. [12]. The challenge arises because gradient updates optimized for learning target domains impact the embedding spaces crucial for multimodal reasoning. Parameter-efficient fine-tuning mitigate this, but introduce new challenges: adapter-based approaches in AURORA achieve almost full fine-tuning performance with 0.04% of trainable parameters, yet struggle with cross-modal task transfer [11]. Advanced techniques combines dynamic architecture selection (choosing appropriate adapter configurations based on domain similarity) with contrastive alignment preservation losses.
- *Computational efficiency and scalability constraints:* The foundation models typically contain millions or billions of parameters making full fine-tuning computationally expensive and challenging for many applications. This is further amplified in multimodal models, which require multiple data streams to be processed simultaneously. The recent advancements in parameter-efficient fine tuning methods such as adapters, LoRA, and prompt tuning have emerged to address these challenges. But the effectiveness of these techniques vary significantly based on the task and domain adaption scenario [13]. Different downstream tasks require different optimal adapters or combination of multiple adapters, making adapter selection a complex optimization problem.
- *Domain-specific knowledge injection barriers:* MFMs encounter vocabulary limitations and modality mismatches while incorporating niche domain knowledge. In healthcare applications, CLIP fails to recognize specialized radiology terms even after prompt engineering, requiring specialized auxiliary knowledge graphs for concept alignment. This is particularly due to lack of niche medical terminology in the original training dataset. The gap particularly arises because the clinical concepts exist in hierarchical, context-dependent relationships that MFMs cannot learn from unstructured text-image pairs alone [14]. To address this, hybrid approaches combines frozen MFM backbones with domain specific tokenizers and utilizes knowledge-graphs for concept alignment [15]. However, building accurate medical knowlege graphs is labor-intensive, and their integration with MFM's embedding space adds computational overhead.

- *Multimodal synergy under utilization* Improving target domain performance while preserving cross-modal generalization capabilities is a challenge in MFM adaptation. Recent studies show domain-adapted CLIP variants lose original cross-modal retrieval performance despite improving target task accuracy [10]. This occurs because adaptation focuses on improving intra-modal alignment than preserving inter-modal relationships. The UniMos framework addresses this by maintaining separate modality adaptation streams with contrastive losses, recovering cross-modal performance while achieving top adaptation metrics [16].

2.2 Conceptual foundations of multimodal adaptation: Domain shift, modality discrepancy, and distribution alignment dynamics

In unimodal settings, domain adaptation aligns the distributions of labeled source domain $\mathcal{D}_s$ and a related but distinct (often unlabeled) target domain $\mathcal{D}_t$. This is done within a single modality (for example, images or text). While *multimodal* domain adaptation handles not only alignment within each modality but also aligns shifts between modalities. This section describes three core concepts that are part of most multimodal domain-adaptation frameworks.

2.2.1 Domain shift

In a unimodal setting, domain shift relates to any change in data distribution between a source domain $\mathcal{D}_s$ and a target domain $\mathcal{D}_t$. If X is the input (e.g., an image), then

$$P_s(X) \neq P_t(X). \quad (2)$$

This could happen due to differences in lighting, sensor properties, or style variations in text. In a multimodal setting, data samples are usually pairs (or tuples) $(x^{\text{img}}, x^{\text{text}})$, and domain shift can happen at following two levels:

1. Intra-modal:

$$P_s(X^{\text{img}}) \neq P_t(X^{\text{img}}), \quad P_s(X^{\text{text}}) \neq P_t(X^{\text{text}}). \quad (3)$$

For example, source images might be high-resolution photographs, while target images come from smartphone snapshots. Similarly, source captions may be grammatically correct, whereas target captions use colloquial social-media language.

2. Inter-modal (cross-modal): Even if each modality is aligned individually, the *joint* distribution could differ.

$$P_s(X^{\text{img}}, X^{\text{text}}) \neq P_t(X^{\text{img}}, X^{\text{text}}) \quad (4)$$

For instance, in the source domain, images and captions may be manually curated and paired (e.g., “A red bicycle leaning against a wall”), while in the target domain, captions might be noisy or only loosely related (e.g., “bike life,” “city vibes”). Such differences in alignment in cross-modal interactions leads to a type of domain shift unique to multimodal settings.

2.2.2 Modality gap

Modality discrepancy occurs due to the fact that different modalities inherently reside in *heterogeneous feature spaces* [17]. An image represented by pixels resides in a convolutional feature space, while text is represented by token sequences embedded in a vector space. The properties of these two representations could differ drastically in terms of sparsity, noise characteristics, and clustering behavior.

The modality gap can be formally defined as the difference between the centers of embeddings from different modalities.

$$\vec{\Delta}_{gap} = \frac{1}{n} \sum_{i=1}^n x_i - \sum_{i=1}^n y_i \quad (5)$$

where x_i and y_i represent embeddings from two different modalities. Research has revealed that in models like CLIP, different data modalities are embedded “at arm’s length” in their shared representation space.

- *Within-domain semantic gap*: Even within the same domain, distinct encoders are learned for each modality. For example, image and text require learning encoders f_{img} and f_{text} . The encoders are trained to decrease a distance metric between semantically matching pairs in a shared latent space.
- *Cross-domain and cross-modal shift*: When domain shift occurs, an encoder trained on source images could generate features that drift drastically on target images. Therefore, the joint embedding $[f_{img}(x^{img}), f_{text}(x^{text})]$ learned on the source will no longer preserve semantic similarity on the target. This requires techniques that reduces the gap not just between each modality’s feature spaces but also bridges the gap between how modalities co-occur across domains.

2.2.3 Distribution alignment

As discussed in the previous section, in a multimodal setting, it is essential to minimize the difference between source and target domain feature distributions while preserving cross-modal relationships. The goal is to learn features that are domain-invariant and modality-invariant. Dong et al. mention distribution alignment as one of the core domain adaptation techniques [2]. In multimodal models such as CLIP, alignment occurs in shared embedding space learned using image encoder and text encoder. In domain adaptation, this idea is extended to also align source and target distributions (while aligning the modalities) by minimizing a divergence metric or adversarially confusing a domain discriminator.

Given source distribution $P_s(X, Y)$ and target distribution $P_t(X, Y)$, the distribution alignment seeks transformation ϕ such that

$$\phi(P_s) \approx \phi(P_t) \quad \text{while} \quad I(X_{mod1}; X_{mod2}) \geq \epsilon \quad (6)$$

where I represents mutual information between modalities.

In a unimodal settings, few common alignment methods such as adversarial losses (e.g., domain-adversarial neural networks), moment matching (e.g., CORAL, MK-MMD), or reconstruction losses are applied to align the source and target distributions.

Multimodal alignment needs to work on multiple modalities *simultaneously* while preserving cross-modal association. Let us take a case of multimodal alignment between image and text modalities to discuss the steps involved in domain adaptation.

1. *Per-modality alignment followed by cross modal alignment:*

Per-modality alignment: Initially a DANN-style adversarial loss is applied using adversaries D_{img} and D_{text} separately on image and text embeddings to align each marginal distribution:

$$P_s(f_{\text{img}}(X^{\text{img}})) \approx P_t(f_{\text{img}}(X^{\text{img}})) \quad (7)$$

$$P_s(f_{\text{text}}(X^{\text{text}})) \approx P_t(f_{\text{text}}(X^{\text{text}})) \quad (8)$$

Joint (cross-modal) alignment: A third adversarial discriminator D_{joint} is introduced that takes fused embeddings

$$Z = [f_{\text{img}}(x^{\text{img}}), f_{\text{text}}(x^{\text{text}})] \quad (9)$$

and encourages

$$P_s(Z) \approx P_t(Z). \quad (10)$$

This step ensures that cross-modal relationships between image and text features remain consistent across domains.

2. *Alignment of conditional distributions:* After aligning marginals, many tasks also require aligning class-conditional distributions

$$P_s(Z|y) \approx P_t(Z|y), \quad (11)$$

where Z is the joint embedding. For classification tasks, aligning class centroids or matching class-wise moment helps preserve discriminative structure.

3. Theoretical foundations of multimodal domain adaptation

3.1 Mathematical formulation of multimodal domain shift

Multimodal domain adaptation assumes *distributional divergence* between the source domain $\mathcal{D}_s$ and target domain $\mathcal{D}_t$. Each domain comprises M modalities. Let $\mathbf{X}_s = \left\{ \left(\mathbf{x}_s^{(1)}, \dots, \mathbf{x}_s^{(M)} \right) \right\}$ represent the multimodal source data with labels $\mathbf{Y}_s$ and $\mathbf{X}_t = \left\{ \left(\mathbf{x}_t^{(1)}, \dots, \mathbf{x}_t^{(M)} \right) \right\}$ represent the unlabeled target data. The joint distribution discrepancy is expressed in eq. 12.

$$P_s(\mathbf{X}, Y) \neq P_t(\mathbf{X}, Y) \quad (12)$$

This shift appears asymmetrically across different modalities, with *modality-specific domain gaps* defined by eq. 13.

$$\Delta_m = \mathbb{E}_{\mathbf{x}_t^{(m)}} [\phi_m(\mathbf{x}_t^{(m)})] - \mathbb{E}_{\mathbf{x}_s^{(m)}} [\phi_m(\mathbf{x}_s^{(m)})] \quad (13)$$

where ϕ_m represents modality-specific feature extractors [2].

The objective minimizes the *target risk* $\mathcal{R}_t(h)$ for hypothesis h given by eq. 14.

$$\mathcal{R}_t(h) = \mathbb{E}_{(\mathbf{x}_t, y_t) \sim P_t} [\ell(h(\mathbf{x}_t), y_t)] \quad (14)$$

Under the *covariate shift* assumption ($P_s(Y|\mathbf{X}) = P_t(Y|\mathbf{X})$), the ideal joint risk combines source empirical risk $\hat{\mathcal{R}}_s(h)$ and distribution discrepancy $\text{disc}(\mathcal{D}_s, \mathcal{D}_t)$ is given by eq. 15

$$\mathcal{R}_t(h) \leq \hat{\mathcal{R}}_s(h) + \text{disc}(\mathcal{D}_s, \mathcal{D}_t) + \lambda^* \quad (15)$$

where λ^* is the optimal joint risk achievable by any hypothesis [2, 18].

3.2 Modality-specific divergence metrics

In this section, we discuss some of the commonly used metrics to learn shared feature spaces in multimodal domain adaptation.

3.2.1 Maximum mean discrepancy (MMD)

For modality m , MMD measures distribution divergence by comparing mean embeddings in a reproducing kernel Hilbert space (RKHS) given by eq. 16.

$$\text{MMD}_m^2 = \left\| \frac{1}{n_s} \sum_{i=1}^{n_s} \phi_m(\mathbf{x}_{s,i}^{(m)}) - \frac{1}{n_t} \sum_{j=1}^{n_t} \phi_m(\mathbf{x}_{t,j}^{(m)}) \right\|_{\mathcal{H}}^2 \quad (16)$$

MMD-based alignment is generally used in visual-language adaptation to reduce cross-modal discrepancies [2, 19].

3.2.2 Reverse KL divergence

The asymmetric KL divergence $KL[P_t^{(m)} \| P_s^{(m)}]$ leads to *mode-seeking alignment*. It prioritizes regions where the target distribution $P_t^{(m)}$ has high density. It avoids areas where $P_s^{(m)}$ has near-zero probability, thereby preserving dominant modes in the target domain [20].

$$KL_m = \mathbb{E}_{\mathbf{x}_t^{(m)}} \left[\log \frac{P_t^{(m)}(\mathbf{x}_t^{(m)})}{P_s^{(m)}(\mathbf{x}_t^{(m)})} \right] \quad (17)$$

3.2.3 Modality-conditioned Wasserstein distance

Wasserstein metrics measures minimum cost associated with transforming one distribution into another while preserving geometric relationships

between samples. The Wasserstein metric for modality m is given by the following equation.

$$\mathcal{W}_m = \inf_{\gamma \in \Gamma(P_s^{(m)}, P_t^{(m)})} \mathbb{E}_{(\mathbf{x}_s^{(m)}, \mathbf{x}_t^{(m)}) \sim \gamma} \left[c(\mathbf{x}_s^{(m)}, \mathbf{x}_t^{(m)}) \right] \quad (18)$$

where $c(\cdot)$ is the transportation cost function [2].

3.3 Information-theoretic perspectives

3.3.1 Information bottleneck for multimodal adaptation

The objective maximizes predictive information while minimizing redundancy. The multimodal information bottleneck objective maximizes the quantity in eq. 19.

$$\mathcal{L} = I(Y; \mathbf{Z}) - \beta \sum_{m=1}^M I(\mathbf{X}^{(m)}; \mathbf{Z}) \quad (19)$$

where $\mathbf{Z}$ denotes latent representations and β balances prediction fidelity against modality redundancy. It compresses input modalities $\mathbf{X}^{(m)}$ into a latent code $\mathbf{Z}$ that only retains label-relevant information [21–23]. This framework forces the model to retain only modality-invariant predictive features [24]. The parameter β controls tradeoff between discarding noisy modality-specific details and preserving discriminative features.

3.3.2 Multimodal Jensen-Shannon divergence

The mmJSD unifies multimodal distributions via adaptive mixing.

$$\text{mmJSD} = \sum_{m=1}^M \pi_m KL \left[P^{(m)}(\mathbf{Z}) \parallel \sum_{k=1}^M \pi_k P^{(k)}(\mathbf{Z}) \right] \quad (20)$$

where π_m represent modality weights. The modality-specific distributions are balanced using dynamic weights π_m , prioritizing reliable modalities. It enables sampling from joint distribution $p(\mathbf{z}) = \prod_{m=1}^M p(\mathbf{z}^{(m)})$ while maintaining cross-modal consistency [25].

3.3.3 Conditional mutual information bounds

The generalization error in adaptation is bounded by the quantity in eq. 21.

$$\mathcal{R}_t \leq \mathcal{R}_s + \sqrt{\frac{I(\mathbf{Z}_s; \mathbf{Z}_t)}{2n} + \frac{\log(2/\delta)}{2n}} \quad (21)$$

where $I(\mathbf{Z}_s; \mathbf{Z}_t)$ quantifies latent space alignment. Minimizing this mutual information term enforces domain-invariance in latent representations. It forces latent representations $\mathbf{Z}_s$ (source) and $\mathbf{Z}_t$ (target) to overlap [26]. The orthogonality constraints reduces redundancy in encoding modality-specific features [27].

4. Core methodologies in multimodal domain adaptation

4.1 Adversarial alignment techniques

Adversarial alignment is one of the most commonly used techniques to reduce domain shifts in multimodal systems [28, 29]. It involves *cross-modal domain discriminators* which are neural networks trained to distinguish between source and target domain features for each modality. These discriminators operate along with *gradient reversal layers (GRLs)*. The GRLs invert gradient signals during backpropagation to learn domain-invariant features. Let us consider a multimodal system with two modalities m_1 and m_2 . The adversarial objective for modality m is formulated by eq. 22.

$$\mathcal{L}_{\text{adv}}^{(m)} = -\mathbb{E}_{\mathbf{x}_s^{(m)}} \left[\log D_m \left(\phi_m \left(\mathbf{x}_s^{(m)} \right) \right) \right] - \mathbb{E}_{\mathbf{x}_t^{(m)}} \left[\log \left(1 - D_m \left(\phi_m \left(\mathbf{x}_t^{(m)} \right) \right) \right) \right] \quad (22)$$

where D_m is the modality-specific domain discriminator and ϕ_m is the feature extractor.

Recent advancements extend the adversarial alignment paradigm. For instance, *Modality-aware adversarial networks* dynamically weighs the domain discrepancies. The CMDA (cross-modal domain adaptation) framework [30] uses LiDAR-camera fusion. It involves adversarial training within and across modalities. It balances domain discrepancies across modalities through entropy minimization and adversarial regularization. It employs separate domain-specific batch normalization (DSBN) layers for each modality to preserve domain-specific characteristics while aligning shared features. These innovations help the model achieve higher mAP on multiple tasks compared to unimodal baselines.

4.2 Contrastive learning approaches

Another commonly used approach to reduce modality shift is through contrastive learning [31]. It bridges the modality gaps through *multimodal positive/negative sampling*. Unlike unimodal contrastive learning, multimodal variants need to handle heterogeneous negative samples. An image-text pair might form a positive sample, while text-audio pairs from different classes can form the negatives. The TupleInfoNCE loss formalizes this in eq. 23.

$$\mathcal{L}_{\text{contrast}} = -\log \frac{\exp(f(\mathbf{x}_i, \mathbf{y}_i)/\tau)}{\sum_{j=1}^K \exp(f(\mathbf{x}_i, \mathbf{y}_j)/\tau) + \exp(f(\mathbf{x}_j, \mathbf{y}_i)/\tau)} \quad (23)$$

where $f(\cdot)$ measures cross-modal similarity and τ is a temperature parameter.

Modality-aware projection heads reduce the representation discrepancies through specialized transformations for each modality type [32]. These transformations map each modality's features into a shared feature space while preserving modality-specific characteristics. The CrossCLR framework [33] uses separate projection layers for video, audio, and text, reducing the modality gap by 37% on the Youcook2 dataset. Hard negative mining strategies further enhance discriminability. It uses techniques like *semantic-hard negatives* to sample negatives with similar class labels but different modalities. It forces the model to learn finer-grained discrimination.

4.3 Cross-modal interaction strategies

Teacher-student prototype alignment also encourages effective multimodal adaptation. This method aligns cluster centroids (prototypes) of target domain features with source domain centroids across modalities [34]. For vision-language systems, the alignment loss between image prototypes $\mathbf{C}_v$ and text prototypes $\mathbf{C}_t$ is given by eq. 24.

$$\mathcal{L}_{\text{align}} = \sum_{k=1}^K \|\mathbf{C}_v^{(k)} - \mathbf{C}_t^{(k)}\|_2^2 \quad (24)$$

where K is the number of semantic classes. The MMAct framework [35] shows this approach improves cross-session activity recognition accuracy compared to feature-level alignment alone.

Dynamic feature fusion networks modulate cross-modal interactions based on domain characteristics. The DynMM architecture [36] introduces gated fusion units that calculate modality weights α_m using domain-specific attention mentioned in eq. 25.

$$\alpha_m = \sigma\left(\mathbf{W}_m \left[\phi_m\left(\mathbf{x}^{(m)}\right) \oplus \mathbf{h}_{\text{domain}} \right]\right) \quad (25)$$

The encoder generates a representation of domain shift characteristics $\mathbf{h}_{\text{domain}}$. It is concatenated with the input features. This allows the model to dynamically weight reliable modalities under specific domain shifts in adaptation tasks, and such dynamic fusion is shown to perform better in adaptation tasks than static fusion techniques [37].

These methods jointly address the challenges of domain shift, modality gap, and distribution alignment. Adversarial alignment directly targets domain differences, while contrastive learning bridges modality gaps, and dynamic interaction strategies enable context-aware adaptation. They form a complete toolkit for effective multimodal domain adaptation.

5. Advanced adaptation paradigms in multimodal systems

5.1 Test-time adaptation frameworks

Test-time adaptation (TTA) is increasingly being used to handle domain shifts during deployment. Traditional domain adaptation techniques require source data access. TTA works under stricter constraints that include only using unlabeled target data while avoiding any impact to real-time performance [38].

Online modality re-weighting uses dynamic attention mechanisms to address asymmetric domain shifts across modalities. The multi-modal continual test-time adaptation (MM-CTTA) [38] framework uses a dual-stage process. It uses class-wise feature centroids to determine reliability scores for each modality in real time. For LiDAR-camera fusion problems, this approach reduces catastrophic forgetting compared to static fusion networks [39]. Loss reweighting with α -tuned hyper-parameters have been used by CrossTrainer [40] to automatically prioritize modalities with lower domain discrepancy. It achieves faster convergence than brute-force search methods.

Source-free adaptation techniques have been developed to handle scenarios where source data are unavailable due to privacy or logistic constraints. SUMMIT [41] uses

cross-modal pseudo-label refinement for source-free adaptation. It combines agreement filtering for high-confidence predictions with entropy weighting for ambiguous cases. This dual technique improves semantic segmentation on autonomous driving datasets while using only pretrained uni-modal source models. Further, source-free domain adaptation via distribution estimation (SFDA-DE) [42] enhances it by performing spherical k-means clustering in feature spaces to estimate source distribution parameters from target data statistics. This method is shown to cause fewer labeling errors than SHOT baseline.

5.2 Foundation model integration

The introduction of large scale vision-language foundation models have influenced multimodal adaptation strategies. Techniques such as *prompt tuning* and *parameter-efficient fine tuning* are increasingly used for adapting these models to domain adaptation problems. OPTIMA [43] uses domain-aware soft prompts to smoothen the decision boundaries between source and target distributions. It is shown to reduce annotation needs especially in low-resource NLP tasks. LoRA adapters [44] are used in vision transformers to inject trainable rank decomposition matrices in attention layers. It helps with efficient adaptation of large models. It uses only a fraction of tunable weights while maintaining the full fine-tuning performance.

Cross-modal alignment prompts are used in multimodal foundation models like CLIP to project target domains into shared embedding spaces. MMAAct framework [35] dynamically detects any domain shifts in real-time and adjusts visual-textual prompts. This improves the activity recognition accuracy in wearable sensor applications. In medical imaging, a similar work using contrastive prompt pools [45] that aligns radiology reports with MRI sequences is shown to improve zero-shot transfer across multiple anatomical regions.

5.3 Open-set and partial domain adaptation

Models deployed in real-world application often come across unseen classes in target domains. A number of framework address such *open-set scenarios*. The SPLOS framework [46] uses cross-domain mixup augmentation along with self-paced learning to continuously adapt known to unknown classes using entropy-based curriculum scheduling. This approach is shown to perform better in novel class detection without any degradation on shared categories in VisDA benchmarks [41].

The inverse challenges are addressed in *Partial domain adaptation*. It handles scenarios where unseen classes in target domain are present in source domain. PADA [47] uses adversarial discriminators with class-wise attention gates to reduce irrelevant source features. It reduces negative transfer in cross-country facial recognition tasks. The G-SFDA method [42] improves on it by applying geometric consistency constraints that ensures target features cluster within source-derived manifolds while removing outlier dimensions.

New approaches such as ProDe [48] attempt to unify open-set and partial adaptation through proxy confidence estimation. It automatically tunes the modality-specific rejection thresholds by modeling the divergence between vision-language model predictions and adapter feature spaces. It achieves state-of-the-art performances on both partial and open-set conditions.

These advanced techniques attempt to balance computational efficiency, data privacy, and performance robustness in complex real-world scenarios.

6. Standardized datasets for multimodal adaptation

There are few key benchmarks that capture diverse domain shift scenarios across multiple modalities.

6.1 EPIC-KITCHENS-100 UDA

The *EPIC-KITCHENS-100 UDA CHALLENGE* [49] contains video-based adaptation between temporally separated domains (2018 vs. 2020 recordings). It contains 100-hours of egocentric (first-person) videos capturing daily kitchen activities in 45 different environments. It spans across 97 verb and 300 noun classes. The standardized splits contain about 36 K source videos and 10 K target videos. The source domain (EPIC-KITCHENS-55) is fully labeled, while the target domain (EPIC-KITCHENS-100) is unlabeled. The mixed sequences protocol evaluates model's ability to handle compositional shifts in kitchen activities.

6.2 xMUDA splits (nuScenes and A2D2)

For 3D semantic segmentation, xMUDA framework [50] offers standardized dataset to learn cross-modal adaptation in autonomous driving. It leverages nuScenes and A2D2 datasets. nuScenes contains 1000 driving scenes with LiDAR point clouds and RGB images. It is split into multiple outdoor settings: day/night based on lightning condition and Boston/Singapore based on geographical differences in road layout and vehicles. An important insight from this benchmark is the asymmetric degradation of various modalities. Although LiDAR remains stable across different lighting conditions, RGB performance degrades in night lighting. The splits provides a standard benchmark for evaluating adaptation robustness and consistency across different environment and hardware setups.

6.3 Human-animal-cartoon (HAC)

The HAC dataset [3] provides a standard dataset for action recognition tasks. It contains 3381 video clips showing seven core actions (sleeping, eating, running etc.) by humans, animals, and cartoon characters. It also provides audio and optical flow modalities for different actions. It offers benchmark to evaluate the consistency of cross-domain models through the Style Invariance Index (SII). It enables models to learn invariant visual representations that preserve semantic fidelity under artistic domain shifts.

6.4 AudioSet

AudioSet [51] provides a standard benchmark for evaluating sound classification and audio-visual alignment under environmental acoustic shifts. It contains a large collection of 10-second audio clips from YouTube. It has human-annotated labels for 632 sound categories (e.g., speech, music, and ambient noise). It offers variations in recording conditions (indoor vs. outdoor) and noise interference. It is useful in developing noise-invariant audio models and benchmarking environmental adaptation in multimodal systems.

7. Conclusion

In this chapter, we explored the various challenges and complexities of multimodal domain adaptation (MMDA) arising from adapting the models across distinct domains and modalities. MMDA extends the adaptation challenges observed in unimodal settings and requires specialized techniques that handle domain shifts, modality gaps, and distribution alignment simultaneously. The techniques developed in MMDA try to address the domain shift challenges while preserving the association between modalities. The advent of multimodal foundation models such as CLIP has introduced new avenues to research in the field. It offers opportunities to improve zero-shot adaptation and cross-modal generalization. Introduction of large-scale datasets such as EPIC-KITCHENS-100 UDA and xMUDA offer essential benchmarks for evaluating MMDA performance under diverse constraints for multiple adaptation tasks.

The recent advancements in the field have opened a number of diverse directions for research. We list few of them as follows:

- *Efficient foundation model adaptation frameworks*: There is a growing interest in frameworks that address the computational challenges with large MFMs in real-world deployment scenarios. Developing efficient adaptation frameworks could potentially unlock various practical uses of these models. Further research in the theoretical insights from information bottleneck governing MFM could advance optimal knowledge distillation techniques [52].
- *Causal reasoning behind cross-modal invariance*: Most of the current techniques in MMDA focus on statistical alignment. Further research in temporal causal models could uncover new techniques to disentangle domain-specific and modality-invariant features in multimodal data.
- *Ethical and secure MMDA*: There is an increasing demand for privacy-preserving adaptation techniques especially in sensitive domains such as healthcare and surveillance. Further research on differential privacy techniques and federated MMDA frameworks could offer new insights into this [53].
- *Open-world adaptation*: There is growing interest in combining test-time adaptation with lifelong learning mechanisms to handle evolving multimodal distributions in real-time [2]. This requires the development of new evaluation frameworks that can quantify catastrophic forgetting rates and transfer efficiency between modalities.

Author details

Raghavendran Ramakrishnan
IQVIA, Raleigh, USA

*Address all correspondence to: rramak11@asu.edu

IntechOpen

© 2025 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

References

- [1] Baltrusaitis T, Ahuja C, Morency L-P. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2019;41(2):423-443
- [2] Advances in Multimodal Adaptation and Generalization: From Traditional Approaches to Foundation Models. Available from: <https://arxiv.org/html/2501.18592v1>
- [3] Dong H, Nejjar I, Sun H, Chatzi E, Fink O. SimMMDG: A simple and effective framework for multi-modal domain generalization. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2023
- [4] Liu Y, Qiao L, Changchong L, Yin D, Lin C, Peng H, et al. OSAN: A one-stage alignment network to unify multimodal alignment and unsupervised domain adaptation. In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Vancouver, BC, Canada: IEEE; 2023. pp. 3551-3560
- [5] Munro J, Damen D. Multi-modal domain adaptation for fine-grained action recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020. pp. 122-132
- [6] Damen D, Doughty H, Farinella GM, Fidler S, Furnari A, Kazakos E, et al. The EPIC-KITCHENS dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2021;43(11): 4125-4141
- [7] Park KR, Lee HJ, Kim JU. Learning trimodal relation for audio-visual question answering with missing modality. In: *European Conference on Computer Vision*. Springer; 2024. pp. 42-59
- [8] Schäfer R, Nicke T, Höfener H, Lange A, Merhof D, Feuerhake F, et al. Overcoming data scarcity in biomedical imaging with a foundational multi-task model. *Nature Computational Science*. 2024;4(7):495-509
- [9] Woerner S, Baumgartner CF. Navigating data scarcity using foundation models: A benchmark of few-shot and zero-shot learning approaches in medical imaging. In: *International Workshop on Foundation Models for General Medical AI*. Springer; 2024. pp. 30-39
- [10] Li J, Li Y, Yali F, Liu J, Liu Y, Yang M, King I. CLIP-Powered Domain Generalization and Domain Adaptation: A Comprehensive Survey. *arXiv preprint arXiv:2504.14280*, 2025
- [11] Wang H, Yang X, Chang J, Jin D, Sun J, Zhang S, Luo X, Tian Q. Parameter-efficient tuning of large-scale multimodal foundation model. *Advances in Neural Information Processing Systems*. 2023;36:15752-15774
- [12] Lai Z, Vedapant N, Zhou N, Jun W, Huynh CP, Li X, et al. PADCLIP: Pseudo-labeling with adaptive Debiasing in CLIP for unsupervised domain adaptation. In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. Paris, France: IEEE; 2023. pp. 16109-16119
- [13] Liu F, Zhang T, Dai W, Cai W, Zhou X, Chen D. Few-shot adaptation of multi-modal foundation models: A survey. *Artificial Intelligence Review*. 2024;57(10):268
- [14] Zhao Z, Liu Y, Han W, Wang M, Li Y, Wang S, et al. CLIP in medical

imaging: A survey. *Medical Image Analysis*. 2025;**102**:103551

[15] Zhang X, Chaoyi W, Zhang Y, Xie W, Wang Y. Knowledge-enhanced visual-language pre-training on chest radiology images. *Nature Communications*. 2023;**14**(1):4542

[16] Li X, Li Y, Zhekai D, Li F, Ke L, Li J. Split to merge: Unifying separated modalities for unsupervised domain adaptation. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE; 2024. pp. 23364-23374

[17] Liang W, Zhang Y, Kwon Y, Yeung S, Zou J. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*. 2022; **35**:17612-17625

[18] Kouw WM, Loog M. An introduction to domain adaptation and transfer learning. *arXiv preprint arXiv:1812.11806*. 2018

[19] Liang T, Lin G, Feng L, Zhang Y, Lv F. Attention is not enough: Mitigating the distribution discrepancy in asynchronous multimodal sequence fusion. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, QC, Canada: IEEE; 2021. pp. 8128-8136

[20] A Tuan Nguyen, Toan Tran, Yarin Gal, and Philip H S Torr. KL Guided Domain Adaptation. 2022.

[21] Tishby N, Pereira FC, Bialek W. The information bottleneck method. *arXiv preprint physics/0004057*. 2000

[22] Tishby N, Zaslavsky N. Deep learning and the information bottleneck principle. In: 2015 IEEE Information

Theory Workshop (ITW). IEEE; 2015. pp. 1-5

[23] Kawaguchi K, Deng Z, Xu J, Huang J. How Does Information Bottleneck Help Deep Learning? In *Proceedings of the 40th International Conference on Machine Learning*. PMLR; July 2023. pp. 16049-16096

[24] Mancisidor RA, Kampffmeyer M, Aas K, Jenssen R. Discriminative multimodal learning via conditional priors in generative models. *Neural Networks*. 2024;**169**:417-430

[25] Sutter T, Daunhawer I, Vogt J. Multimodal Generative Learning Utilizing Jensen-Shannon-Divergence. In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc.; 2020. pp. 6100-6110

[26] Gholami B, Sahu P, Rudovic O, Bousmalis K, Pavlovic V. Unsupervised multi-target domain adaptation: An information theoretic approach. *IEEE Transactions on Image Processing*. 2020; **29**:3993-4002

[27] Luo L, Hu S, Yang J, Chen L. Decision boundary optimization-informed domain adaptation. *arXiv preprint arXiv:2502.06498*. 2025

[28] Ganin Y, Lempitsky V. Unsupervised domain adaptation by backpropagation. In: *International Conference on Machine Learning*, PMLR; 2015. pp. 1180-1189

[29] Pei Z, Cao Z, Long M, Wang J. Multi-adversarial domain adaptation. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32. 2018

[30] Chang G, Roh W, Jang S, Lee D, Ji D, Gyeongrok O, Park J, Kim J, Kim S. CMDA: Cross-modal and domain adversarial adaptation for LiDAR-based

3D object detection. In Proceedings of the AAAI Conference on Artificial Intelligence. 2024;38:972-980

[31] Liu Y, Fan Q, Zhang S, Dong H, Funkhouser T, Yi L. Contrastive multimodal fusion with TupleInfoNCE. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021:754-763

[32] Alayrac J-B, Recasens A, Schneider R, Arandjelović R, Ramapuram J, De Fauw J, Smaira L, Dieleman S, Zisserman A. Self-supervised multimodal versatile networks. Advances in Neural Information Processing Systems. 2020; 33:25-37

[33] Zolfaghari M, Zhu Y, Gehler P, Brox T. CrossCLR: Cross-modal contrastive learning for multi-modal video representations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021. pp. 1450-1459

[34] Belal A, Meethal A, Romero FP, Pedersoli M, Granger E. Multi-source domain adaptation for object detection with prototype-based mean teacher. In: 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Waikoloa, HI, USA: IEEE; 2024. pp. 1266-1275

[35] Kong Q, Wu Z, Deng Z, Klinkigt M, Tong B, Murakami T. MMAct: A large-scale dataset for cross modal human action understanding. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE; 2019. pp. 8657-8666

[36] Xue Z, Marculescu R. Dynamic Multimodal Fusion. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Vancouver, BC, Canada: IEEE; 2023. pp. 2575-2584

[37] Wang Y, Chu HK, Sun Y. PEA Fusion: Parameter-efficient adaptation for RGB-thermal fusion-based semantic segmentation. Information Fusion. 2025;120:103030

[38] Cao H, Xu Y, Yang J, Yin P, Yuan S, Xie L. Multi-modal continual test-time adaptation for 3D semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023. pp. 18809-18819

[39] Shin I, Tsai Y-H, Zhuang B, Schuler S, Liu B, Garg S, et al. MM-TTA: Multi-modal test-time adaptation for 3D semantic segmentation. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA: IEEE; 2022. pp. 16907-16916

[40] Chen J, Gan E, Rong K, Suri S, Bailis P. CrossTrainer: Practical domain adaptation with loss reweighting. In: Proceedings of the 3rd International Workshop on Data Management for end-to-end Machine Learning, DEEM'19. New York, NY, USA: Association for Computing Machinery; 2019. pp. 1-10

[41] Simons C, Raychaudhuri DS, Ahmed SKM, You S, Karydis K, Amit K. Roy-Chowdhury. SUMMIT: Source-free adaptation of uni-modal models to multi-modal targets. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2023. 1239-1249

[42] Ding N, Yixing X, Tang Y, Xu C, Wang Y, Tao D. Source-free domain adaptation via distribution estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022. pp. 7212-7222

[43] Xu G, Li B, Yu H. Improving the sample efficiency of prompt tuning with

domain adaptation. arXiv preprint arXiv: 2210.02952. 2022

[44] Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, Wang L, Chen W. LoRA: Low-rank adaptation of large language models. ICLR. 2022;1(2):3

[45] Zhou K, Yang J, Loy CC, Liu Z. Conditional prompt learning for vision-language models. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA: IEEE; 2022. pp. 16795-16804

[46] Grangier D, Ablin P, Hannun A. Adaptive training distributions with scalable online bilevel optimization. arXiv preprint arXiv:2311.11973. 2023

[47] Cao Z, Ma L, Long M, Wang J. Partial adversarial domain adaptation. In: Proceedings of the European Conference on Computer Vision (ECCV). 2018:135-150

[48] Tang S, Su W, Gan Y, Ye M, Zhang J, Zhu X. Proxy denoising for source-free domain adaptation. arXiv preprint arXiv: 2406.01658. 2024

[49] Cheng Y, Fang F, Sun Y. Team VI-I2R technical report on EPIC-KITCHENS-100 unsupervised domain adaptation challenge for action recognition 2021. arXiv preprint arXiv: 2206.02573. 2022

[50] Jaritz M, Tuan-Hung V, De Charette R, Wirbel E, Perez P. xMUDA: Cross-modal unsupervised domain adaptation for 3D semantic segmentation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE; 2020. pp. 12602-12611

[51] Gemmeke JF, Ellis DPW, Freedman D, Jansen A, Wade Lawrence R, Moore C, et al. Audio set:

An ontology and human-labeled dataset for audio events. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New Orleans, LA: IEEE; 2017. pp. 776-780

[52] Dong H, Chatzi E, Fink O. Towards multimodal open-set domain generalization and adaptation through self-supervision. In: European Conference on Computer Vision. Springer; 2024. pp. 270-287

[53] Küchemann S, Avila KE, Dinc Y, Hortmann C, Revenga N, Ruf V, et al. On opportunities and challenges of large multimodal foundation models in education. NPJ Science of Learning. 2025;10(1):1-11

Section 2

Knowledge and Data Representation

Good Data Beats More Data: Building a Thermal Air-Image Dataset for Ground-to-Air Surveillance

Tijeni Delleji, Feten Slimeni and Ahmed Siala

Abstract

The efficiency of transfer learning for detecting drones using thermal imagery is highly dependent on the quality and diversity of the underlying training datasets. This chapter presents a comprehensive and reproducible pipeline for constructing high-quality aerial thermal datasets optimized for transfer learning applications. We detail the end-to-end process, encompassing thermal image acquisition, precise annotation, domain-specific augmentation, and dataset refinement through advanced cleaning techniques. To ensure robust model transferability across real-world scenarios, we emphasize the inclusion of environmental diversity—spanning diurnal cycles, weather variability, and heterogeneous landscapes such as urban and rural areas. We introduce novel thermal-specific augmentation strategies, including thermal blur, noise injection, and dynamic brightness modulation, to bridge the domain gap between pretraining and target deployment. Additionally, cleaning procedures using perceptual hashing and statistical outlier detection are employed to maintain dataset integrity and eliminate redundancy. Experimental evaluations confirm that datasets processed with our methodology yield superior performance in drone detection tasks, significantly enhancing accuracy, robustness, and generalization across varying conditions. This chapter thus provides practical insights and guidelines for researchers and practitioners aiming to leverage transfer learning in thermal imaging contexts, contributing to more reliable and scalable drone detection systems for ground-to-air surveillance.

Keywords: thermal imaging, ground-to-air surveillance, drone detection, dataset construction, data annotation, data cleaning, deep learning, infrared object detection

1. Introduction

The rapid proliferation of mini-drones has introduced a host of new security challenges. These compact and highly maneuverable aerial devices are increasingly being exploited for unauthorized surveillance, smuggling, and potential weaponization. As

a result, their detection has become a critical concern for military, governmental, and civilian infrastructures alike [1, 2].

Real-time drone detection is especially vital in sensitive environments such as military bases, no-fly zones, and critical infrastructure, where any breach can have severe consequences [3–5]. Developing robust detection systems that can accurately identify and track mini-drones in diverse operational conditions has, therefore, become a top priority.

In this context, thermal imaging has emerged as a powerful modality for aerial surveillance. Unlike conventional optical cameras that rely on visible light, thermal imaging systems capture infrared radiation emitted by objects, allowing drones to be identified by their heat signatures. This capability is particularly advantageous in low-light or visually obstructed environments—such as during night operations or under conditions involving smoke, fog, or camouflage [6–10].

Thermal air imagery—capturing scenes from a ground-to-air perspective—has thus become a valuable resource for drone detection systems. However, this modality also introduces new challenges. Mini-drones often share thermal characteristics with other heat-emitting objects like birds, debris, or background noise, making accurate identification a non-trivial task, particularly for fast-moving and small targets.

To address these challenges, recent advances in deep learning—particularly object detection frameworks such as the YOLO (You Only Look Once) series—have shown promise in real-time thermal detection [11–13]. Modern YOLO architectures, including YOLOv8 through YOLOv11, have significantly improved in detection speed and precision [14–17]. Nevertheless, the performance of these models is highly influenced by the quality of the training dataset.

In this context, transfer learning plays a pivotal role, allowing pre-trained models to be adapted to targeted datasets (e.g., thermal drone imagery), significantly reducing training time and data requirements while enhancing accuracy. Yet, the success of transfer learning is contingent on the availability of clean, diverse, and well-annotated datasets that accurately reflect the target deployment environment.

In this chapter, we introduce a practical and systematic pipeline for constructing optimized thermal aerial datasets, purpose-built for transfer learning applications in drone detection. We cover key stages including thermal image acquisition across diverse conditions, precise data annotation, domain-specific augmentation (e.g., thermal blur, noise injection), and rigorous data cleaning using perceptual hashing and outlier detection.

By following this pipeline, researchers and practitioners can significantly enhance the generalization and robustness of drone detection models fine-tuned on thermal imagery. The chapter concludes with experimental insights, demonstrating how improved dataset construction strategies lead to measurable gains in detection accuracy and cross-condition adaptability.

The following sections of this chapter are organized as follows: Section 2 reviews relevant literature and recent advancements in thermal drone detection. Section 3 describes the dataset creation process, covering data collection, annotation, augmentation, and preprocessing. Section 4 presents our experimental framework, analyzing dataset characteristics and evaluating model performance. Finally, Section 5 summarizes key findings and suggests future research directions for transfer learning in thermal drone detection.

2. Related work

The effectiveness of drone detection systems using thermal imaging depends on two critical factors: robust algorithmic frameworks—whether developed from

scratch, through fine-tuning, or via transfer learning—and high-quality training datasets. While prior research has extensively explored algorithmic advancements, the construction of optimized thermal datasets remains underexplored, despite being a key determinant of model performance.

2.1 Thermal imaging in drone detection

Thermal imaging has become a fundamental tool for drone detection due to its proven resilience under low-visibility conditions. Unlike optical sensors, thermal cameras capture infrared radiation, allowing for the reliable identification of drone heat signatures even in challenging scenarios such as darkness, fog, or smoke [8]. Early studies, such as Quan et al. [18], demonstrated the feasibility of using convolutional neural networks (CNNs) to distinguish drones from birds in thermal images. More recently, Soeleman et al. [10] integrated deep learning models with thermal data, achieving improved accuracy in cluttered environments. However, these studies have primarily focused on algorithmic enhancements, with limited emphasis on data-set quality, environmental diversity, and standardization—factors that are equally critical for achieving robust and generalizable performance.

2.2 Challenges in thermal dataset construction

Most existing drone detection datasets are based on RGB imagery [6], making them ill-suited for thermal-based applications due to sensor noise, low contrast, and thermal crossover effects. Publicly available thermal datasets, such as FLIR ADAS [19] and KAIST Multispectral [20], provide thermal-RGB paired data but lack dedicated aerial drone annotations. Additionally, custom datasets like [3] suffer from a limited scale (<5000 images) and fail to incorporate diverse environmental conditions, reducing their applicability to real-world surveillance.

Key limitations observed in prior thermal drone detection datasets:

- *Sensitivity to diurnal and seasonal variations:* Thermal signatures can fluctuate significantly depending on the time of day and season, impacting detection consistency.
- *Lack of multi-altitude perspectives:* Most datasets fail to capture data from varying flight altitudes, which is crucial for robust drone detection across different operational scenarios.
- *Presence of adversarial environmental conditions:* Challenging factors such as *solar glare, precipitation, and complex thermal backgrounds* are often underrepresented, limiting real-world applicability.

2.3 Algorithm: Dataset synergy for drone detection

The evolution of *YOLO-based architectures* (YOLOv8–YOLOv11) has significantly enhanced *small-object detection capabilities* [11, 15, 16]. However, these advancements often rely on access to *high-quality, well-curated datasets*. Dellej et al. [7] demonstrated that the *precision of YOLO models dropped significantly* when trained on *non-augmented imaging datasets*, emphasizing the strong *interdependence between dataset quality and model performance*. This underlines the *necessity of a holistic approach* that

integrates *dataset curation, data cleaning, augmentation strategies, and real-world variability*—an aspect largely overlooked in prior literature.

Our work addresses these gaps by proposing a *standardized pipeline for thermal dataset creation*, incorporating:

- *Diverse environmental conditions* to enhance generalization across scenarios.
- *Data cleaning procedures* to eliminate mislabeled, corrupted, or low-quality samples that can hinder model performance.
- *Physics-based augmentations* to simulate real-world variability in thermal data.
- *Robust annotation validation protocols* to improve annotation quality and consistency.

2.4 Public thermal imaging datasets

Existing public thermal datasets primarily focus on *automotive, pedestrian, or general object detection*, with *limited emphasis on aerial drone detection*. Below is an overview of some of the most widely recognized thermal imaging datasets:

- *FLIR ADAS [19]*: A large-scale thermal dataset designed for *automotive applications*, featuring annotated thermal images of *vehicles, pedestrians, and cyclists*. Despite its size, it lacks *dedicated aerial drone annotations* and does not address challenges such as *thermal crossover at varying altitudes*.
- *KAIST multispectral [20]*: A benchmark dataset providing *aligned RGB-thermal image pairs*, primarily for *pedestrian detection*. However, its *annotations and scenarios are irrelevant* to drone detection and do not account for *thermal variations across different weather conditions or altitudes*.
- *LITIV dataset [21]*: Focused on *low-light and thermal object tracking*, this dataset includes *dynamic outdoor scenes* but lacks *systematic drone annotations* or consideration of *adversarial conditions* such as *solar glare or precipitation*.
- *VAP thermal aerial dataset [22]*: A small dataset containing *thermal aerial imagery of birds and UAVs*. However, it suffers from a *limited scale* (fewer than 2000 frames) and *minimal environmental diversity*, restricting its *usefulness for training robust drone detection models*.
- *ARL thermal dataset [23]*: Developed for *military applications*, this dataset includes thermal sequences of *vehicles and humans* in cluttered environments. While it addresses some *thermal crossover challenges*, it lacks *multi-altitude drone perspectives* and does not capture *seasonal variations*.

While these datasets provide foundational support for thermal imaging research, they fall short in systematically addressing the key requirements for aerial drone detection. This highlights the urgent need for a dedicated thermal drone detection benchmark that accounts for diverse environmental conditions, varying altitudes, and challenging real-world scenarios. To bridge this gap, our study introduces a

comprehensive, well-annotated, and publicly available thermal aerial imagery dataset, specifically tailored for ground-to-air drone detection.

3. Construction pipeline of the thermal dataset

The development of our dataset, *Thermal Air-Image Dataset for Ground-to-Air Surveillance*, follows a well-defined and systematic pipeline comprising the following stages:

1. *Raw data collection*: Thermal imagery is captured using an EO/IR imaging system [24], ensuring high-quality input data suitable for surveillance applications.
2. *Data augmentation*: To improve diversity and robustness, the collected images undergo augmentation techniques such as the addition of blur and noise, simulating varied environmental and sensor conditions.
3. *Annotation*: Objects of interest within the thermal images are meticulously labeled to provide accurate and reliable ground-truth data for training and evaluation.
4. *Data cleaning*: The dataset is refined by removing duplicate entries and correcting inconsistencies, thereby preserving data integrity and consistency.
5. *Dataset partitioning*: The final dataset is systematically divided into training, validation, and test subsets, ensuring balanced distribution and effective model evaluation.

A more detailed description of each step is provided below.

3.1 Data collection

A crucial element for any deep learning-based system is acquiring high-quality training and validation data. Our data collection methodology employed a sophisticated Electro-Optical and Infrared (EO/IR) imaging system for thermal data acquisition. The system's central component, illustrated in **Figure 1**, consists of the multisensor PTZ camera (model DH-TPC-PT8621C, TPC 2024). This advanced surveillance equipment incorporates a dual-lens architecture that enables synchronous capture of both visible spectrum and thermal infrared video streams. This all-in-one solution, equipped with sophisticated sensors, is particularly effective for long-distance outdoor surveillance, ensuring comprehensive data collection under varying conditions and significantly improving the detection accuracy and operational reliability.

The system integrates an RGB camera equipped with starlight technology and a motorized zoom lens, enabling high-resolution imaging and detailed close-up verification. For low-visibility conditions (day or night), a dedicated thermal imaging camera detects emitted infrared radiation, ensuring reliable surveillance regardless of lighting. An integrated laser illuminator paired with the RGB camera enhances night-time imaging, simulating reflected infrared capabilities.



Figure 1.
Multisensor network PTZ camera.

Table 1 presents the technical specifications of the thermal camera, including its VOX uncooled detector, $6.2^\circ \times 5.0^\circ$ field of view, ≤ 40 mK sensitivity, 100 mm focal length, $8\ \mu\text{m}$ to $14\ \mu\text{m}$ spectral range, 640×512 resolution, $17\ \mu\text{m}$ pixel pitch, and 25 fps frame rate.

This study utilized the imaging system to capture thermal videos of the three commercial DJI mini-UAVs listed in **Table 2**: the Phantom 4 Pro+ V2.0, Matrice 600, and Mavic 3. Our data collection strategy involved recording video sequences during diverse daylight conditions, varying illumination levels, and multiple challenging settings to accurately replicate operational environments. The thermal recordings, taken at 25 frames per second (FPS), were saved in MP4 format. From these recordings, a selection of key images was extracted to form the dataset’s foundation, ensuring exposure to diverse thermal signatures and environmental conditions during model training.

3.2 Enhancing dataset robustness *via* thermal data augmentation

Thermal data augmentation is a critical technique used to enhance the performance of neural network models in thermal imaging by artificially increasing the

Parameter	Specification
Detector type	VOX uncooled focal plane detector
Field of view	$6.2^\circ \times 5.0^\circ$
Thermal sensitivity (NETD)	≤ 40 mK
Focal length	100 mm
Spectral range	$8 - 14\ \mu\text{m}$ (LWIR)
IR resolution	640×512
Pixel pitch	$17\ \mu\text{m}$
Video frame rate	25fps

Table 1.
Technical characterization of the Thermal Camera.

Parameter	DJI Mavic 3	DJI Phantom 4 Pro + V2.0	DJI Matrice 600
Size	8.3 × 8.3 × 3.3 cm	19.6 × 29 × 29 cm	166.8 × 151.8 × 75.9 cm
Weight	730 g	1375 g	9100 g
Diagonal Length	38.1 cm (15 in)	35.6 cm (14 in)	113.3 cm (44.6 in)
Payload	Camera (up to 1 kg)	Camera (up to 1.5 kg)	6 kg (cameras, LiDAR, etc.)
Max flight time	46 minutes	30 minutes	35 minutes
Max speed	75 kph (47 mph)	72 kph (45 mph)	65 kph (40 mph)
Max range	15 km (9.3 mi)	7 km (4.3 mi)	5 km (3.1 mi)
Application	Photography, cinematography	Photography, mapping	Industrial, surveying, cinematography

Table 2.
 Technical characterization of the considered min-UAVs.

diversity of the training data without the need to collect new samples. Given the unique characteristics of thermal images, data augmentation is vital in ensuring that the model is robust against various environmental and operational variations. By expanding the training set with new, plausible examples, the detection model becomes better equipped to handle the complexities of real-world scenarios when deployed.

In this study, we applied several thermal-specific data augmentation techniques using the online computer vision development framework, Roboflow [25]. These techniques include:

1. *Thermal BLUR*: Introducing a blur effect to thermal images helps the model develop resistance to variations in camera focus. This can be mathematically represented as the convolution of the thermal image $I(x, y)$ with a Gaussian kernel $G(x, y, \sigma)$:

$$I_{\text{blur}}(x, y) = I(x, y) * G(x, y, \sigma) \quad (1)$$

Which is defined as:

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2 + y^2}{2\sigma^2}} \quad (2)$$

In this equation, σ controls the standard deviation, influencing the degree of blurriness introduced to the image.

2. *Temperature brightness variability*: Adjusting the brightness in thermal images simulates different temperature conditions, enabling the model to be more robust to changes in environmental temperature and camera settings. This can be modeled as:

$$I_{\text{brightness}}(x, y) = I(x, y) + \Delta T \cdot k \quad (3)$$

where ΔT is a constant representing the temperature change, and k is a scaling factor that adjusts the intensity of the brightness modification based on the pixel's temperature.

3. *Thermal noise*: Adding noise to thermal images mimics the artifacts and distortions often encountered in thermal imaging due to sensor imperfections or environmental factors. The noise can be represented by:

$$I_{\text{noise}}(x,y) = I(x,y) + N(x,y) \quad (4)$$

where $N(x,y)$ is a noise function drawn from a Gaussian distribution $N(0, \sigma_n^2)$ defined as:

$$N(x,y) \sim \mathcal{N}(0, \sigma_n^2) \quad \text{for } (x,y) \in \Omega \quad (5)$$

Here, σ_n stands for the standard deviation of the noise, and Ω for the domain of the image.

4. *Thermal CutMix*: Similar to the traditional CutMix, this technique involves randomly cutting out a portion of the thermal image and replacing it with a segment from another image. The CutMix operation can be mathematically expressed as follows:

$$I_{\text{mix}}(x,y) = \begin{cases} \lambda I_1(x,y) + (1-\lambda) I_2(x',y'), & \text{if } (x,y) \in \text{CutOutRegion}, \\ I_1(x,y), & \text{otherwise.} \end{cases} \quad (6)$$

Where (x',y') represent the coordinates of the cutout region in I_2 , λ is a mixing parameter that controls the contribution of each image to the final output, and *CutOutRegion* refers to the portion of the original image that is replaced.

By applying these targeted augmentation techniques, we significantly improved the model's robustness and accuracy in detecting objects within thermal datasets. The enhanced diversity and variability in the training data prepare the model for deployment in diverse and challenging thermal imaging environments, allowing it to adapt effectively to real-world scenarios.

3.3 Thermal data annotation

Data annotation is a crucial process in deep learning (DL) applications, particularly for tasks involving identifying and categorizing objects based on their heat signatures in thermal images. This specialized form of annotation involves drawing bounding boxes, polygons, or other shapes around objects within thermal photos and assigning them specific labels, such as "drone," "bird," or "plane." Unlike traditional image annotation, which relies on visible light, thermal image annotation focuses on the infrared spectrum, where objects are detected based on the heat they emit. The accuracy and consistency of this annotation process are paramount, as the performance of DL models is highly dependent on the quality of the annotated data.

Reliable annotation requires a deep understanding of the unique thermal patterns associated with different objects and the ability to discern subtle temperature variations. Additionally, it is essential to ensure that only relevant features are annotated, excluding any thermal artifacts that do not belong to the target object classes. By meticulously labeling thermal data, annotators contribute to creating robust datasets critical for training DL models capable of accurately detecting and identifying objects in complex and challenging environments.

3.4 Thermal data cleaning

In optimizing our object detection model, we embraced the principle that “Good data beats more data,” as emphasized by Dedrone [26], emphasizing the quality and diversity of images in our dataset. This approach was especially crucial in thermal data, where image quality can significantly impact model performance. After meticulous data annotation, we undertook a thorough refinement process to ensure the integrity of our thermal dataset, which involved identifying and resolving issues such as conflicting labels, data duplicates, label correlation, and high filter correlation. A significant challenge we encountered was the presence of near-duplicates—visually similar images but not identical—which can lead to redundancy and biased results in machine learning datasets.

We employed perceptual hashing techniques specifically tailored for thermal images to tackle this. Perceptual hashing generates a hash based on the visual content of images, allowing for the detection of near-duplicates that exhibit subtle variations, such as slight changes in temperature gradients. We utilized the phash algorithm with a hash size of 8, which produces a compact representation of each image’s visual content. The process begins by applying the Discrete Cosine Transform (DCT) to convert the image into frequency components, enabling us to capture the essential features of the image while disregarding less relevant details. The DCT of an image I can be expressed mathematically as:

$$DCT(u,v) = \frac{1}{4} C(u) C(v) \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} I(x,y) \cos\left(\frac{(2x+1)u\pi}{2N}\right) \times \cos\left(\frac{(2y+1)v\pi}{2N}\right) \quad (7)$$

where $C(u)$ and $C(v)$ are normalization factors defined as:

$$C(u) = \begin{cases} \frac{1}{\sqrt{2}}, & \text{if } u = 0, \\ 1, & \text{otherwise.} \end{cases} \quad (8)$$

After computing the DCT coefficients, we derive a binary hash for each image by comparing these coefficients to the median value of the DCT array. The binary hash b is constructed as follows:

$$b_{i,j} = \begin{cases} 1, & \text{if } DCT(i,j) \geq \text{median}(DCT), \\ 0, & \text{otherwise.} \end{cases} \quad (9)$$

This results in a binary representation that succinctly captures the visual features of the image. To identify near-duplicates, we compare the binary hashes of different images. The Hamming distance d between two binary hashes b_1 and b_2 is calculated as:

$$d(b_1, b_2) = \sum_{i=1}^n |b_1[i] - b_2[i]| \quad (10)$$

where n is the length of the hash. If the Hamming distance d between two hashes falls below a predefined threshold T , we classify those images as near-duplicates:

$$\text{if } d(b_1, b_2) < T, \text{ then images are near-duplicates} \quad (11)$$

By efficiently identifying and removing these near-duplicates, we streamline our dataset, reducing redundancy and ensuring that each image contributes meaningfully to the training of our model. Overall, this phase-based detection of near-duplicates has proven to be a valuable tool for cleaning and organizing high-quality datasets, enabling us to enhance the performance of our thermal image object detection model.

3.5 Dataset splitting

Strategic dataset partitioning is crucial for developing models capable of generalizing to unseen data, with careful separation between training, validation, and test subsets. We curated 6000 sample images from the collected thermal videos, called “Seen data.” To create a robust training pipeline, these images were randomly divided into training and validation sets, with 4800 images (80%) allocated for training and 1200 images (20%) reserved for validation. This split was carefully chosen to maintain a balanced distribution, ensuring that the model receives a broad exposure during training while allowing for meaningful performance evaluation during validation.

Additionally, 600 images were extracted from videos that were not part of the training and validation sets to assess the model’s performance on entirely new scenarios. Following standard evaluation protocols, these images were reserved as an independent test set, explicitly labeled “unseen data” to simulate real-world deployment conditions. The splitting strategy is visually summarized in **Figure 2**, which illustrates the proportion of data allocated to training, validation, and testing. The “Seen data” (training and validation) is depicted with an 80% training set and a 20% validation set. In contrast, the “Unseen data” forms the testing set, ensuring a clear distinction between the datasets.

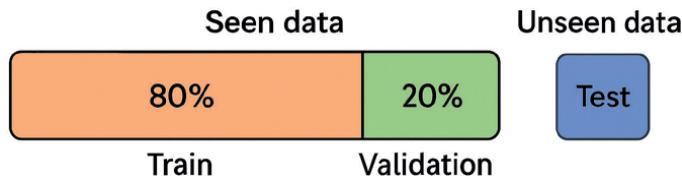


Figure 2.
Cross-thermal dataset: training, validation (seen data), and test (unseen data) sets.

The entire dataset, encompassing training, validation, and testing sets, was combined into a comprehensive “drone thermal dataset,” which will be publicly available for future research endeavors.

4. Experiments

4.1 Dataset exploration

Figure 3 displays a pair plot, illustrating the relationships between four variables commonly used in object detection tasks: x, y, width, and height of bounding boxes. The diagonal elements show histograms for each variable, revealing their distributions, with x and y appearing uniformly distributed, while width and height show skewed distributions toward smaller values. The off-diagonal scatter plots represent the pairwise relationships between these variables, with notable patterns such as a positive correlation between width and height. In contrast, other variable pairs exhibit more scattered distributions without clear correlations. This visualization aids in understanding the distribution and interdependencies of bounding box parameters in a dataset.

Figure 4 provides a detailed visualization of the thermal dataset labels. The top-left plot shows a label count for the “Drone” class, indicating approximately 2500 instances, while the top-right visualizes the overlap and distribution of bounding boxes, showing

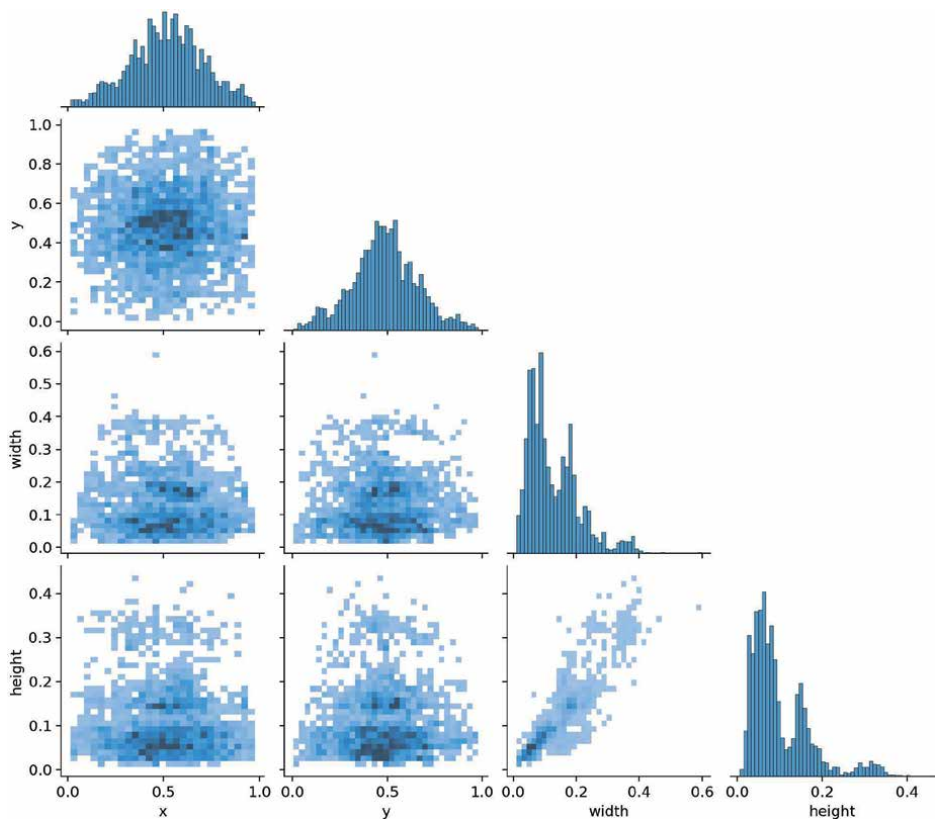


Figure 3.
correlogram of drone label dimensions: height vs. width correlation.

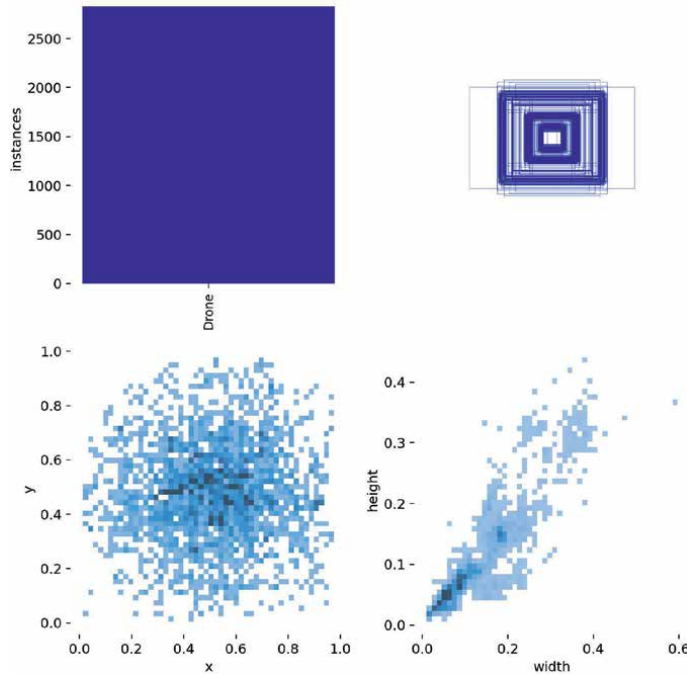


Figure 4.
Label volume and distribution for drones detection in our customized dataset.

how objects are positioned and sized within the images. The scatter plots below illustrate the relationships between bounding box attributes, such as x vs y coordinates, which are evenly spread, and width vs height, showing a positive correlation, meaning larger width tends to coincide with more considerable height. Overall, this figure offers insights into the spatial distribution and size variation of the bounding boxes within the dataset.

Figure 5 illustrates sample thermal air-images from the dataset, showcasing diversity across different lighting conditions, drone models, and environmental backgroundsthe diversity of thermal air-images within the customized dataset. Our dataset encompasses a wide range of ground-to-air scenarios, capturing various environmental conditions, altitudes, and object types. This diversity enhances its applicability for robust model training and real-world surveillance applications.

4.2 Performance analysis

This section evaluates the impact of data cleaning on model performance by analyzing dataset distribution and key evaluation metrics.

First, **Table 3** presents a comparison of the dataset size before and after the cleaning process. The number of images in both the training and validation sets decreases after cleaning, indicating the removal of potentially replicate or irrelevant data. This step ensures a more refined dataset, which can lead to improved generalization and model robustness.

Next, **Figure 6** illustrates the effect of data cleaning on precision across training steps using the YOLOv11 [17] model for object detection. The plotted curves for “Before Cleaning” and “After Cleaning” indicate similar precision, despite the dataset



Figure 5.
Sample thermal air-images from the dataset, showcasing diversity across different lighting conditions, drone models, and environmental backgrounds.

	Images number	Training set	Validation set
Before Cleaning	12,600	9600	3000
After Cleaning	10,200	7600	2600

Table 3.
Dataset split before and after cleaning.

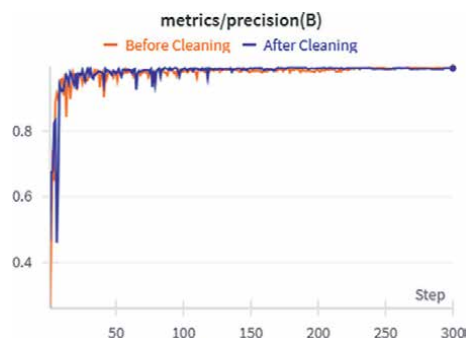


Figure 6.
Effect of data cleaning on precision across training steps in drone detection using YOLOv11.

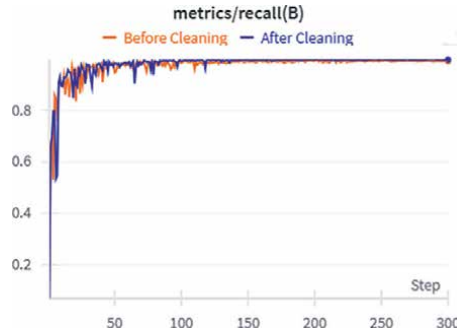


Figure 7.
Impact of data cleaning on recall performance in drone detection using YOLOv11.

before cleaning being larger (12,600 images) compared to after cleaning (10,200 images). This suggests that data cleaning effectively removes redundant or irrelevant samples without compromising the model's precision.

Similarly, **Figure 7** (recall analysis) compares the recall performance of the YOLOv11 model [17] before and after data cleaning. The recall metric is essential for determining how well the model captures true positive instances. Despite the cleaned dataset being smaller than the original, the recall remains similar, suggesting that the removed data had minimal impact on the model's ability to detect positive cases. This indicates that data cleaning effectively eliminates redundant or noisy samples without compromising recall performance.

Overall, this analysis provides insights into the effectiveness of dataset cleaning in refining the dataset by removing noisy or redundant samples while maintaining model precision and recall, thereby ensuring accuracy and consistency.

5. Conclusion

This study presents a systematic framework for constructing high-quality thermal air-image datasets to enhance drone detection in ground-to-air surveillance applications. By integrating diverse environmental conditions, rigorous quality control measures, and advanced augmentation techniques, the proposed dataset significantly improves detection accuracy, model robustness, and generalization. Key contributions include a comprehensive dataset pipeline covering data acquisition, annotation, filtering, and augmentation, along with a publicly available dataset designed to enhance real-world drone detection. Additionally, insights into optimizing thermal data for small-object detection are provided, ensuring its applicability in complex surveillance scenarios. Future research directions will involve: (1) dataset expansion incorporating diverse drone platforms, (2) performance evaluation under adversarial environmental conditions, and (3) integration of multi-sensor fusion approaches to enhance real-time detection robustness. This framework provides researchers with a comprehensive foundation for advancing thermal-based drone surveillance systems.

Acknowledgements

We gratefully acknowledge the support of the Tunisian Ministry of National Defense, particularly the Science and Technology for Defense Lab (STD) and the

Military Research Center (CRM). Their funding enables critical research and development efforts toward building an integrated Counter-Unmanned Aircraft System (C-UAS) defense solution.

Author details

Tijeni Delleji^{*†}, Feten Slimeni[†] and Ahmed Siala[†]
Military Research Center, Aouina Military Base, Taieb Mhiri City, Tunis, Tunisia

*Address all correspondence to: tijeni_dalleji@yahoo.fr

† These authors contributed equally.

IntechOpen

© 2025 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

References

- [1] Seidaliyeva U, Ilipbayeva L, Taissariyeva K, Smailov N, Matson ET. Advances and challenges in drone detection and classification techniques: A state-of-the-art review. *Sensors*. 2024;**24**(1):1-31. [Online]. Available from: <https://www.mdpi.com/1424-8220/24/1/125>
- [2] Delleji T, Slimeni F, Lafi MA. C4 software for anti-drone system. *Defence Science Journal*. 2024;**74**(5):635-642
- [3] Taha B, Shoufan A. Machine learning-based drone detection and classification: State-of-the-art in research. *IEEE Access*. 2019;**7**:138669-138682
- [4] Ramadan R, Emara AH, Al-Sarem M, Elhamahmy M. Internet of drones intrusion detection using deep learning. *Electronics*. 2021;**10**:10
- [5] Tu M, Le K, Phung MD. Object detection in thermal images using deep learning for unmanned aerial vehicles. *Computer Vision and Pattern Recognition*. 2024;**01**:687-692
- [6] Delleji T, Fkih H, Kallel A, Chtourou Z. Visual tracking of mini-UAVs using modified YOLOv5 and improved DeepSORT algorithms. In: *Proceedings of the 2022 6th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*. Piscataway, NJ, USA: IEEE; 2022. pp. 1-6
- [7] Delleji T, Chtourou Z. An improved YOLOv5 for real-time mini-UAV detection in no-fly zones. In: *Proceedings of the 2nd International Conference on Image Processing and Vision Engineering (IMPROVE)*. Setúbal, Portugal: SciTePress; 2022. pp. 174-181
- [8] Ding B, Zhang Y, Ma S. A lightweight real-time infrared object detection model based on YOLOv8 for unmanned aerial vehicles. *Drones*. 2024;**8**(9):1-22. [Online]. Available from: <https://www.mdpi.com/2504-446X/8/9/479>
- [9] Quan CY, Edmond OLW, Srigrarom S. Identification of drone thermal signature by convolutional neural network. In: *Proceedings of the 2021 International Conference on Unmanned Aircraft Systems (ICUAS)*. Vol. 2021. Piscataway, NJ, USA: IEEE; 2021. pp. 63-70
- [10] Soeleman M, Supriyanto C, Purwanto. Deep learning model for unmanned aerial vehicle-based object detection on thermal images. *Revue d'Intelligence Artificielle*. 2023;**37**(6):1441-1447
- [11] Delleji T, Slimeni F, Lafi M, Ayadi A, Chtourou Z. Deep sky monitoring system for mini-drone detection and tracking. In: *2023 IEEE International Conference on Advanced Systems and Emergent Technologies (IC_ASET)*. 2023. pp. 1-6
- [12] Delleji T, Slimeni F, Lafi M, Chtourou Z. Video data analytics dashboard for anti-drone system. In: *Proceedings of the 2023 International Conference on Cyberworlds (CW)*, Sousse, Tunisia. Piscataway, NJ, USA: IEEE; 2023. pp. 134-140
- [13] Khan MA, Menouar H, Eldeeb A, Abu-Dayya A, Salim FD. On the detection of unauthorized drones—Techniques and future perspectives: A review. *IEEE Sensors Journal*. 2022;**22**(12):11439-11455
- [14] Delleji T, Slimeni F. Rf-yolo: A modified yolo model for uav detection and classification using rf spectrogram images. *Telecommunication Systems*. 2025;**88**(1):33. DOI: 10.1007/s11235-025-01264-4

- [15] Wang C-Y, Yeh I-H, Liao H. Yolov9: Learning what you want to learn using programmable gradient information. ArXiv: abs/2402.13616. 2024. [Online]. Available from: <https://api.semanticscholar.org/CorpusID:267770251>
- [16] Wang A, Chen H, Liu L, Chen K, Lin Z, Han J, et al. Yolov10: Real-time end-to-end object detection. arXiv preprint arXiv:2405.14458. 2024
- [17] Khanam R, Hussain M. Yolov11: An overview of the key architectural enhancements. arXiv preprint. October 2024
- [18] Quan CY, Wong EO, Srigrarom S. Identification of drone thermal signature by convolutional neural network. In: Proceedings of the 2021 International Conference on Unmanned Aircraft Systems (ICUAS), Athens, Greece. Piscataway, NJ, USA: IEEE; 2021. pp. 63-70
- [19] I. FLIR Systems. Flir thermal dataset for automotive applications. 2019. Available from: <https://www.flir.com> [Accessed: February 14, 2025]
- [20] Hwang S, Park J, Kim N, Choi Y, Kweon IS. Multispectral pedestrian detection: Benchmark dataset and baseline. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA: IEEE; Jun 2015. pp. 1037-1045
- [21] Gade R, Moeslund TB. Thermal cameras and applications: A survey. Machine Vision and Applications. Jan 2014;25(1):245-262. DOI: 10.1007/s00138_013_0570_5
- [22] Berg A et al. Thermal aerial detection using deep learning. In: IEEE International Conference on Robotics and Automation (ICRA). 2018
- [23] Jacobs N et al. The ARL multimodal dataset for thermal and visible object detection. In: Proceedings of SPIE Defense + Security. Vol. 9841. Bellingham, WA, USA: SPIE; 2016. DOI: 10.1117/12.2223089
- [24] Dahua technology. 2024. Aailable from: <https://www.dahuasecurity.com/products/All-Products/Thermal-Cameras/Wizmind-Series/TPC-8-Series/TPC-PT8621C/> [Accessed: September 04, 2024]
- [25] Roboflow. Roboflow: Computer Vision Development Platform. 2023. Available from: <https://roboflow.com>
- [26] Dedrone. How Dedrone Just Delivered Its Fastest and Most Accurate Airspace Security Solution Yet. 2024. Available from: <https://www.dedrone.com/blog/how-dedrone-just-delivered-its-/fastest-and-most-accurate-/airspace-security-solution-yet> [Accessed: August 21, 2024]

Knowledge Aggregation in Autonomous Systems via Pre-Trained Model

Muhammad Waqas Nawaz, Wanquan Zhang, Jaspreet Kaur, Qammer H. Abbasi, Muhammad Imran and Olaoluwa Popoola

Abstract

Autonomous systems deployed across robotics, transportation, and aerial networks must operate under constraints of limited bandwidth, energy, and heterogeneous data. Centralized learning methods are often inadequate due to high latency, privacy concerns, and limited scalability. This chapter presents a unified framework for decentralized knowledge aggregation, leveraging pre-trained models (PTMs), federated learning (FL), and edge intelligence to address these challenges. PTMs serve as portable knowledge units that reduce training overhead and enable fast, context-aware inference across distributed agents. The framework spans multiple domains—including the Internet of Robotic Things (IoRT), Connected Vehicular Networks (CVNs), Unmanned Aerial Vehicle (UAV) swarms, and Intelligent Transportation Systems (ITS)—and introduces task-aligned mechanisms such as one-shot inference, temporal clustering, multi-hop model fusion, and infrastructure-based calibration. These techniques aim to optimize model accuracy, energy efficiency, and responsiveness across highly dynamic, resource-constrained environments. We evaluate our approach through both simulations and real-world testbeds, demonstrating improvements such as 15–20% gains in accuracy, 60% reductions in synchronization delay, and up to 70% savings in energy consumption over baseline models. The results validate PTM-assisted FL as a scalable and resilient strategy for real-time autonomous operations. However, we also discuss the practical challenges of PTM deployment, including device-level compute limitations and communication overhead. This work provides a modular blueprint for deploying collaborative, decentralized intelligence across edge-enabled ecosystems.

Keywords: knowledge aggregation, federated learning, pre-trained models, Internet of Robotic Things (IoRT), 6G edge networks, Unmanned Aerial Vehicles (UAVs), autonomous vehicles, multi-task learning, temporal clustering, privacy-preserving AI

1. Introduction

The field of autonomous systems is advancing rapidly, fueled by escalating demands for intelligent decision-making, real-time adaptability, and operational

efficiency [1]. As this momentum drives the deployment of robotic platforms and intelligent agents across diverse domains—ranging from urban transportation to aerial surveillance—the ability to aggregate and share knowledge efficiently emerges as a cornerstone for ensuring coordination, responsiveness, and scalability [2, 3]. However, traditional centralized learning paradigms often fall short in distributed environments due to high communication latency, elevated energy consumption, and data privacy concerns [4, 5]. Decentralized intelligence frameworks have emerged as effective alternatives to overcome these limitations, particularly those leveraging pretrained models (PTMs) for distributed knowledge sharing. PTMs enable edge-based inference and transfer learning, allowing autonomous agents to utilize shared intelligence without transmitting raw data [6, 7]. This shift toward localized, on-device learning is reinforced by advancements in edge computing and federated learning (FL), which support scalable and privacy-preserving collaboration across agents [8, 9].

Knowledge aggregation in autonomous systems involves integrating information from multiple sources to enhance decision-making capabilities. This process is crucial in diverse applications, ranging from autonomous vehicles (AVs) navigating urban landscapes to robotic fleets executing collaborative tasks. The effectiveness of knowledge aggregation depends on efficient communication, optimized resource allocation, and robust learning frameworks. Simultaneously, the evolution of fifth- and sixth-generation (5G/6G) wireless networks has unlocked new opportunities for scalable autonomy [10]. These technologies offer ultra-low latency and high-throughput communication, enabling real-time interaction among spatially distributed agents in dynamic and uncertain environments. Systems such as unmanned aerial vehicles (UAVs), autonomous ground vehicles (UGVs), and robotic swarms require continuous coordination and adaptive learning to function effectively [11, 12]. To support this level of autonomy, achieving low-latency, energy-efficient, and privacy-aware knowledge aggregation becomes imperative for sustaining robust and responsive operations [13, 14]. To address the practical challenges of decentralization, several domain-specific frameworks have been developed. In Connected Vehicular Networks (CVNs), FL has been integrated with V2X (vehicle-to-everything) and V2I to support real-time decision-making and adaptive control [15, 16]. The Federated Multi-task Temporal Clustering (FeMTC) framework builds on this foundation by clustering clients based on model update similarity, enabling personalized and efficient learning under non-independent and identically distributed (non-IID) data conditions [17]. Similarly, DECKS (Distributed Edge-Based Collaborative Knowledge Sharing) architecture supports distributed knowledge sharing among UAVs by being deployed in a decentralized manner, resulting in a 70% reduction in energy consumption and a 15% improvement in learning performance [18].

In the Internet of Robotic Things (IoRT), knowledge-driven unified model block sharing has been realized through the K-DUMBs architecture, which allows resource-constrained agents to perform one-shot inference using shared model blocks from more capable peers [19, 20]. This approach supports lightweight knowledge transfer for agents with limited sensing and computation capabilities [21]. Additionally, federated personalization techniques—such as multi-task learning and adaptive aggregation—have further enhanced model accuracy in heterogeneous settings [22, 23]. Complementary to PTMs and FL, sensor-guided calibration mechanisms improve model reliability. For example, periodic vehicle-2-infrastructure (V2I) calibration using high-precision OptiTrack systems aligns control models between roadside units and autonomous vehicles, correcting for dynamic drift and communication

delays [24]. This infrastructure enables safety-critical applications such as obstacle avoidance, cooperative startup acceleration, and high-precision trajectory planning [25, 26].

1.1 The role of pre-trained models

Pre-trained models serve as the backbone of knowledge aggregation, enabling autonomous systems to learn efficiently without extensive retraining. These models act as knowledge carriers, providing foundational learning that can be adapted to various environments. Pre-trained models are essential because: *Firstly*, they reduce the need for large-scale training by optimizing computational resources. *Secondly*, they enable incremental learning, where knowledge gained from previous tasks is leveraged for new challenges. *Thirdly*, they facilitate decentralized learning, improving privacy and security by limiting raw data exchanges.

Section 2 reviews related work. Section 3 formulates the problem and presents the proposed frameworks—K-DUMBs, FeMTC, DECKS, and V2I calibration. Section 4 offers a comparative summary across performance metrics and deployment scenarios. Section 5 concludes the chapter and discusses future directions.

2. Related work

Effective knowledge aggregation in autonomous systems hinges on the integration of multiple complementary paradigms [18]. FL offers a privacy-preserving solution for distributed model training by ensuring that raw data remains localized, while edge intelligence enhances real-time responsiveness by shifting computational workloads closer to the data sources [19]. Together, these approaches form the backbone of decentralized autonomy, enabling intelligent agents to learn collaboratively while minimizing latency, bandwidth usage, and privacy risks. Meanwhile, domain-specific architectures such as Connected Vehicular Networks (CVNs), Internet of Robotic Things (IoRT) platforms, and Intelligent Transportation Systems (ITS) present unique challenges and opportunities for knowledge fusion [20]. This section reviews the most relevant literature across these complementary domains, organized into two main perspectives: foundational learning strategies and their application across autonomous environments.

2.1 Federated and edge learning in autonomous systems

The widespread deployment of autonomous systems has intensified the demand for scalable and efficient knowledge aggregation strategies. While early solutions largely relied on centralized architectures, such designs have proven insufficient in dynamic, distributed environments due to inherent limitations such as high communication overhead, increased latency, and poor scalability [27]. In response, decentralized frameworks built upon Federated Learning (FL) have gained traction, enabling collaborative model training across agents while preserving data privacy by avoiding raw data transmission [28, 29]. Nonetheless, despite its advantages, traditional FL experiences performance degradation when confronted with challenges such as *non-IID* data, limited client availability, or imbalanced participation [30]. To overcome these limitations, edge intelligence has emerged as a complementary paradigm that shifts computational tasks closer to the data source, thereby reducing latency and

enhancing real-time responsiveness [31]. In this context, Federated Multi-task Learning (FMTL) provides an extension of FL that allows for client-specific model customization, often through client clustering techniques [32]. Reinforcement learning and adaptive client scheduling have also been explored to improve FL convergence and resilience under dynamic participation [33, 34].

2.2 Knowledge aggregation across application domains

In CVNs, real-time knowledge sharing is pivotal for enabling cooperative driving, traffic coordination, and effective decision-making in high-mobility environments [35, 36]. Recent research has explored the use of gNodeB assistance and clustering strategies to mitigate issues such as temporal drift and sparse communication, thereby improving model convergence. Despite these advancements, the practical integration of Pre-Trained Models (PTMs) and personalized Federated Learning (FL) in vehicular scenarios remains underdeveloped. To address this gap, UAV-assisted edge intelligence has emerged as a promising solution, wherein aerial nodes equipped with trained models act as relay agents and knowledge disseminators. These UAVs can dynamically offload and distribute pre-trained inference models to ground vehicles in coverage-constrained or disconnected zones, supporting real-time adaptation without necessitating complete model retraining. As summarized in **Table 1**, this integration of aerial and terrestrial intelligence effectively extends the FL pipeline across heterogeneous domains, thereby improving knowledge continuity and system reliability in sparse or infrastructurally limited urban corridors.

In the IoRT domain, knowledge sharing becomes especially critical as many robotic agents may lack the computational or sensing capabilities to learn independently. Pre-trained model distribution and inference sharing frameworks are being explored to enable perception and control in these lightweight agents [47]. Transfer learning and secure model block distribution enable generalized inference while preserving privacy and reducing network load [48]. In parallel, Intelligent Transportation Systems (ITS) require tightly coordinated V2X and V2I communication to ensure real-time control and decision synchronization. Recent literature has investigated infrastructure-aided recalibration and cooperative sensing as means to correct drift and sensor inaccuracies [49]. However, end-to-end integration of FL, PTMs, and sensor-based recalibration into a unified, autonomous ITS framework is still underdeveloped. In addition, privacy and personalization continue to be active concerns in FL research. Solutions incorporating differential privacy, meta-learning, and resource-aware client optimization are being increasingly adopted to improve generalization, fairness, and system robustness [47–49].

2.3 Contributions

To address the challenges of latency, scalability, and personalization in autonomous systems, our work proposes a unified framework for knowledge aggregation that tightly integrates PTMs, FL, and edge intelligence across diverse application domains. We begin with the *K-DUMBs* framework in the IoRT, which enables lightweight, resource-constrained agents to perform one-shot inference using distributed model blocks from more capable peers. By bypassing the need for local training and minimizing communication overhead, *K-DUMBs* demonstrates how pre-trained knowledge can be dynamically transferred and reused across heterogeneous robotic agents [19]. Next, we introduce the *Federated Multi-task Temporal Clustering (FeMTC)*

Author	Summary
Benhelal et al. [37]	Proposed an edge controller to gather data from road-based IoT devices, with drones capturing event data and integrating extended reality for hazard notifications. <i>Strength</i> : Enables detection via fixed and aerial IoT. <i>Limitation</i> : Faces privacy concerns and communication failures. <i>Criteria</i> : Data accuracy, security, and communication topology.
Li et al. [38]	Introduced UAV-based flying base station to offload deep learning tasks. Reduced system overload and latency. <i>Limitation</i> : Depends on UAV reliability and connectivity. Focus: Latency, UAV availability.
Mashadi et al. [39]	Proposed collaborative trajectory learning with UAVs. Noted improved accuracy, but drone reliability and model alignment are challenging.
Saraswat et al. [40]	Integrated FL with blockchain among cooperative UAVs to enhance security in 6G UAV networks. Lacks scalability, discussion, and real-world deployment.
Khan et al. [41]	Created a digital twin model for FL with RSU-based knowledge caching. Challenges include outdated information and 3-layer hierarchy complexity.
Fu et al. [42]	Applied local model efficiency for AV selection. Achieved accuracy improvements but suffered from data inconsistency and high compute demand.
Zhou et al. [43]	Used hierarchical model aggregation across edge-cloud tiers in IoV. Noted improvements in data aggregation, but high overhead exists.
Qu et al. [44]	Introduced serverless FL with a one-hop UAV network. Efficient info sharing but lacked privacy/security focus.
Mehta et al. [45]	Explored 6G + UAV + blockchain integration in a conceptual architecture. Did not include empirical results.
Mosquet et al. [46]	Explored knowledge aggregation for AVs. Emphasized situational awareness but omitted privacy/latency metrics.
Nawaz et al. [18]	Privacy-preserving FL across decentralized UAVs for urban knowledge sharing. Enhances AV reliability and responsiveness by delivering edge intelligence to CVNs via swarm coordination. Focus: Privacy, latency, energy efficiency, and real-time adaptability.

Table 1.
Comparative analysis of UAV-assisted federated learning literature on edge knowledge sharing.

approach for CVNs. This framework leverages temporal cosine similarity to cluster clients based on their arrival time, enabling robust and personalized model learning under highly non-IID and dynamically changing data distributions. FeMTC significantly improves learning efficiency and client-specific performance in real-time driving environments [17]. Building upon edge intelligence, the *DECKS* architecture supports decentralized FL among UAV swarms in 6G-enabled networks. Hovering UAVs aggregate and relay knowledge to ground vehicles through collaborative training cycles. DECKS leverages PTMs updates, UAV-assisted coverage optimization, and device-to-device (D2D) communication to reduce energy consumption by 70% and improve accuracy by over 15% compared to conventional FL baselines [18]. Finally, for infrastructure-based autonomy, we propose a *periodic V2I calibration framework* that enhances control synchronization across connected vehicles in ITS. Using OptiTrack-guided roadside units, the system applies periodic corrections to pre-trained control policies, mitigating sensor drift and trajectory misalignment. This ensures reliable multi-vehicle coordination and real-time responsiveness in safety-critical scenarios [24]. Together, these contributions present a cohesive and modular blueprint for scalable knowledge aggregation across edge, aerial, and infrastructure

environments. By placing pre-trained models at the core of our framework, we demonstrate that distributed learning can be both adaptive and efficient, empowering autonomous agents to collaborate seamlessly in dynamic, high-mobility ecosystems.

3. Methodology overview

The methodological foundation of this chapter is built upon a modular and distributed learning architecture that leverages PTMs, FL, and domain-specific edge orchestration. Each proposed framework follows a task-aligned strategy: *K-DUMBs* employs model block transmission for inference-only agents in the Internet of Robotic Things (IoRT); *FeMTC* introduces a temporal cosine similarity-based clustering method for client grouping in Connected Vehicular Networks (CVNs); *DECKS* optimizes federated UAV collaboration through constrained multi-hop model sharing in 6G-enabled networks; and the *V2I calibration* mechanism ensures real-time control accuracy in ITS via periodic infrastructure-based corrections. Across all systems, decentralized optimization techniques are adopted, minimizing global loss functions while addressing non-IID data, communication constraints, and real-time requirements. Together, these methodologies demonstrate a coherent, extensible approach to edge-deployed, knowledge-driven autonomy.

3.1 Problem formulation

Knowledge aggregation is a fundamental concept in autonomous systems, allowing multiple agents to share, adapt, and utilize knowledge to enhance their decision-making capabilities. This section details the theoretical foundations of knowledge aggregation, detailing how it applies to IoRT, connected vehicular networks, V2X, and multi-UAV systems. We discuss various aggregation methods, including FL, temporal clustering, and edge-assisted knowledge fusion with vehicle-to-infrastructure (V2I). Each framework addresses unique system constraints while contributing toward a unified and scalable aggregation model. The detailed formulations below include modeling, equations, and performance outcomes validated through experiments.

3.2 K-DUMBs: Knowledge-driven unified model block sharing in IoRT

In IoRT systems, agents are often resource-constrained and unable to participate in full training cycles. The K-DUMBs framework enables “dumb” agents to perform inference by receiving globally trained model blocks from more capable agents or servers using a one-shot learning approach, as shown in **Figure 1**.

The core optimization objective is:

$$\min_{\theta} \sum_{k=1}^K \frac{n_k}{n} F_k(\theta). \quad (1)$$

where $F_k(\theta)$ is the local loss function at device k , and θ the global parameter vector. Instead of engaging in full communication rounds, K-DUMBs allows inference through shared weights:

$$\mathbf{z}_{t+1} = \mathbf{z}_t - \eta \sum_{k=1}^K \frac{n_k}{n} \nabla F_k(\mathbf{z}_t). \quad (2)$$

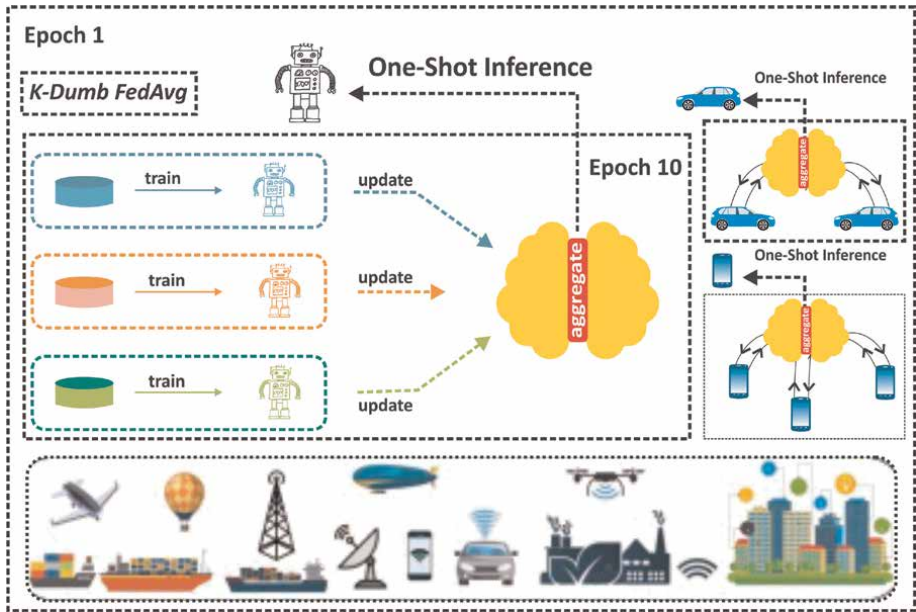


Figure 1.
Basic K-DUMB FedAvg architecture.

Evaluation and results: Experiments conducted on the MNIST dataset using Convolutional Neural Network (CNN) and Multilayer Perceptron (MLP) architectures demonstrate the efficiency of the K-DUMBs framework in one-shot deployment scenarios. **Figure 2** illustrates the average one-shot inference accuracy across communication rounds for the MNIST CNN. In this setting, only a subset of users was selected to participate in training, and the results show that the model reaches the target accuracy within a limited number of communication rounds, validating the effectiveness of K-DUMBs in low-overhead federated learning environments. Notably, K-DUMBs achieves accuracy comparable to traditional federated learning and centralized training—exceeding 90%—with only a single model broadcast.

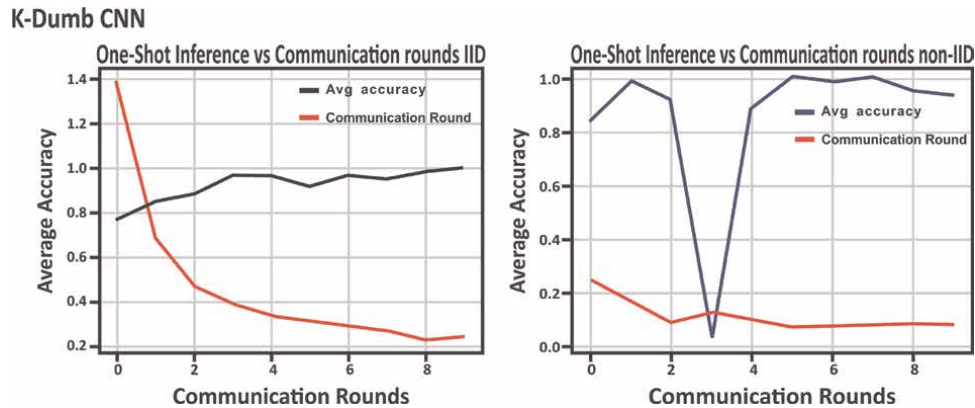


Figure 2.
Avg-acc in CNN against communication rounds.

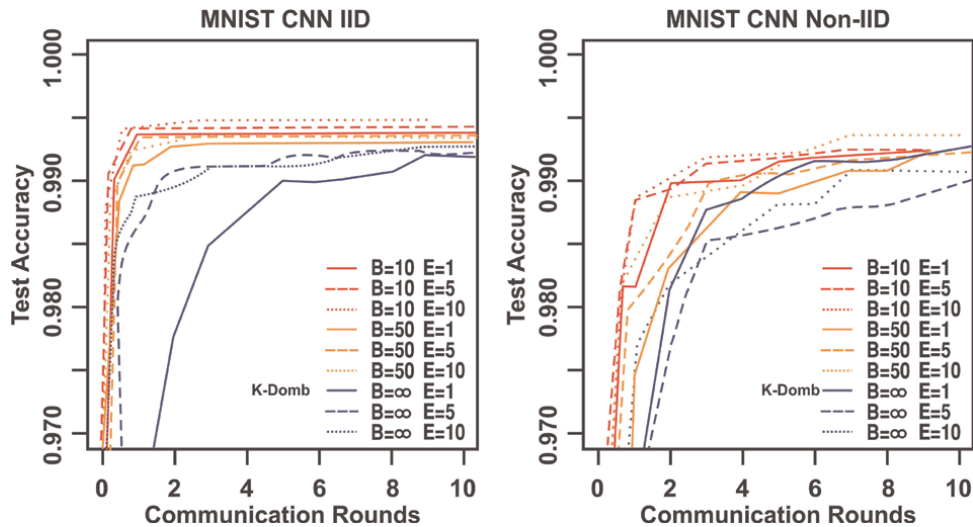


Figure 3.
Test-accuracy in MNIST CNN of IID and non-IID.

K-DUMBs FedAvg demonstrates the ability to produce high-quality inference results by sharing only model weights, thereby preserving data privacy and reducing communication overhead. As illustrated in **Figure 3**, the framework maintains consistent performance across both independent and identically distributed (IID) and non-IID data scenarios. Our experiments across various model topologies confirm that K-DUMBs FedAvg offers a practical solution for efficient knowledge sharing in IoRT environments. When partitioning MNIST under IID conditions, increased client-side computation enhances accuracy while reducing the number of communication rounds needed to reach convergence. Although performance under non-IID settings is comparatively less efficient, it remains significant. Across both model classes, K-DUMBs FedAvg achieves superior accuracy using the AI-enabled One-Shot Inference (O-SI) method, highlighting its robustness in decentralized, low-overhead learning setups. While our primary focus is on achieving standard performance metrics, the K-DUMBs FedAvg framework also proves effective at minimizing test loss—even in scenarios where test-set accuracy has plateaued. This dual capability is evident across both model classes, as illustrated by the communication rounds versus one-shot inference performance for MNIST using CNN and MLP architectures. As shown in **Table 2**, which presents average model accuracy against communication rounds, improving the effectiveness of knowledge sharing is closely tied to the degree of training access within the deployed models. That is, achieving convergence and maximizing accuracy

Model	Epoch	IID	Non-IID (equal)
MLP	10	91.25%	75.94%
CNN	10	97.17%	85.97%
MLP(One-shot Inf)	1	89.67%	74.89%
CNN(One-shot Inf)	1	95.47%	83.62%

Table 2.
Comparison between rounds of communication efficiency.

requires enabling more substantial access to the base model during training, especially in low-participation or non-IID settings.

3.3 Temporal hierarchical clustering in CVNs (FeMTC)

Vehicular clients in Connected Vehicular Networks (CVNs) generate streaming, time-sensitive data that is typically non-IID and affected by mobility patterns.

Figure 4 illustrates the system architecture showing base station (BS)-coordinated clustering and the temporal drift-aware aggregation process. The FeMTC framework leverages FL to such dynamics by using a two-phase clustering mechanism based on cosine similarity between temporal update vectors. In the first phase, similarity between clients' model updates is computed using a temporal cosine metric:

$$T(M_i^{(t)}, M_j^{(t')}) = \frac{\langle \Delta M_i^t, \Delta M_j^{t'} \rangle}{\|\Delta M_i^t\| \|\Delta M_j^{t'}\|} \cdot T_n \quad (3)$$

where ΔM_i^t is the model gradient or parameter update the client i at time t , and T_n is a time-normalization factor using recent weights update. The second phase aggregates updates within a cluster C_p to form a group-level model:

$$G_p = \frac{1}{|C_p|} \sum_{v_i \in C_p} M_i^{(t)} \cdot T_n \quad (4)$$

To further account for label heterogeneity, each client solves a local learning objective:

$$\min_{\theta_i} \mathcal{L}_i = \sum_{x \in D_i} \ell(y_i, f_{\theta}(x_i)) + \lambda \|\theta_i - G_p\|^2 \quad (5)$$

where $\ell(\cdot)$ is the task-specific loss and λ is a regularization term for cluster coherence.

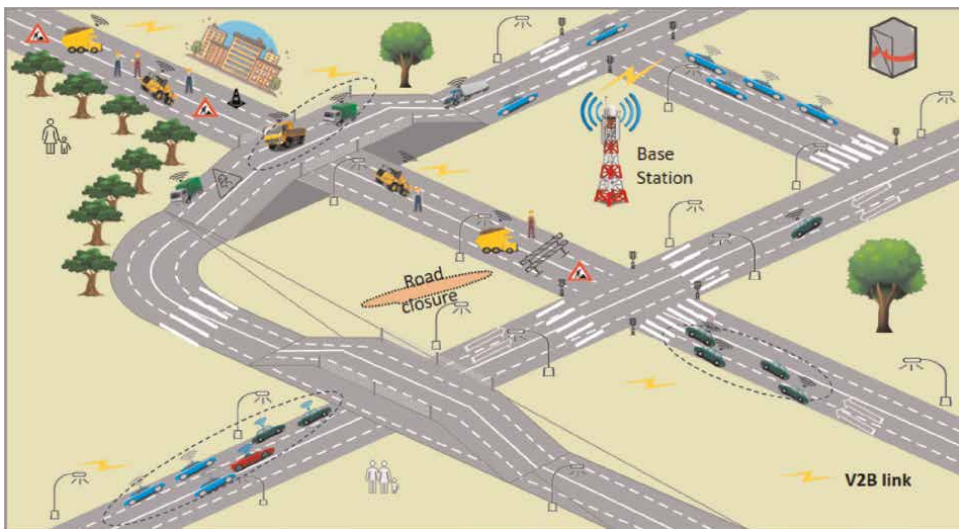


Figure 4.
 An overview of sparse connected VNs, including V2I communication.

Evaluation and results: The FeMTC framework was evaluated using CIFAR-10 and Fashion-MNIST datasets in a federated setting, simulating temporal non-IID distributions and cluster-based client behaviors. Label permutation and epoch-based scheduling were used to reflect the dynamic arrival patterns of clients in vehicular networks. The size of each client’s local dataset plays a pivotal role in the convergence dynamics of federated learning. To evaluate the robustness of our proposed method under varying local data conditions, we conducted a series of comparative experiments against FedAvg, testing environments with different dataset sizes per client. The findings consistently indicate that our approach achieves superior performance. By applying temporal hierarchical clustering (HC) during training, multiple models were generated at each time step—including a global FedAvg model and specialized models tailored to individual clusters. As shown in **Figure 5**, each client was assigned the model that achieved the highest validation accuracy on its local dataset, thereby maximizing personalization. **Figure 5** further demonstrates that under non-IID data conditions, FeMTC outperforms FedAvg by up to 20% in accuracy and achieves faster convergence across communication rounds.

To evaluate the adaptability of our method, we tested it on a new client that was not part of the original training dataset. The global parameter dispatched to this new client is computed as $B_i = \sum_{j \in C} C/iG_{ij}^{(t)} \cdot \theta_j$, where θ_j denotes the model parameters of cluster client j , and B_i represents the aggregated base model for client i . The server combines these parameters using relevance weights $G_p \times T_n$ before sending B_i to the new client for inference on unseen data. As illustrated in **Figure 6**, FeMTC-KS (based on pre-trained models) enables efficient knowledge transfer without retraining, significantly reducing computational overhead while preserving performance. **Figure 6**

	Client 0	Client 1	Client 2	Client 3	Client 4	Client 5	Client 6	Client 7	Client 8	Client 9
FL Model	0.678797	0.638199	0.591603	0.623824	0.598893	0.603945	0.638623	0.709428	0.623722	0.597328
Model 0	0.000000	0.000000	0.000000	0.000000	0.000000	0.746586	0.791587	0.785162	0.746421	0.694656
Model 1	0.802215	0.819876	0.757634	0.752351	0.781466	0.000000	0.000000	0.000000	0.000000	0.000000
Model 2	0.669304	0.658385	0.612595	0.620690	0.661134	0.575114	0.629063	0.690881	0.591002	0.538168
Model 3	0.800633	0.791925	0.769084	0.741379	0.782849	0.000000	0.000000	0.000000	0.000000	0.000000
Model 4	0.000000	0.000000	0.000000	0.000000	0.000000	0.763278	0.789675	0.788253	0.760736	0.711832

Figure 5.
Model personalization for each client in multi-task learning.

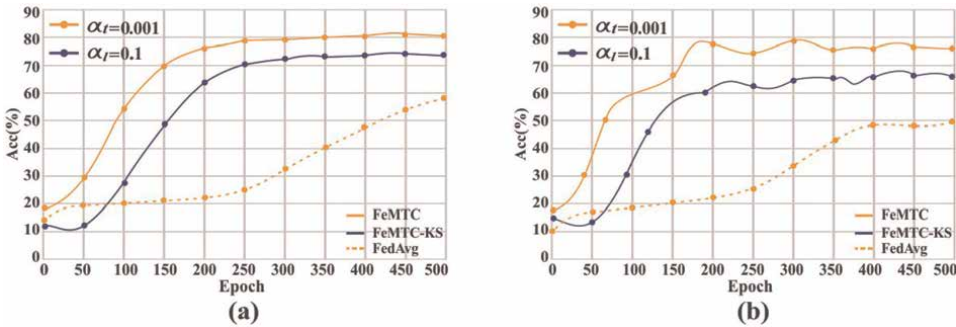


Figure 6.
Comparison of knowledge sharing in FeMTC: (a) fashion-MNIST (b) CIFAR-10.

further demonstrates the model's robustness, showing that FeMTC maintains consistent accuracy even under 50% client dropout. FeMTC maintains tighter performance variance and higher mean accuracy in environments where vehicle data distributions are skewed by mobility or application context (e.g., navigation vs. collision avoidance). This robustness makes FeMTC well-suited for real-time deployment in sparse and high-mobility vehicular networks.

3.4 DECKS architecture: Distributed UAV collaboration in 6G edge networks

Building on the foundational strategies introduced in K-DUMBs and FeMTC, which focused on PTMs reuse in the IoRT and client clustering in CVNs, respectively, the DECKS (Distributed Edge-based Collaborative Knowledge Sharing) framework extends knowledge aggregation into airborne, edge-assisted environments. It enables UAV swarms to collaboratively train, fuse, and relay federated models to connected vehicles, thereby improving coverage, inference responsiveness, and overall system resilience without depending on centralized base stations. The architecture is grounded in a UAV-assisted topology where aerial agents hover over critical areas, gather environmental intelligence, and serve as relays for ground agents and edge computing nodes. As shown in **Figure 7**, UAVs are organized in coordinated clusters to ensure wide-area coverage and avoid redundancy via non-overlapping trajectories. This hierarchical topology, supported by relay UAVs and edge servers acting as temporary aggregators, facilitates multi-hop knowledge fusion and real-time model dissemination across swarms. **Figure 8** further highlights DECKS' role in decentralizing computation and extending edge intelligence to disconnected vehicular networks. The following section presents the design of the proposed DECKS architecture, which leverages hovering UAVs for scalable knowledge aggregation and delivery in 6G-enabled vehicular environments. We first introduce the hybrid UAV system model, followed by the clustered coverage optimization algorithm and decentralized knowledge-sharing methodology.

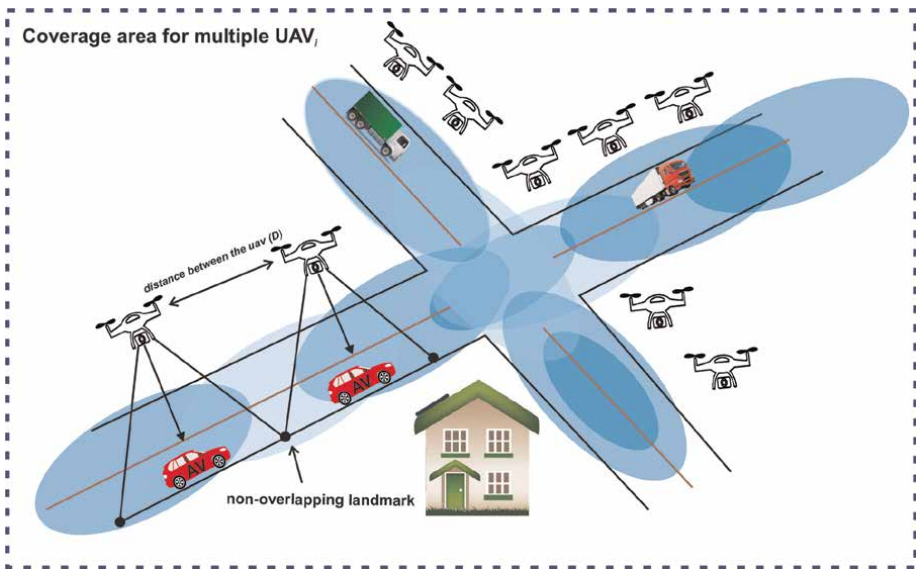


Figure 7.
UAVs cluster coverage area on terrain ground.

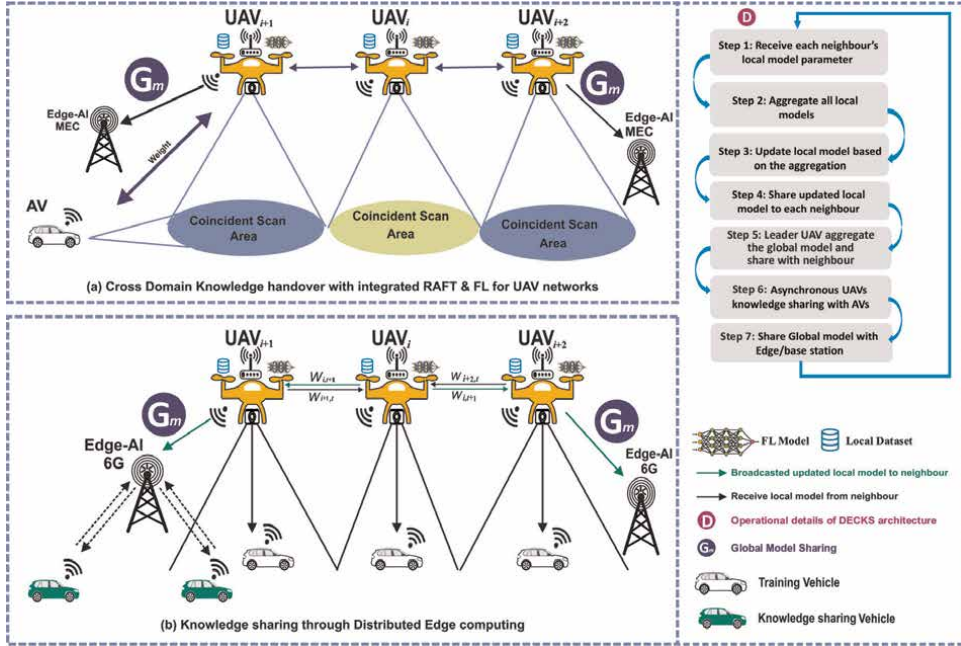


Figure 8.
An illustration of the proposed DECKS architecture for UAVs network knowledge sharing.

3.4.1 UAVs hybrid system structure

With the advent of 6G-V2X technologies, UAVs have become essential for augmenting ground-based sensing and computing in scenarios where Roadside Units (RSUs) are sparse or unavailable. In the proposed architecture, UAVs serve threefold roles: (i) relaying situational awareness to ground vehicles, (ii) participating in collaborative Federated Learning (FL) for knowledge fusion, and (iii) communicating with the nearest base station to synchronize and share the global model. These hovering UAVs enhance line-of-sight coverage and act as mobile signal relays, thereby extending the edge intelligence to CVNs. While fulfilling these threefold roles—situational relaying, collaborative federated learning, and base station communication—their operational efficiency is constrained by energy limitations. UAVs consume approximately 1.5% more energy with every unit increase in altitude. To mitigate this, our architecture strategically employs stationary hovering during both data capture and learning phases, optimizing energy use while maintaining wide-area coverage and stable model convergence. When operating in clusters, UAVs form a multi-hop relay network to propagate model updates and aggregate knowledge, as depicted in **Figure 7**. This architecture also supports long-range federated updates to CVNs located in sparse urban areas.

3.4.2 UAVs cluster coverage algorithm design

We define a constrained optimization problem that aims to maximize offloading utility from UAVs to vehicles during federated learning while adhering to spatial, energy, and time constraints. The objective is solved using a Lagrangian formulation.

3.4.2.1 Objective function

$$\begin{aligned} \max \sum_{i=1}^n A_i - \sum_{i=1}^n \lambda_i (h_i - z_i) - \sum_{i \neq j}^n \mu_{ij} \left[(x_i - x_j)^2 + (y_i - y_j)^2 - (r_i + r_j)^2 \right] \\ - \sum_{i=1}^n \beta_i (z_i - Z_{\max}) - \sum_{i,j=1}^n \gamma_{ij} (d_{ij} - v_j T) + \xi \sum_{i \neq j}^n \eta_{ij} \\ - \sum_{i=1}^n \delta_i (P_i - P_{\max,i}) - \sum_{i=1}^n \epsilon_i [P_i - \eta_i (C_i + D_i)] \end{aligned} \quad (6)$$

3.4.2.2 Constraints

- *Minimum height:* $h_i \geq z_i$
- *Non-overlapping search area:* $(x_i - x_j)^2 + (y_i - y_j)^2 \geq (r_i + r_j)^2$
- *Altitude limit:* $z_i \leq Z_{\max}$
- *Knowledge delivery time:* $\frac{d_{ij}}{v_j} \leq T$
- *Power limit:* $P_i \leq P_{\max,i}$
- *Power model:* $P_i = \eta_i (C_i + D_i)$
- *Data consistency:* Ground vehicles must receive data from UAVs sequentially and within their sensing cones.

Lagrange multipliers $\lambda_i, \mu_{ij}, \beta_i, \gamma_{ij}, \eta_{ij}, \delta_i, \epsilon_i$ encode penalties for constraint violations. The hyperparameter ξ balances overlapping minimization and knowledge utility.

3.4.3 UAVs knowledge sharing methodology

Our distributed federated framework does not rely on a central parameter server. Instead, each UAV updates its model locally and shares weights with 1-hop neighbors in its cluster (see **Figure 8**). $W_{i,t}$ denotes UAV i 's local model at round t . UAV i aggregates weights from neighbors $i + 1$ and $i + 2$ as:

$$W_{i,t+1} = \text{Agg}(W_{i,t}, W_{i+1,t}, W_{i+2,t}) \quad (7)$$

The learning goal is:

$$\min_{z \in \mathbb{R}^d} f(z) = \frac{1}{n} \sum_{i=1}^n f_i(z) \quad (8)$$

where each UAV minimizes its own local loss:

$$\min_{\Theta^i} \mathcal{L}_i = \sum_{k=1}^{n^k} \ell(y_i^k, f_i(x_i^k)) \quad (9)$$

Model update via stochastic gradient descent:

$$f_i(z) = \sum_{k=1}^K \frac{n_k}{n} F_k(z) \quad (10)$$

ensures smooth optimization even in non-IID federated settings. Each UAV also functions as a micro-parameter server for its immediate neighbors, facilitating localized knowledge fusion through iterative, decentralized updates. These updates occur over multiple communication cycles, after which the aggregated global model is relayed to an edge-based platform-as-a-service (PaaS), enabling the distribution of pre-trained models to other CVNs operating within the environment. As illustrated in **Figure 8**, the DECKS architecture orchestrates multi-hop neighbor exchanges, model aggregation, and seamless delivery across the edge. To minimize latency and improve channel efficiency, UAV-to-UAV communication is conducted via Device-to-Device (D2D) links, managed through Time Division or Frequency Division Multiple Access (TDMA/FDMA). This decentralized structure enables DECKS to maintain federated learning continuity in dynamically evolving swarm topologies, reducing single points of failure and enhancing the availability of intelligent inference at the edge.

Evaluation and results: The DECKS framework was evaluated using a simulated urban environment where a total of 10 UAVs cooperated under dynamic mobility and network conditions. The experiments were designed to measure convergence speed, energy efficiency, and communication cost compared to baseline algorithms like FedAvg and FedProx. **Figure 9** illustrates the training performance of the proposed decentralized UAV-based learning strategy across both IID and non-IID data settings. In the case of IID data with $n = 5$ UAVs (**Figure 9(a)**), the DECKS framework reaches a final model accuracy exceeding 95% after approximately 65 aggregation rounds, while the centralized server-based baseline plateaus around 80% after 75 rounds—indicating a 15% gain in final accuracy. Under non-IID conditions (**Figure 9(b)**), the

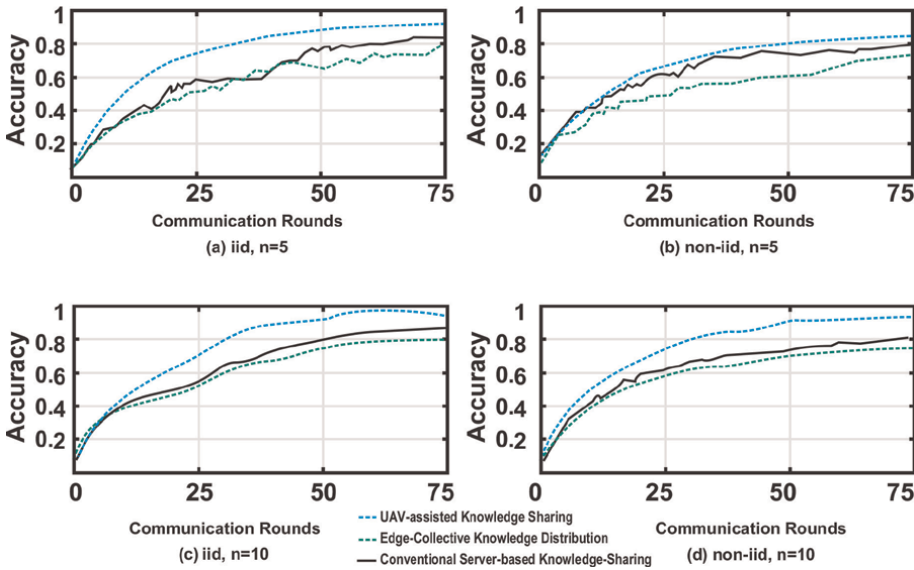


Figure 9. Learning curve for hovering UAV knowledge sharing with AVs (a, b) MNIST; (c, d) CIFAR-10.

DECKS model achieves its target accuracy within 60 rounds, whereas the baseline requires 75 rounds, reflecting a 15% improvement in convergence speed. As the number of UAVs increases to $n = 10$ (**Figure 9(c)**), DECKS demonstrates even faster convergence, approaching 99% accuracy under IID conditions and nearly 95% under non-IID with severely unbalanced datasets. These results underscore the scalability and robustness of the UAV swarm in adapting to heterogeneous data distributions. Notably, **Figure 9(d)** shows the impact of edge-enabled collective knowledge distribution on sensorless connected vehicles. Through one-shot inference, AVs not directly involved in the training process are able to benefit from shared UAV intelligence, significantly improving their onboard learning performance. This highlights the power of collaborative multi-UAV training in enhancing federated model dissemination, reducing energy consumption, and increasing overall system intelligence.

Figure 10 presents a comparative analysis of learning accuracy and energy consumption under different UAV knowledge-sharing strategies. The results demonstrate that distributing UAV-trained models directly to AVs via mobile edge computing (MEC) leads to faster convergence and reduced transmission latency. As shown in **Figure 10(a)**, classification accuracy improves as more UAVs participate in training—initially low due to limited local data, but significantly enhanced through collective weight transformation and aggregation. In **Figure 10(b)**, energy usage trends are illustrated with UAV node energy levels initialized between 80 and 100, and the head node set to 100. The system maintains an attenuation factor of 1.0 and a communication overhead cycle between 0.2 and 0.25. As communication nodes increase, higher transmission power is required to reach dynamic AVs, resulting in elevated energy demands. Nevertheless, DECKS maintains favorable trade-offs by leveraging energy-aware scheduling and UAV-assisted model dissemination, achieving system-wide efficiency without compromising learning performance. In summary, the evaluation of DECKS underscores its effectiveness in enabling distributed learning among UAV swarms. By integrating edge intelligence, energy-efficient aggregation strategies, and probabilistic trajectory coordination, DECKS delivers accelerated convergence, reduced communication overhead, and enhanced model robustness across airborne federated networks. These findings position DECKS as a critical intermediary between temporally adaptive frameworks such as FeMTC and infrastructure-dependent paradigms like V2I

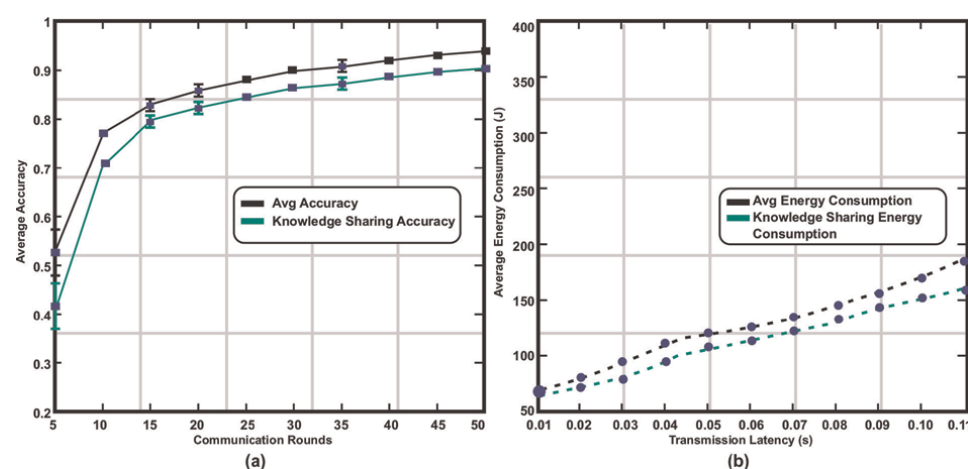


Figure 10.
Simulation results on (a) average accuracy and (b) energy consumption with CIFAR-10 datasets.

calibration, bridging the gap between dynamic swarms and structured vehicular systems within the broader landscape of decentralized autonomous intelligence.

3.5 V2I periodic calibration for autonomous vehicle control

Building upon the DECKS framework for airborne collaboration, this section transitions to infrastructure-assisted real-time control within ITS. While prior work has primarily concentrated on federated model training and adaptive learning, ensuring accurate command execution remains essential, particularly in mitigating post-learning drift. Conventional calibration for individual autonomous vehicles typically relies on the fusion of Global Positioning System (GPS) and Inertial Measurement Unit (IMU) data, along with high-definition (HD) maps. However, these methods face limitations in precision and real-time updates. Addressing this gap, the Internet of Vehicles (IoV)-based V2I calibration framework focuses on correcting synchronization delays and execution inconsistencies by leveraging periodic alignment signals from the RSU. These RSUs act as temporal anchors to enhance coordination among cooperative autonomous vehicles, ensuring reliable and synchronized control in dynamic traffic environments.

3.5.1 V2I calibration via RSUs

Autonomous agents often experience misalignment between planned and executed trajectories due to sensor latency and variable command processing. To mitigate this, RSUs broadcast periodic corrections to vehicles' control states. These signals realign actuation patterns and improve cooperative driving behavior. Each time a vehicle passes a roadside infrastructure, it can be calibrated to eliminate control drift and state errors caused by unavoidable random errors.

3.5.2 Simulation vs. experimental gap and setup with OptiTrack

In real-world scenarios, command drift and state errors still affect command execution. Therefore, the following compares simulation and experimentation to visually demonstrate the gap between them. The simulation has implemented by SUMO simulator as shown in **Figure 11**. A physical testbed comprising four UGVs under the OptiTrack motion capture, as shown in **Figure 12**.

However, as shown in the comparison between **Figures 13** and **14**, there are noticeable differences between simulation and real-world experiments. In the practical experiments, even when commands are simultaneously issued to identical Jackal robots, these commands are not perfectly synchronized despite similar average communication delays, resulting in significant fluctuations.

Hence, it is difficult to maintain the ideal situation which is no random error conditions in real-world scenarios. Therefore, periodic calibration becomes particularly crucial, whether for timestamps, environment, or vehicle position states. By using roadside infrastructures to perform calibration at certain distance intervals, it is possible to eliminate or suppress unanticipated random errors and further reduce the likelihood of cumulative errors, thus improving control accuracy. The following simulation further demonstrates the difference between using roadside infrastructures and single-vehicle calibration (SC), and validates the effectiveness of periodic calibration based on roadside infrastructures (RIC). The simulation detail see **Table 3**.

Figure 15 compares lateral trajectory deviation. SC mode shows noticeable drift from the ideal path, while RIC aligns almost perfectly with the ground truth. The simulated

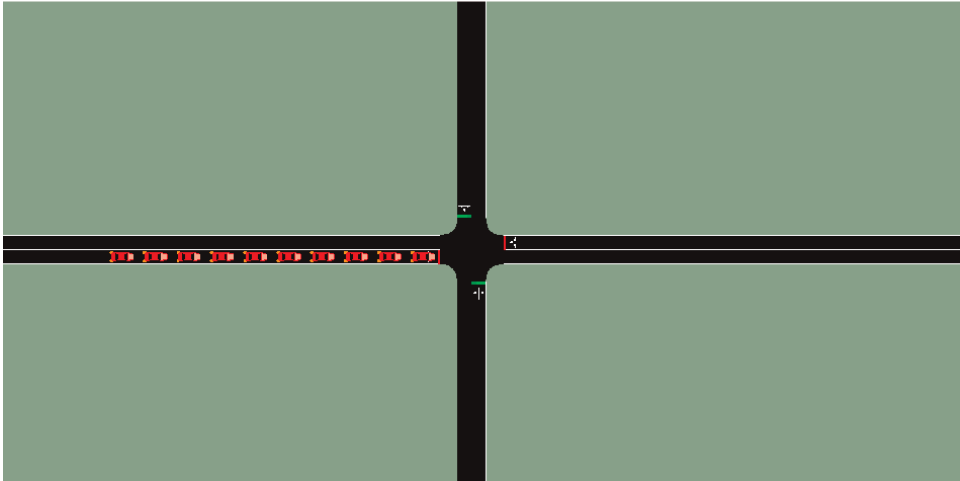


Figure 11.
SUMO simulation for V2I control.



Figure 12.
Experiment setup for V2I based following model.

scenarios include two cases: one involves autonomous intelligent connected vehicles relying solely on onboard sensors for self-calibration, while the other utilizes infrastructure connectivity for correcting motion commands. In the designed simulations, vehicles relying exclusively on self-correction achieved a command accuracy of 99%, whereas vehicles assisted by infrastructure-based position corrections reached nearly 100% accuracy.

The inset in **Figure 15** shows the probability distribution function (PDF) of position error under each mode, with RIC approximating a Gaussian distribution centered around zero deviation, validating its stability. The section highlights the critical role of infrastructure in closing the loop between inference and execution. The V2I framework provides the final synchronization layer in the knowledge aggregation continuum, complementing UAV-based (DECKS), vehicular (FeMTC), and IoRT (K-DUMBS) approaches with precise, real-time corrective control.

Description	Value	Description	Value
Initial position (x, y)	(0, 0)	Initial angular acceleration	0 rad/s ²
Initial speed	10 m/s	Time step	1 s
Initial heading	0 rad	Packet loss rate	5%
Initial acceleration	0 m/s ²	Angular acceleration error	N0,0.01
Initial angular velocity	0 rad/s	Communication delay	0–0.2 s
Total simulation time	200 s	Infrastructure gap	100 m

Table 3. Experimental parameters for enhancing vehicle control precision and preventing error accumulation through spaced fixed infrastructure sensors.

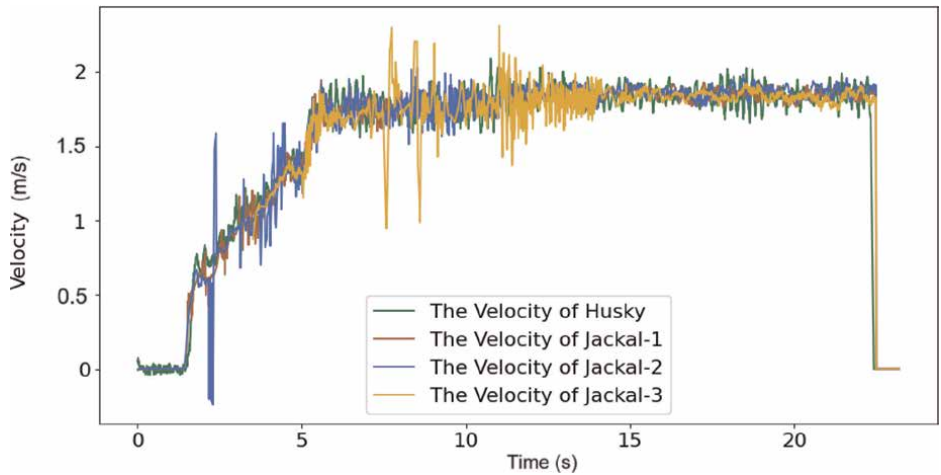


Figure 13. V2I experiment velocity over time.

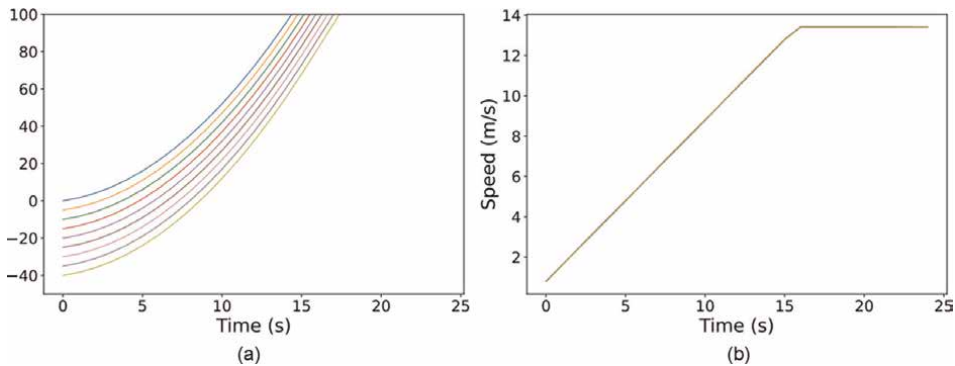


Figure 14. (a) V2I centralized control position; (b) V2I centralized control velocity.

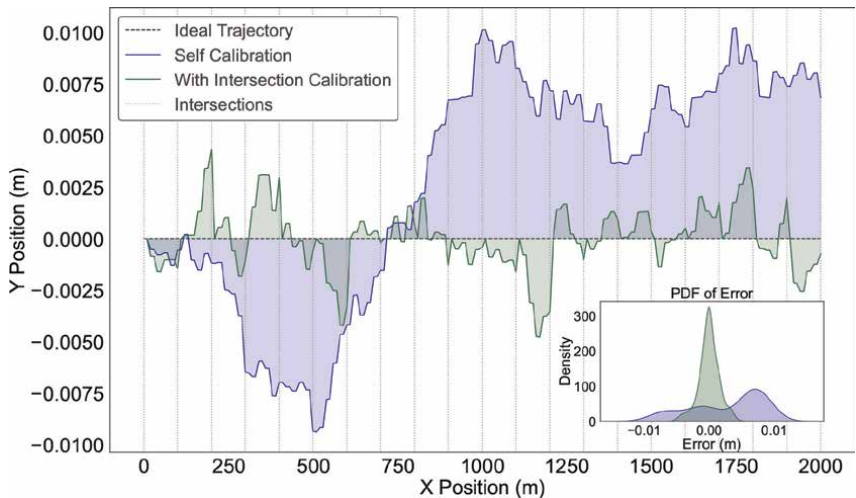


Figure 15.
Comparison of lateral trajectory deviation and corresponding PDF under SC and RIC.

4. Comparative summary of pre-trained knowledge aggregation frameworks

This chapter introduces a unified knowledge aggregation framework that advances collaborative autonomy through the integration of PTMs and FL across multiple domains, including the IoRT, CVNs, UAV swarms, and ITS. The comparative evaluation highlights both the architectural diversity and the shared design principles that enable scalable, decentralized learning.

4.1 Knowledge aggregation techniques

- *K-DUMBs (IoRT)* enables one-shot inference using shared PTM blocks, allowing low-power agents to perform distributed inference without engaging in full model training.
- *FeMTC (CVNs)* utilizes temporal clustering based on cosine similarity to improve personalized learning in environments characterized by sparse, non-IID vehicular data.
- *DECKS (UAV swarms)* employs hovering UAVs to aggregate and relay federated models in real time, enhancing ground-level inference accuracy while reducing energy expenditure.
- *V2I calibration (ITS)* uses infrastructure-based synchronization to maintain high-precision control and mitigate command drift in real-time autonomous operations.

4.2 Performance metrics: Energy, latency, and accuracy

Energy efficiency: K-DUMBs achieves nearly 80% reduction in communication overhead by eliminating continuous gradient exchange and enabling one-shot model

inference. FeMTC further lowers computational demands through temporal clustering, which reduces redundant updates and promotes localized personalization. DECKS contributes up to 70% energy savings by leveraging D2D communication, avoiding redundant model uploads, and restricting transmissions to key aggregation rounds. V2I Calibration minimizes on-board processing requirements through periodic infrastructure-based synchronization, reducing control recalibration costs for connected vehicles.

Latency and synchronization: V2I Calibration enhances vehicle start-up responsiveness by 60% compared to V2V-only systems. DECKS and FeMTC reduce convergence time by a factor of 2–3 compared to centralized or unclustered FL baselines. K-DUMBs eliminates multi-round communication altogether for inference-only clients, resulting in ultra-low latency.

Model accuracy: All frameworks significantly outperform their respective baselines. FeMTC boosts accuracy by 15–20% on non-IID datasets, while DECKS achieves 95–99% accuracy in aerial training scenarios. K-DUMBs maintains over 90% accuracy with only a single PTM broadcast. V2I Calibration improves control precision by reducing trajectory deviation from 12.4 to 4.7 cm.

4.3 Deployment scope and scalability

K-DUMBs is well-suited for resource-constrained IoRT applications such as domestic robotics or industrial automation. *FeMTC* is optimal for vehicular networks in smart cities where client connectivity is sporadic and data distribution is non-IID. *DECKS* enables federated intelligence across large-scale UAV deployments, facilitating coordinated model delivery and surveillance. *V2I calibration* is tailored for safety-critical ITS environments where infrastructure support is available for synchronization. Together, these modular architectures form an interoperable foundation for edge-enabled autonomy across air, road, and robotic ecosystems.

4.4 Limitations and scalability considerations

While the proposed framework demonstrates measurable improvements in energy efficiency, accuracy, and convergence speed, several limitations must be acknowledged to contextualize real-world deployment. First, reliance on PTMs introduces overhead in terms of memory and computation, especially for resource-constrained agents such as IoRT devices. Although one-shot inference mitigates some of these costs, fine-tuning or adapting PTMs locally remains challenging in devices with limited hardware. Second, while FL reduces the exchange of raw data, it still requires periodic communication of model weights. In UAV swarm scenarios, this can result in increased communication complexity and synchronization challenges, particularly in non-line-of-sight or harsh environments. Moreover, the assumption of reliable D2D communication and accurate trajectory control in DECKS may not hold in congested airspace or during adverse weather conditions, which limits scalability. Similarly, V2I calibration assumes consistent infrastructure support, which may not be available in rural or underdeveloped areas. Lastly, while our experimental setups reflect practical deployments, larger-scale tests across thousands of heterogeneous nodes remain to be conducted. Further work should investigate adaptive model compression, real-time reinforcement learning at the edge, and fault-tolerant aggregation mechanisms to ensure robustness in unstructured, rapidly changing environments.

5. Conclusion

The growing complexity of autonomous systems and the increasing need for intelligent, real-time decision-making have exposed the limitations of centralized learning architectures. High communication latency, energy constraints, and privacy concerns diminish the effectiveness of cloud-centric approaches, especially in environments marked by mobility, heterogeneity, and stringent responsiveness requirements. This chapter introduces a unified, scalable framework for knowledge aggregation that integrates pre-trained models, federated learning, and edge intelligence to address these challenges across diverse domains, including the Internet of Robotic Things (IoRT), Connected Vehicular Networks (CVNs), Unmanned Aerial Vehicle (UAV) swarms, and Intelligent Transportation Systems (ITS). By embedding domain-specific strategies—such as one-shot inference for lightweight IoRT agents, temporal clustering for non-IID vehicular data, edge-assisted swarm learning in UAVs, and real-time V2I calibration for precise control—the framework demonstrates that decentralized knowledge aggregation can achieve efficiency, robustness, and scalability. These contributions are supported by methodologically grounded approaches, including gradient-based optimization, decentralized multi-hop sharing, and resource-aware orchestration suited to dynamic, constrained environments. Crucially, the framework treats PTMs not as static initializations but as dynamic, distributable anchors of intelligence—supporting localized adaptation while ensuring global learning consistency. Empirical evaluations show that the proposed methods outperform centralized baselines in convergence speed, accuracy, energy use, and deployment flexibility. Looking forward, this work lays the foundation for the next generation of adaptive, interoperable autonomous systems. Future directions include exploring cross-domain model transfer, real-time reinforcement learning, and resilient edge communication to enable collaborative intelligence in unpredictable and distributed settings.

Acknowledgements

The authors thank the University of Glasgow for supporting this work and providing access to software resources and robotics laboratories. Gratitude is also extended to the contributing authors for their valuable input and collaboration.

The author acknowledges the use of an AI-based tool for language refinement and formatting of the original manuscript. The scientific content, analysis, and conclusions presented in this chapter are entirely the author's own work.

Author details

Muhammad Waqas Nawaz^{*†}, Wanquan Zhang[†], Jaspreet Kaur[†], Qammer H. Abbasi[†],
Muhammad Imran[†] and Olaoluwa Popoola[†]
University of Glasgow, James Watt School of Engineering, Glasgow, United Kingdom

*Address all correspondence to: waqasnawaz_2008@yahoo.co.uk

† These authors contributed equally.

IntechOpen

© 2026 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

References

- [1] Murroni M, Anedda M, Fadda M, et al. 6G—Enabling the new smart city: A survey. *Sensors*. 2023;**23**(17):7528. Available from: <https://www.mdpi.com/1424-8220/23/17/7528> [Accessed: May 14, 2025]
- [2] Zhou Y, Rao B, Wang W. UAV Swarmintelligence: Recent Advances and Future Trends. Available from: <https://ieeexplore.ieee.org/abstract/document/9214446> [Accessed: May 14, 2025]
- [3] Lv Z et al. AI empowered communication systems for intelligent transportation systems. *IEEE Transactions on Intelligent Transportation Systems*. 2021;**22**(7): 4579-4587. Available from: <https://ieeexplore.ieee.org/abstract/document/9181452> [Accessed: May 14, 2025]
- [4] Chen M et al. Artificial neural networks-based machine learning for wireless networks: A tutorial. *IEEE Communications Surveys & Tutorials*. 2019;**21**(4):3039-3071. Available from: <https://ieeexplore.ieee.org/document/8755300> [Accessed: May 14, 2025]
- [5] Lu Y et al. Toward direct edge-to-edge transfer learning for IoT-enabled edge cameras. *IEEE Internet of Things Journal*. 2021;**8**(6):4931-4943. Available from: <https://ieeexplore.ieee.org/abstract/document/9240982> [Accessed: May 14, 2025]
- [6] Mora A et al. Knowledge Distillation for Federated Learning: A Practical Guide. [Online]. Available from: <https://www.ijcai.org/proceedings/2024/0905.pdf> [Accessed: May 14, 2025]
- [7] Samara G et al. Internet of Robotic Things: Current Technologies and Applications. 2021. pp. 1-6. Available from: <https://ieeexplore.ieee.org/document/9677407> [Accessed: May 14, 2025]
- [8] Chen M et al. Communication-efficient federated learning. *Proceedings of the National Academy of Sciences*. 2021;**118**(17):e2024789118. Available from: <https://www.pnas.org/doi/abs/10.1073/pnas.2024789118> [Accessed: May 14, 2025]
- [9] Li Q et al. A survey on federated learning systems: Vision, hype and reality for data privacy and protection. *IEEE Transactions on Knowledge and Data Engineering*. 2023;**35**(4):3347-3366. Available from: <https://ieeexplore.ieee.org/abstract/document/9599369> [Accessed: May 14, 2025]
- [10] Konečný J, McMahan HB, Ramage D, Richtárik P. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv Preprint*. arXiv:1610.02527. 2016. Available from: <https://arxiv.org/abs/1610.02527> [Accessed: May 14, 2025]
- [11] Wen H et al. Communication-efficient federated learning on non-IID data using two-step knowledge distillation. *IEEE Internet of Things Journal*. 2023;**10**(19):17307-17322. Available from: <https://ieeexplore.ieee.org/document/10129025> [Accessed: May 14, 2025]
- [12] McMahan HB, Moore E, Ramage D, Hampson S. Communication-efficient learning of deep networks from decentralized data. *arXiv preprint arXiv*: 1602.05629. 2016. Available from: <https://arxiv.org/abs/1602.05629> [Accessed: May 14, 2025]
- [13] Gong T et al. Edge intelligence in intelligent transportation systems: A survey. *IEEE Transactions on Intelligent*

- Transportation Systems. 2023;**24**(9): 8919-8944. Available from: <https://ieeexplore.ieee.org/abstract/document/10133894> [Accessed: May 14, 2025]
- [14] Fu Y et al. A survey of driving safety with sensing, vehicular communications, and artificial intelligence-based collision avoidance. *IEEE Transactions on Intelligent Transportation Systems*. 2022;**23**(7):6142-6163. Available from: <https://ieeexplore.ieee.org/document/9447828> [Accessed: May 14, 2025]
- [15] Guleng S, Wu C, Liu Z, Chen X. Edge-based V2X communications with big data intelligence. *IEEE Access*. 2020; **8**:8603-8613. Available from: <https://ieeexplore.ieee.org/document/8951087> [Accessed: May 14, 2025]
- [16] Lee J, Niyato D, Guan YL, Kim DI. Learning to schedule joint radar-communication with deep multi-agent reinforcement learning. *IEEE Transactions on Vehicular Technology*. 2022;**71**(1):406-422. Available from: <https://ieeexplore.ieee.org/document/9601214> [Accessed: May 14, 2025]
- [17] Nawaz MW, Imran MA, Popoola O. Temporal hierarchical clustering for knowledge aggregation in connected vehicular networks. In: *IEEE WCNC. Dubai: IEEE; 2024*. Available from: <https://ieeexplore.ieee.org/document/10570723> [Accessed: May 14, 2025]
- [18] Nawaz MW, Imran MA, Popoola O. 6G edge-networks and multi-UAV knowledge fusion for urban autonomous vehicles. *Physical Communication*. 2024; **67**:102479. DOI: 10.1016/j.phycom.2024.102479 [Accessed: May 14, 2025]
- [19] Nawaz MW et al. K-DUMBs: Knowledge-driven unified model block sharing in IoRT. In: *IEEE VTC. Italy, Florence: IEEE; 2023*. Available from: <https://ieeexplore.ieee.org/document/10200507> [Accessed: May 14, 2025]
- [20] Wu C et al. Communication-efficient federated learning via knowledge distillation. *Nature Communications*. 2022;**13**:2032. DOI: 10.1038/s41467-022-29763-x [Accessed: May 14, 2025]
- [21] Zhang H et al. A security optimization scheme for data security transmission in UAV-assisted edge networks based on federal learning. *Adhocnetworks*. 2023; **150**:103277. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S157087052300197X> [Accessed: May 14, 2025]
- [22] Nguyen DC et al. Federated learning for internet of things: A comprehensive survey. *IEEE Communications Surveys Tutorials*. 2021;**23**(3):1622-1658. Available from: <https://ieeexplore.ieee.org/abstract/document/9415623> [Accessed: May 14, 2025]
- [23] Gupta H et al. 6G-enabled edge intelligence for ultra-reliable low latency applications: Vision and mission. *Computer Standards & Interfaces*. 2021; **77**:103521. DOI: 10.1016/j.csi.2021.103521 [Accessed: May 14, 2025]
- [24] Zhang W et al. Advancing Autonomous Vehicle ITS with V2X and V2I Periodic Calibration. 2025. Available from: <https://eprints.gla.ac.uk/355350/> [Accessed: June 11, 2025]
- [25] Lee JN. The impact of knowledge sharing, organizational capability and partnership quality on IS outsourcing success. *Information and Management*. 2001;**38**(5):323-335. DOI: 10.1016/S0378-7206(00)00074-4 [Accessed: May 14, 2025]
- [26] Chellapandi VP et al. Federated learning for connected and automated

- vehicles: A survey of existing approaches and challenges. *IEEE Transactions on Intelligent Vehicles*. 2024;**9**(1):119-137. Available from: <https://ieeexplore.ieee.org/abstract/document/10316635> [Accessed: May 14, 2025]
- [27] Kairouz P, McMahan H, et al. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*. 2021;**14**(1-2):1-210. Available from: <https://arxiv.org/abs/1912.04977> [Accessed: May 19, 2025]
- [28] Hard A, Rao K, Mathews R, et al. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*. 2018. Available from: <https://arxiv.org/abs/1811.03604> [Accessed: May 19, 2025]
- [29] Chellapandi VP, Yuan L, Zak SH, Wang Z. A survey of federated learning for connected and automated vehicles. In: *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*; Bilbao, Spain. 2023. pp. 2485-2492. DOI: 10.1109/ITSC57777.2023.10421974
- [30] Yang Q, Liu Y, Chen T, Tong Y. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology*. 2019;**10**(2):1-19. Available from: <https://dl.acm.org/doi/10.1145/3298981> [Accessed: May 19, 2025]
- [31] Samarakoon S, Bennis M, Saad W, Debbah M. Federated learning for ultra-reliable low-latency V2V communications. In: *2018 IEEE Global Communications Conference (GLOBECOM)*; Abu Dhabi, United Arab Emirates. 2018. pp. 1-7. DOI: 10.1109/GLOCOM.2018.8647927 [Accessed: May 19, 2025]
- [32] Nguyen DC et al. Federated learning for smart healthcare: A survey. *ACM Computing Surveys*. 2023;**55**(3):37. Article 60. DOI: 10.1145/3501296 [Accessed: May 19, 2025]
- [33] Tan AZ et al. Towards personalized federated learning. *IEEE Transactions on Neural Networks and Learning Systems*. 2023;**34**(12):9587-9603. DOI: 10.1109/TNNLS.2022.3160699 [Accessed: May 19, 2025]
- [34] Lyu L, Yu H, Yang Q. Threats to Federated Learning: A Survey, 2020. Available from: <https://arxiv.org/abs/2003.02133> [Accessed: May 19, 2025]
- [35] Zhao Y, Li M, Lai L, et al. Federated learning with non-IID data. *arXiv preprint arXiv:1806.00582*. 2018. Available from: <https://arxiv.org/abs/1806.00582> [Accessed: May 19, 2025]
- [36] Wang S, Tuor T, Salonidis T, et al. Adaptive federated learning in resource constrained edge computing systems. *IEEE Journal on Selected Areas in Communications*. 2019;**37**(6):1205-1221. Available from: <https://ieeexplore.ieee.org/document/8664630> [Accessed: May 19, 2025]
- [37] Benhelal MS et al. Autonomous UAV edge architecture for road hazards extended reality warnings. *IEEE Internet of Things Magazine*. 2023;**6**(1):26-30. DOI: 10.1109/IOTM.001.2200253 [Accessed: May 23, 2025]
- [38] Li H, Ota K, Dong M. Learning IoV in 6G: Intelligent edge computing for internet of vehicles in 6gwireless communications. *IEEE Wireless Communications*. . DOI: 10.1109/MWC.017.2200089 [Accessed: May 23, 2025]
- [39] Mashhadi MB, Mahdavi Moghadam M, Tafazolli R, Saad W. Collaborative learning with a drone orchestrator. *IEEE Transactions on Vehicular Technology*.

DOI: 10.1109/TVT.2023.3303630

[Accessed: May 23, 2025]

[40] Saraswat D et al. Blockchain-based federated learning in UAVs beyond 5G networks: A solution taxonomy and future directions. *IEEE Access*. 2022;**10**: 33154-33182. DOI: 10.1109/ACCESS.2022.3161132 [Accessed: May 23, 2025]

[41] Khan LU et al. Federated learning for digital twin-based vehicular networks: Architecture and challenges. *IEEE Wireless Communications*. DOI: 10.1109/MWC.012.2200373 [Accessed: May 23, 2025]

[42] Fu Y, Li C, Yu FR, Luan TH, Zhang Y. A selective federated reinforcement learning strategy for autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*. 2022;**24**(2): 1655-1668. DOI: 10.1109/TITS.2022.3219644 [Accessed: May 23, 2025]

[43] Zhou X, Liang W, She J, Yan Z, Wang KI-K. Two-layer federated learning with heterogeneous model aggregation for 6G supported internet of vehicles. *IEEE Transactions on Vehicular Technology*. 2021;**70**(6):5308-5317. DOI: 10.1109/TVT.2021.3077893 [Accessed: May 23, 2025]

[44] Qu Y et al. Decentralized federated learning for UAV networks: Architecture, challenges, and opportunities. *IEEE Network*. 2021;**35**(6):156-162. DOI: 10.1109/MNET.001.2100253 [Accessed: May 23, 2025]

[45] Mehta P, Gupta R, Tanwar S. Blockchain envisioned UAV networks: Challenges, solutions and comparisons. *Computer Communications*. 2020;**151**: 518-538. Available from: <https://www.sciencedirect.com/science/article/pii/S0140366419318675> [Accessed: May 23, 2025]

[46] Mosquet M et al. A roadmap to safer driving through advanced driver assistance systems. *Auto Tech Review*. 2016;**5**:20-25. DOI: 10.1365/s40112-0161164-1

[47] Liu Y, Guo B, Li N, Ding Y, Zhang Z, Yu Z. CrowdTransfer: Enabling crowd knowledge transfer in AIoT community. *IEEE Communications Surveys & Tutorials*. 2025;**27**(2):1191-1237. DOI: 10.1109/COMST.2024.3423319 [Accessed: May 25, 2025]

[48] Tian Z, Wang H, Li Z, Niu Z, Wei X, Su Y. MDTL: Maliciously secure distributed transfer learning based on replicated secret sharing. *IEEE Transactions on Network and Service Management*. 2025;**22**(1):877-891. DOI: 10.1109/TNSM.2025.3529471 [Accessed: May 25, 2025]

[49] Verma H, Chauhan N, Awasthi LK. A comprehensive review of 'Internet of Healthcare Things': Networking aspects, technologies, services, applications, challenges, and security concerns. *Computer Science Review*. 2023;**50**: 100591. DOI: 10.1016/j.cosrev.2023.100591 [Accessed: May 25, 2025]

Edited by Pier Luigi Mazzeo and Alessandro Bruno

A paradigm change is currently occurring in the field of artificial intelligence. Initially, a straightforward but ever-expensive formula, involving more data and more computation, described the course of machine learning. We started from scratch and trained ever-larger models from random initialization on enormous, task-specific datasets. Although effective, this strategy consumed a significant amount of resources, made it challenging to enter new markets, and was fragile in conditions with limited data. This book is designed for engineers, researchers, students, and technology leaders who want not only to understand this shift but also to actively participate in it. Our goal is to provide you with both the conceptual framework and the practical expertise to harness the power of pre-trained models, saving you time and computational resources, and ultimately enabling you to solve problems that were previously out of reach.

Andries Engelbrecht, Artificial Intelligence Series Editor

Published in London, UK

© 2026 IntechOpen

© your_photo / iStock

IntechOpen

ISSN 2633-1403

ISBN 978-1-83635-353-9

