University of Catania

Department of Mathematics and Computer Science

PhD in Computer Science XXXVIII Cycle

Mattia Litrico

Enhancing Deep Learning Methodologies under Limited Labelled Data

---

Ph.D. Thesis

Supervisor: Prof. Sebastiano Battiato

Co-supervisor: Dr. Mario Valerio Giuffrida

Anno Accademico 2024 - 2025

Abstract

Deep learning has achieved remarkable success across a wide range of tasks, yet its dependence on large-scale, fully annotated datasets remains a critical bottleneck, especially in scenarios where annotations are expensive, ambiguous, or hard to obtain. This thesis investigates strategies to mitigate annotation scarcity by developing methods that learn effectively from limited supervision, with a particular focus on unsupervised domain adaptation, open-set recognition, and weak supervision. Firstly, this thesis addresses the challenge of source-free open-set unsupervised domain adaptation, where no access to source data is available and the target domain contains previously unseen classes. This research proposes a methodology that takes advantage of the granularity of unknown categories by segregating their samples into multiple unknown classes, enabling robust adaptation with new semantics. Moreover, this thesis explores the role of textual information for guiding the adaptation of visual models under complex domain shifts. We show that integrating textual data provides robustness to complex shifts, improving the unsupervised adaptation on target domains. Finally, this thesis explores the possibility to perform density estimation requiring in the absence of location-level annotations. We introduce a novel approach that leverages global image-level information to predict spatially meaningful density maps, achieving competitive results with significantly reduced annotation cost. Overall, this thesis contributes to improving deep learning training in scenarios where annotated data is scarce or only partially available, enabling broader real-world applications in resource-limited domains, such as healthcare, agriculture or autonomous systems.

Contents

Chapter 1

Introduction

1.1 Motivation

Deep learning has penetrated in several fields by achieving state-of-the-art performance across a wide range of applications [1, 2, 3, 4, 5, 6, 7, 8]. However, these successes are largely correlated with the availability of large-scale, fine-grained and fully annotated datasets. In many real-world applications, acquiring such detailed annotations is a significant bottleneck due to the immense cost, time, and expertise required. This dependency on extensive labelled data limits the scalability and accessibility of deep learning solutions, especially in domains where data annotation is expensive, ambiguous, or even infeasible.

For instance, in medical imaging, annotations often require highly trained specialists, such as radiologists or pathologists, to label complex anatomical structures or pathological findings [8]. The annotation of each image can take substantial time, and when scaled to thousands or millions of images, the cumulative effort extends over months or years, leading to considerable financial costs [9]. Moreover, ensuring annotation quality often involves multiple rounds of review to correct errors and reduce subjectivity, imposing further operational expenses [10]. Finally. the variability in expert annotations can also reduce the effectiveness and reliability of the training of such deep learning models [11]. Similarly, autonomous driving and robotics domains generate enormous volumes of data, such as sensors acquisition, images, videos or LIDAR scans. This data requires detailed annotations to be effectively used by deep learning algorithm for solving a large number of tasks. Annotating such datasets manually is extremely labor-intensive and costly and lags behind with data collection rates [10]. Other fields such as remote sensing, natural

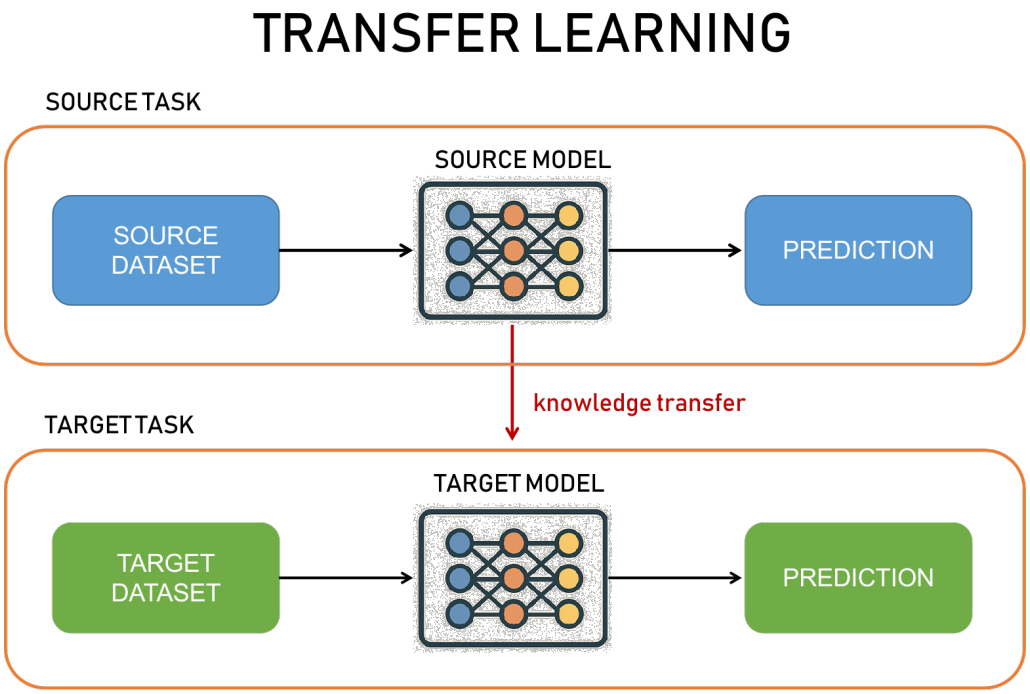

Figure 1.1: Transfer learning aims at transferring knowledge acquired by a model on a labeled source task, to make it working effectively on a target task without requiring labeled data for the target task.

language processing, and industrial inspection face analogous challenges. Satellite imagery annotation demands significant domain knowledge to label land use patterns or environmental changes accurately. Audio and video data require precise temporal labeling, demanding both expertise and intensive labor [11]. As a result, reducing the annotation burden is both a practical necessity and a major focus of research in deep learning. Indeed, reducing the annotation cost allows a rapid spread of deep learning-based technologies, obtaining tangible benefits in multiple critical domains.

One key strategy that has emerged is transfer learning, which leverages knowledge acquired from labeled data in a related source task to improve learning in a new target task with fewer annotations. A schematic representation of transfer learning is offered in Figure 1.1.

Within transfer learning, domain adaptation (DA) specifically addresses the case where the source and target tasks remain the same, but the underlying data distributions differ due to what is commonly referred to as domain shift. Domain shift

may arise from variations in visual appearance, differences in data acquisition protocols, or gaps between synthetic and real data. In this setting, a deep learning model is trained on a labeled source domain and adapted to generalize to an unseen, unlabeled target domain [12]. By aligning the source and target distributions, domain adaptation methods reduce the reliance on large-scale annotation in every new environment, thereby lowering annotation costs, improving scalability, and accelerating the deployment of AI systems in real-world applications.

Other paradigms to reduce the burden of collecting annotated data are weak supervision and semi-supervision that leverage less detailed or noisier labels, such as image-level or partial annotations, or even completely unlabeled data combined with a small amount of high-quality labels [13, 14]. These methods aim to extract maximal learning signals from minimal annotations, thus reducing the time and expense associated with labeling while maintaining competitive performance.

More recently, the advent of large foundation models [15] has further revolutionized the applicability of AI frameworks in real world applications. These models are pre-trained on vast amounts of unlabeled data across multiple domains and modalities and they provide powerful, generalized representations that can be fine-tuned efficiently on target tasks with only modest annotation, drastically lowering the annotation barrier. A pictorial representation of differences between transfer learning, domain adaptation, weak-supervision and foundation models is shown in Figure 1.2.

This thesis is motivated by the pressing need to overcome the bottleneck of limited data annotation and it focuses on learning paradigms that operate with minimal supervision, including unsupervised domain adaptation, open-set recognition, and weakly supervised learning. These approaches aim to enhance model adaptability, robustness, and generalization, even when available annotations are partial or outdated.

1.2 Objectives

This thesis explores solutions for addressing fundamental challenges associated with the cost of collecting fine-grained annotated dataset and the necessity of training deep learning models with limited human-derived supervision. With this purpose

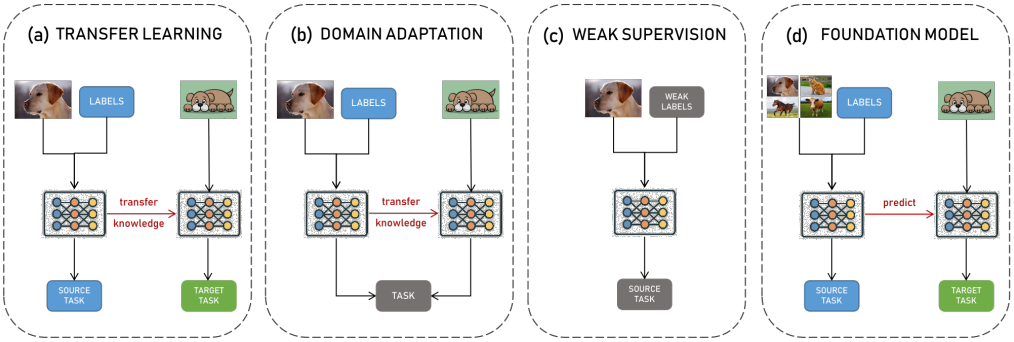

Figure 1.2: (a) Transfer Learning transfers knowledge from a labeled source task to a target task; (b) Domain Adaptation adapts a model trained on a labeled source domain to an unlabeled target domain under domain shift; (c) Weak-supervision leverages weak or noisy labels instead of full supervision; (d) Foundation Models are pretrained on large-scale labeled data and directly applied to downstream tasks.

this thesis aims at improving the understanding and effectiveness of deep learning solutions achieving the following objectives:

- Enhance the effectiveness of deep learning methods in scenarios with limited annotated data and the presence of novel and unseen categories: we propose a novel source-free open-set unsupervised domain adaptation approach that significantly improves performance under these conditions. Our goal is to improve knowledge transfer from pretrained models without requiring access to annotated source datasets, even n presence of novel and unseen categories. Achieving this objective will expand the applicability of deep learning solutions across industrial domains where annotated data is scarce and the ability to handle unseen objects is essential.

- Explore the integration of language to enhance deep learning models across diverse geographical regions: we propose a novel unsupervised domain adaptation method that incorporates the language modality to improve the robustness of deep learning models when handling data from different countries worldwide. Our objective is to facilitate knowledge transfer from pretrained models trained on data collected in one region to effectively adapt to data from entirely different countries. This approach aims to enable the deployment of deep learning solutions in underdeveloped regions where obtaining annotated datasets is particularly challenging.

- Develop a weakly supervised crowd density estimation pipeline that does not require location-level annotations: we propose a novel framework for crowd density estimation capable of predicting spatially meaningful density maps while relying solely on count-level annotations during training. This approach significantly reduces the annotation cost and effort typically associated with fine-grained labelling, as count-level data can be more easily collected across various domains, including surveillance and medical scenarios. By minimizing the need for location-level annotations, our goal is to enable a wider use of crowd density estimation technologies in contexts where the cost and effort of acquiring fine-grained annotated datasets is unsustainable.

- Investigate generative models for effortless dataset augmentation in medical imaging: we explore diffusion models to temporally model MRI trajectories associated with neurodegenerative diseases. Our objective is to demonstrate how generative models can serve as effective tools for synthesizing new, unseen data in contexts where data acquisition is challenging, such as medical imaging.

By achieving these objectives, this thesis seeks at promoting to encourage the broader use of deep learning frameworks in challenging scenarios characterized by limited labelled data, the presence of novel and unseen categories, cross-geographical data variability and weaker forms of supervision. Moreover, this thesis provides solutions that enhance the applicability of these technologies in critical domains such as surveillance and medical imaging.

1.3 Contributions

The primary contributions of this thesis can be summarized as follows:

- A novel source-free open-set unsupervised domain adaptation framework: We introduce a method that enables effective knowledge transfer from pretrained models without access to source domain data, while robustly handling previously unseen classes through segregation of unknown categories.

- A multimodal domain adaptation approach integrating language modality: This work demonstrates how incorporating textual information improves the

robustness and generalization of unsupervised domain adaptation across diverse geographical and cultural data distributions.

- A weakly supervised crowd density estimation pipeline: We propose a framework that relies solely on count-level annotations to generate spatially meaningful density maps, significantly reducing annotation costs by eliminating the need for fine-grained location-level labels.

- Exploration of diffusion-based generative models for medical imaging data augmentation: The thesis investigates how diffusion models can temporally model MRI trajectories in neurodegenerative diseases to synthetically augment datasets.

The findings and contributions of this thesis have been disseminated through the following publications:

- Mattia Litrico, Davide Talon, Sebastiano Battiato, Alessio Del Bue, Mario Valerio Giuffrida, Pietro Morerio, "Uncertainty-guided Open-Set Source-Free Unsupervised Domain Adaptation with Target-private Class Segregation" [16]. International Journal of Computer Vision, 2025.

- Mattia Litrico, Mario Valerio Giuffrida, Sebastiano Battiato, Devis Tuia, "TRUST: Leveraging Text Robustness for Unsupervised Domain Adaptation" [17]. AAAI Conference on Artificial Intelligence, 2026.

- Mattia Litrico, Feng Chen, Michael Pound, Sotirios A Tsaftaris, Sebastiano Battiato, Mario Valerio Giuffrida, "COUNT2DENSITY: Crowd Density Estimation without Location-level Annotations" [18]. Under Review, Pattern Recognition, 2025.

- Mattia Litrico, Francesco Guarnera, Mario Valerio Giuffrida, Daniele Ravì, Sebastiano Battiato, "TADM: Temporally-aware Diffusion Model for Neurodegenerative Progression on Brain MRI" [19]. International Conference on Medical Image Computing and Computer-Assisted Intervention, 2024.

- Mattia Litrico, Francesco Guarnera, Mario Valerio Giuffrida, Daniele Ravì, Sebastiano Battiato, "Temporally-Aware Diffusion Model for Brain Progression

Modelling with Bidirectional Temporal Regularisation" [20]. Under Review, Computerized Medical Imaging and Graphics, 2025

1.4 Thesis Outline

The reminder of this thesis is organized as following:

- Chapter 2 presents an in-depth analysis of the current state-of-the-art.
- Chapter 3 describes our study on using pseudo-labels for source-free open-set unsupervised domain adaptation.
- Chapter 4 presents our proposed solution to unsupervised domain adaptation under complex shifts leveraging the language modality.
- Chapter 5 explores our study on using only count-level annotations for density estimation.
- Chapter 6 concludes the thesis by summarizing key findings and outlining future research in the domain.
- Appendices A and B present two other studies where we propose two solutions to use generative models for efficient dataset augmentation in medical imaging.

Chapter 2

Related Works

2.1 Unsupervised Domain Adaptation

Unsupervised Domain Adaptation (shorten in UDA) aims to adapt a model trained on a labelled source domain to generalise well on an unlabelled target domain, in the presence of domain shift across the two domains [21]. Earlier UDA approaches focused on aligning statistical distributions [22]. More recent methods proposed domain alignment strategies, including Maximum Mean Discrepancy [23, 24], adversarial learning [25, 26], clustering [27, 28] and self-training [29, 30]. Other works leveraged generative models [31] and transformers [32, 33] to learn the underlying distributions of data using the generation process or the attention mechanism. However, all these methods required the access to both source and target data during the adaptation.

2.1.1 Source-free Unsupervised Domain Adaptation

Recently, some Source-free adaptation methods have been proposed, adapting to the target domain using only the source model and unlabelled target data [34, 35, 36, 37, 38, 39, 40, 29]. TENT [35] and SHOT [36] introduced entropy minimisation and pseudo-labeling. On-target [40] proposed to combine a pseudo-labels generation and a self-supervised task, but the pseudo-labels are not refined during the adaptation. Instead, [29] proposed a self-supervised strategy to refine the pseudo-labels.

2.1.2 Open-set Domain Adaptation

The assumption that both source and target domains share the same label space may not hold in several real-world scenarios, where the target domain can contain classes that are not included in the source domain. In the literature, such classes are referred to as *target private* (or simply *private*) or *unknown*. Open-set domain adaptation aims to transfer knowledge from a source domain to a target domain containing private classes. Recent works mainly focus on separating the known and unknown samples in the target domain [41, 42, 43, 44] or aligning the known distributions [43, 45]. Differently, [46] proposed to use word-prototypes based on mid-level features, which are more robust to the negative transfer to perform a subsidiary prototype-space alignment. UADAL [42] proposed to segregate the unknown classes leveraging a 3-way adversarial learning among the source, the shared, and unknown classes of the target. However, all these methods assume the availability of source data during the adaptation, *i.e.* they are not source-free.

More recently, [47, 39, 48] proposed source-free strategies for the open-set domain adaptation setting. However, these approaches rely on ad-hoc pre-training strategies on the source domain that assume the awareness of the presence of unknown classes in the target during the pretraining.

For example, OneRing [47] pretrained a classifier on the source data with an extra class to empower the model to better segregate target-private samples during adaptation. In [39, 48], the authors leverage a weighting scheme and feature-splicing to avoid negative transfer, but they generate extra training samples of unknown categories during pretraining on the source domain, facilitating the separation of private classes in the target domain. Differently, our method makes no ad-hoc assumptions on the training in the source domain, providing a more general approach to this challenging setting, widening its application.

The use of ad-hoc training strategies has been considered as a weaknesses also in the recent literature. For example, SHOT [36] and AaD [49] use a separation criterion based on the entropy of predictions, to distinguish between shared and unknown classes. Instead, USD [50] proposed to segregate private samples based on Jenson-Shannon divergence. However, since these separation criteria are computed on model outputs, they are easily affected by confirmation bias. More recently, GLC [51] introduces a global adaptive one-versus-all clustering to initially assign

pseudo-labels for known vs unknown classes in the target. To enhance robustness, a local k-NN refinement step smooths assignments among neighbors, mitigating noise and improving separation. Differently, LEAD [52] uses orthogonal feature decomposition orthogonal decomposition to construct source-known and unknown space. To identify target-private samples, it computes instance-level decision boundaries based on distances in these constructed spaces.

2.1.3 Self-supervised Learning and Pseudo-Labelling

Self-supervised methods are successful in learning transferable representations of visual data [53, 54, 55, 56, 57, 58, 59, 60, 61]. Specifically, [54, 56, 58] use contrastive-based pretext tasks to enhance the generalisation ability of deep models. Moreover, several recent self-supervised approaches have been utilised within UDA [62, 63] and SF-UDA [34, 40, 29, 64] settings. On the other hand, pseudo-labelling is a simple but effective technique used in semi-supervised [65, 66] and self-supervised learning [53], as well as in domain adaptation [36, 67, 40, 29]. It consists in using labels predicted by the model as self-supervision. Fix-Match [66] and On-target [40] take advantage of pseudo-labels without any refinement during training.

2.1.4 Learning from Noisy Labels

Noisy training sets often result in poorly trained models. In [68], the authors demonstrated that a deep neural network can easily overfit an entire dataset with any ratio of corrupted labels, resulting in a poor generalisation on test data. To address this problem, different approaches have been proposed focusing on noise-robust losses [69, 70, 71, 72], estimation of the noise-transition matrix [73], reweighting of the loss based on the reliability of samples [74, 75, 76], as well as the selection of clean data from noisy samples [77, 78]. In [77, 78], the authors proposed sample selection strategies based on the loss, but they require multiple deep neural networks to make the selection more robust, leading to an increase of the computational cost of the method. Furthermore, the sample selection is based on the loss, which is computed with noisy labels that are not refined during the training. This implies that the amount of noise in the labels will not be reduced and the selection will be highly affected by the noise. Lastly, NEL [79] combined a *negative learning* loss with a

pseudo-labels refinement framework based on ensembling. Negative learning [72] refers to an indirect learning method which uses complementary labels to combat noise.

2.1.5 Textual Guidance for Unsupervised Domain Adaptation

Typically, UDA approaches solely operate on the image space, which has been demonstrated to be suboptimal for complex shifts [80]. Differently, Chen et al. [81] proposed to integrate a large language model (LLM) for UDA. Instead, other works [82, 83, 84, 85] leveraged the language modality for domain generalisation, but they rely only on short class descriptors, which are not semantically rich as crowd-sourced text. For example, Goyal et al. [86] used class names as text descriptors, which are less semantically informative than free-form texts. To overcome these limitations, LaGTran [87] used image captions to generate pseudo-labels as a source of supervision for target samples. However, given the uncurated nature of the captions, the generated supervision may be incorrect, often leading to overfitting the introduced noise.

2.1.6 Visual Language Models in Knowledge Transfer

Vision-Language Models (VLMs), such as CLIP [88] and ALIGN [89], demonstrated great success in capturing modality-invariant features, opening new avenues for knowledge transfer. Some works [90] leveraged the zero-shot ability of VLMs, directly applying them to the target domain. In [91], the authors used prompt or adaptor learning to fine-tune the VLM to the target domain in a semi-supervised fashion. DAPL [92] learned domain-specific prompts to separate domain and class information in the CLIP visual feature space. DIFO [93] used prompt learning to adapt the VLM to the target domain and then distilled the VLM knowledge to a target model.

2.2 Learning from Weak Supervision

Weakly supervised learning (WSL) has garnered significant attention in the computer vision community in recent years. A wide range of methods have been proposed over the past decade to tackle complex tasks such as semantic segmentation [94], object detection [95] or 3D reconstruction [96], using only limited supervision. In WSL, the learning process relies on partial or coarse annotations—such as image-level labels or sparse object location information—available only for a subset of the training data. Compared to fully supervised learning, which requires extensive pixel-level or bounding-box annotations, WSL significantly reduces the burden of manual labeling by leveraging weaker forms of supervision. This is particularly advantageous in scenarios where acquiring fine-grained annotations is prohibitively expensive, time-consuming, or impractical at scale. This thesis focuses on exploring the WSL paradigm in the context of crowd density estimation.

2.2.1 Crowd Density Estimation

Density estimation has achieved great success in crowd counting due to its robustness in handling crowded scenes and complex backgrounds. From a computer vision perspective, perspective distortion, caused by varying filming angles and distances, poses a significant challenge. Some research has addressed this distortion by using perspective maps for additional supervision [97], while other studies have proposed perspective-map-free approaches [98]. In [98], a multi-column network with different convolutional kernel sizes in each branch of the model was introduced to capture multi-scale information. In [99], the authors proposed a context-aware model to adaptively handle scale variations. Recent works [100] also leverage transformer-based models to capture global relationships in crowd counting. More recently, *CrowdDiff* was proposed, where a diffusion model is trained to generate multiple hypotheses of crowd density [101]. These approaches require location-level supervision during training. Given the difficulty and cost of collecting such annotations, other methods have been developed to learn from limited supervision.

2.2.2 Crowd Density Estimation from Limited Location-level Annotations

Unsupervised strategies for crowd density estimation have been proposed, but they generally yield unsatisfactory performance [102]. Semi-supervised approaches, on the other hand, have achieved better results. L2R [103] leverages the abundant availability of unlabelled crowd images by learning to rank. MATT [104] uses both location-level and count-level annotations, promoting consistency between predictions generated by multiple auxiliary tasks. IRAST [105] is a self-training method that incorporates inter-relationships between different predictors to produce reliable pseudo-labels for semi-supervised learning. GP [106] is another semi-supervised crowd counting method that estimates pseudo-labels through a Gaussian Process-based iterative learning mechanism. AC-AL [107] introduces an active learning framework to gradually collect point annotations. A different approach is taken by PAL [108], where a deep neural network for density estimation is trained with partial annotations, requiring only a small, annotated region in each image. While semi-supervised methods reduce the reliance on extensive location-level annotations, they still necessitate a subset of location-level labels to produce meaningful density maps. Another way to alleviate the need for location-level annotations is through cross-domain learning, such as domain adaptation [109], where a model is trained on a fully annotated source domain and then adapted to an unlabelled target domain. Unlike these methods, COUNT2DENSITY generates meaningful density maps without requiring any location-level annotations.

2.2.3 Crowd Density Estimation from Count-level Annotations

Approaches that rely solely on count-level annotations are typically categorised as weakly supervised methods [110, 111]. These methods are primarily regression-based, meaning they predict the total number of people in a scene. Given their significance in the field, it is important to review the literature on regression-based approaches. In [112], the authors propose a regression-based weakly supervised method that uses a sorting network to improve crowd counting. In [113], multiple specialized CNNs are used, which are adaptively selected based on the appearance

of an image through a gating network. In [110], the authors introduce a strategy for crowd counting using a large number of ranking labels and a few images annotated with count labels. Transformer-based weakly supervised approaches have also been proposed, such as [114]. Although these methods predict accurate crowd counts, they infer density maps from feature activations. However, this approach to estimating density maps loses quantitative spatial information, as feature map values exhibit different statistical characteristics than density values, which impedes subregion counting. In contrast, our method leverages only count-level annotations and directly predicts meaningful density maps, where values are correlated with crowd density. This also enables subregion counting by integrating over areas of interest.

Chapter 3

Leveraging Pseudo-labels for Source-free Open-set Unsupervised Domain Adaptation

This chapter presents a novel source-free open-set unsupervised domain adaptation framework designed to address the challenges of lack of source data during the adaptation and the presence of unseen categories in the target domain. This chapter details the development and evaluation of a method that enable effective knowledge transfer from pretrained models without requiring access to source domain data, while robustly segregating unknown classes. In particular, we expose limitations of existing approaches that isolate novel classes in a single unknown category, encouraging a poor granularity in the learned features.

This work is the extension of our previous conference paper.[1] The finding and contributions of this chapter have been accepted to an international journal. [2]

3.1 Introduction

Deep learning methods achieve remarkable performance in visual tasks when training and test data share a similar distribution. However, learning approaches struggle to generalise in presence of *domain shift* [115, 116], *i.e.* data coming from different

[1] Mattia Litrico, Alessio Del Bue, Pietro Morerio, "Guiding Pseudo-Labels With Uncertainty Estimation for Source-Free Unsupervised Domain Adaptation". IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.

[2] Mattia Litrico, Davide Talon, Sebastiano Battiato, Alessio Del Bue, Mario Valerio Giuffrida, Pietro Morerio, "Uncertainty-guided Open-Set Source-Free Unsupervised Domain Adaptation with Target-private Class Segregation" [16]. International Journal of Computer Vision, 2025.

distributions, such as different illumination conditions, imaging modalities or styles. To this end, Domain Adaptation (DA) takes advantage of the knowledge learned on a *source domain* to aid the learning in a different yet related *target domain*. As the considerable cost of data collection and annotation limits the availability of large amounts of target data for supervised adaptation, DA focuses on leveraging unsupervised target data for knowledge transfer across domains. In particular, Unsupervised Domain Adaptation (UDA) aims to transfer the knowledge learned on a labelled source domain to an unseen target domain without requiring any target label [21, 117, 118, 119, 120, 121, 122].

Most UDA techniques typically rely on two strong assumptions on the data. On the one hand, they require simultaneous access to source and target data during the adaptation, hindering the deployment in applications where data privacy or transmission bandwidth become critical issues. To overcome this assumption, recent efforts focused on the setting of Source-free Domain Adaptation (SF-DA) [30, 34, 35, 36, 37], where *source data is no longer accessible* during the adaptation phase, but only unlabelled target data is available. On the other hand, most UDA algorithms assume a *closed-set* setting, *i.e.* the two domains share the same class space. Despite the adaptation to the new domain, the transferred source model has no capability to make predictions for novel/unseen classes. This severely limits their applicability in real-world scenarios [44, 123], where the target domain might contain unknown classes that are not present in the source data. To overcome this second assumption, Open-set Domain Adaptation (OSDA) [41, 42, 43, 44, 123] aims to empower the adapting model with the ability to recognise novel-class samples in the target domain as *unknown*. In this scenario, the target dataset contains both classes already seen in the source domain (*shared* classes), as well as novel classes that are not present in the source (*private* classes).

In [43, 44], the authors used adversarial learning to align the source domain with the portion of the target domain sharing the same classes, excluding the target private classes, leading to a poor decision boundary. To avoid this issue, Jang *et al.* [42] included target private classes in the adversarial learning to align the source domain with shared classes, while segregating private samples in unknown classes. However, all these previous works are not source-free, learning from source and target data simultaneously.

The Source-Free Open-set Domain Adaptation (SF-OSDA) addresses both challenges, assuming that the adaptation process has no access to source data in presence of private classes in the target domain. Although this joint setting is underexplored, some recent approaches [36, 49, 50] have been proposed. However, these approaches consider samples from target-private classes as outliers and cluster them into one "*unknown*" class. We hypothesize that gathering semantically different target-private samples into a single class induces a degradation of the feature space. Additionally, to segregate samples of unknown classes, previous approaches rely on a statistic-based separation criterion computed on the outputs of the adapted model, often leading to the so called *negative transfer*, a consequence of the well-known *confirmation bias* [124]. This means that the model relies on its noisy outputs during the adaptation process, increasing the confidence in wrong predictions. For all intents and purposes, the model tends to overfitting its own noise.

In this work, we propose a novel Source-free Open-set Domain Adaptation approach that, leveraging the granularity of target-private classes, segregates their samples into *multiple* clusters, resulting in a more effective adaptation compared to previous works. As a by product, this strategy also allows the discovery of the underlying semantics of novel classes. As an extension of [30], our method produces pseudo-labels for all target samples and progressively refines them using the consensus from neighbours samples. However, differently from [30], the initial pseudo-labels assignment (which also includes target private classes) is obtained by clustering the features space spanned by the source model, rather than directly using its predictions. A subset of the obtained clusters centroids is matched with the source model prototypes computed from the classifier weights. The unmatched centroids are treated as new prototypes of the target private classes and used to initialise the columns of the classifier weights corresponding to the unknown classes. Differently from the state-of-the-art, this strategy does not require any separation criterion to distinguish between samples of shared and unknown classes, limiting the negative effects of the confirmation bias.

The pseudo-labels produced during the adaptation are inevitably affected by the so called *shift noise* [125], namely the noise caused by the domain shift, here further intensified by the presence of novel classes in the target. Moreover, the initial clustering-based assignment could also introduce some noise, potentially leading to

classify an entire shared class as unknown or viceversa. To mitigate the effects of such noise, we perform an uncertainty-based samples selection to discard sample with unreliable (and possibly noisy) pseudo-labels. To that end, we propose a novel uncertainty estimation solution, which extends the one used in [30] that was not devised to handle this scenario. The uncertainty of the pseudo-labels is measured by analysing the consensus of neighbours samples predictions, and the relative distances of samples with respect to the class prototypes. Such newly introduced strategy assigns low uncertainty values when samples are close to one prototype, while being far from the others. Contrarily, when the distance values are similar, samples will have a high uncertainty and their pseudo-labels are considered unreliable.

To obtain a regularised features space where neighbours samples are semantically similar, we leverage a self-supervised contrastive learning framework together with a negative pairs exclusion strategy. Differently from [30], we introduce the *NL-InfoNCELoss*, a novel contrastive loss that integrates the principles of negative learning in the standard contrastive loss [29]. By empowering the standard loss with the well known benefits of the negative learning paradigm [126, 79], this formulation increases the robustness of the contrastive framework to the noise inevitably affecting the pseudo-labels.

Extensive experiments demonstrate the effectiveness of our method in the challenging setting of SF-OSDA, obtaining competitive results on three major benchmark datasets. Specifically, our method achieves comparable performance with the most recent baselines on Office31 [127] and Office-Home [128] and it sets the new state-of-the-art on the large-scale dataset VisDA-C. Ablation studies show the effectiveness of the new components in handling the issues introduced by the presence of novel classes in the target domain. Notably, additional analyses show the ability of our model to classify samples from target-private classes, meaning the model learned the underlying semantics of novel classes without being explicitly optimised for that task.

3.2 Proposed Method

This section presents our approach to the SF-OSDA problem in classification tasks, which is depicted in Fig. 3.1. Differently than other approaches [47, 39, 48], our

model is firstly trained on a source domain *without* using any ad-hoc training strategies. Then, at the beginning of the adaptation process, the pretrained source model extracts the features for each of the unlabelled target samples. A clustering-based strategy on target features roughly detects samples from private classes and provides an initial pseudo-labels assignment on the target samples (Section 3.2.2). Due to the domain shift between source and target domains and the presence of unknown class samples, this initial assignment is unreliable. Therefore, we mitigate the impact of the noise in the pseudo-labels by refining them and selecting reliable samples for the training. On the one hand, the nearest neighbors voting scheme in Section 3.2.3 progressively refines the pseudo-labels according to a majority rule. On the other, in Section 3.2.4 we select reliable pseudo-labels with low uncertainty, which are then used in the adaptation process. The contrastive scheme proposed in Section 3.2.6 enforces the underlying assumption of the pseudo-labels refinement process: samples of the same class should be close in the target feature space. In Section 3.2.5, we leverage the pseudo-labels to adapt the model to the target domain through a negative learning classification loss. In Section 3.2.7, we provide the overall training objective that progressively refines the noisy pseudo-labels and segregates target-private samples, while aligning the pretrained model to the target domain.

Problem formulation. Let $\mathcal{D}_s$ be the labelled source dataset including pairs $\{x_s, y_s\}$, where $x_s \in \mathcal{X}_s$ and $y_s \in \mathcal{Y}_s$ are images and ground-truth labels, respectively. Due to the source-free constraint, the source data $\mathcal{D}_s$ is not available during adaptation. Similarly, let $\mathcal{D}_t$ be the target data composed of images $\{x_t\}$ only, with $x_t \in \mathcal{X}_t$. The open-set scenario assumes that $\mathcal{Y}_s \subset \mathcal{Y}_t$, where $C_S := \mathcal{Y}_s \cap \mathcal{Y}_t$ and $C_P := \mathcal{Y}_t \setminus \mathcal{Y}_s$ are the set of shared and private classes, respectively. The set of all classes C is given by $C = C_S \cup C_P$. The source model $g_s(\cdot) = h_s(\varphi_s(\cdot))$ is composed of a features extractor $\varphi_s : \mathcal{X} \to \mathbb{R}^D$ and a classifier $h_s : \mathbb{R}^D \to \mathbb{R}^{|C_S|}$, where D is the size of the features space. Without loss of generality, we assume that the weights of the classifier are characterised by a matrix $W \in \mathbb{R}^{D \times |C_S|}$. We will refer to each column i of the weight matrix $W[i]$ as a shared class *prototype*.

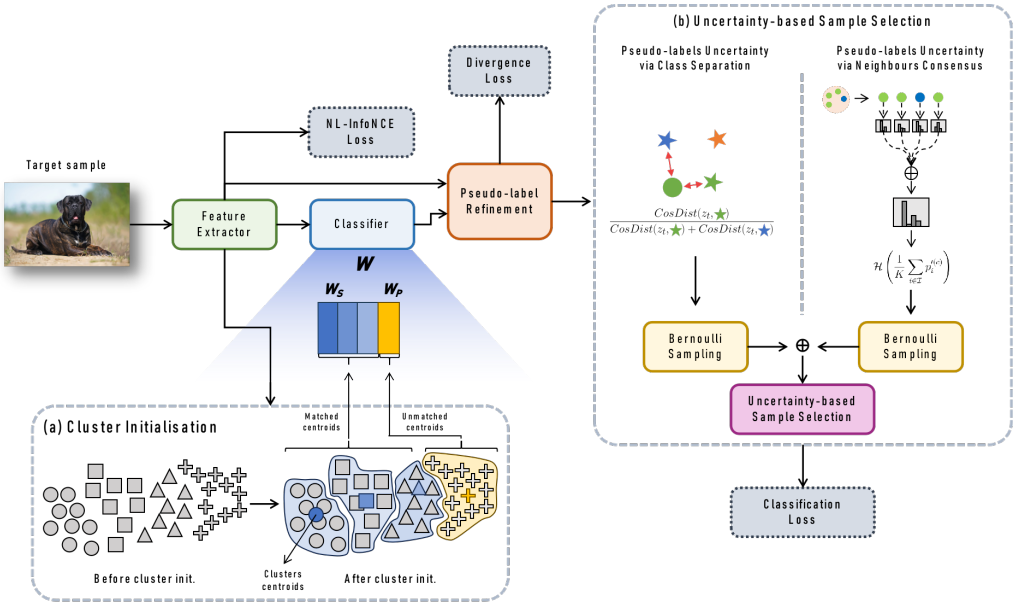

Figure 3.1: Overview of the proposed adaptation approach. **(a)** Target samples are clustered based on the features extracted from the pretrained source model, to provide an initial pseudo-labels assignment. Next, clustering allows the identification of class centroids (coloured shapes), which are matched against the most similar prototypes of shared classes. Such prototypes are taken as the columns of weight matrix W of the classifier (BLUE clusters). Since the target domain has more classes than the source domain, some clusters are left out from this matching, which will be treated as target-private classes (YELLOW clusters). **(b)** After the refining of the pseudo-label, its uncertainty is estimated using two approaches. *Neighbours consensus:* the uncertainty is determined by analysing the consensus of neighbours on the refined pseudo-label. *Class separation:* the model uncertainty is estimated based on the distance w.r.t. the two closest class prototypes. A novel contrastive loss (NL-InfoNCELoss) gathers same-class samples in the features space, while being robust to noisy pseudo-labels.

3.2.1 Leveraging the granularity of target-private classes

Open-set source-free domain adaptation considers a two-steps approach. In the first stage, a model is trained on large labelled source data. Without further access to the source samples, the second step adapts the source model to the unlabelled target domain. During the (pre-)training on the source dataset, the classifier h_s is trained to accommodate the classes in C_S, *i.e.* the shared classes. The target domain contains not only the classes in C_S, but also includes a set of private classes C_P.

Previous works usually train models to segregate samples from target-private

classes into a single "unknown" class, which is more affine to outlier-detection. While samples from shared classes are separated with respect to their semantics, this strategy forces the model to aggregate samples from target-private classes into a single class, independently from their semantics. Our hypothesis is that this affects the learned feature space, leading to a suboptimal geometry, where samples with possibly very different semantics are grouped together. Moreover, this strategy hinders the possibility to discover the underlying semantics of novel classes. To solve this issue, we design our model to leverage the granularity of target-private classes by segregating their samples into multiple unknown classes. This choice has the side effect of learning a features space where samples from private classes are aggregated based on their semantics, as discussed in Section 3.3.2. Hence, before the adaptation process, we craft a new classifier h_t whose weight matrix is made of two parts, such that $W_T = [W_S|W_P]$. The set of weights W_S is initialised by taking the pretrained weights from h_s, whereas W_P is randomly initialised with values sampled from a uniform distribution. This means that we augment the pretrained classifier by adding a certain number of columns matching a plausible number of private classes $|\widehat{C}_P|$ in the target domain. Note that $|\widehat{C}_P|$ is arbitrarily chosen, as the real cardinality of the target private classes is unknown. In Section 3.3.2, we analyse the performance with different values for $|\widehat{C}_P|$. Using the extended classifier h_t, we use the model $h_t(\varphi_s(\cdot))$ for an initial pseudo-label prediction in the target dataset $\mathcal{X}_t$ (Fig. 3.2(a)).

3.2.2 Clustering-based Target Model Initialisation

Since the pretrained model has never seen any of the private classes C_P and the classifier h_t is randomly initialised for its extended portion, most of the samples from private classes will be predicted into the shared classes. This will produce a certain amount of incorrect pseudo-labels, as shown in Fig. 3.2(a). To alleviate this problem, we propose a clustering-based solution to discover suitable prototypes for the private classes C_P, and thus initialize W_P in a more convenient way. This will produce a better initial pseudo-labels assignment. The idea of using a clustering approach builds on the observation that the features of target samples (both from shared and private classes) are characterised by a low intra-class variability (*c.f.* Fig. 3.2(a)). Therefore, a standard unsupervised clustering algorithm can, to some

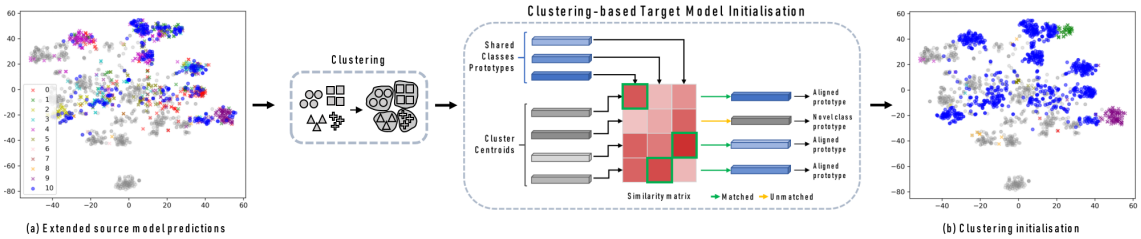

Figure 3.2: The extended source model h_t incorrectly classifies samples from private classes as belonging to shared ones. Nonetheless, samples from both shared and private classes exhibit a low intra-class variability. Building on this observation, we cluster the feature space to perform an initial pseudo-labels assignment on the target domain. The figure shows the initial identification of target-private samples using clustering. In (a) and (b) coloured samples belong to target-private classes, while GRAY samples belongs to shared classes. The colour of samples represents the predicted class. For visualisation purposes, all samples predicted in a target-private class are represented in the BLUE class. The • symbol represents correct predictions, whereas the × represents incorrect predictions. While the source model cannot correctly assign target-private samples, the clustering leverages the structure in the features space to aggregate private samples and provides a good pseudo-labels initialisation. In fact, after the clustering, more target-private samples are correctly classified in the BLUE class. Nonetheless, pseudo-labels still contain noise, which is progressively reduced during training.

extent, group both shared and private target samples, providing the model with an improved initial pseudo-labeling. Differently from other approaches [42, 43, 49], this strategy avoids to rely on a separation criterion that hardly distinguishes whether samples belong to shared or target-private classes.

However, a major drawback in utilising an unsupervised clustering algorithm is the *class misalignment*, as there is no correspondence between the clusters and the model classes and one needs to find the correct permutation which maps cluster ids to class ids.[3] To alleviate this problem, we leverage the pretrained source model g_s to align the shared classes C_S with a subset of clusters obtained by the clustering algorithm. The remaining unmatched clusters are treated as unknown classes and their centroids as novel prototypes for the initialisation of W_P, as depicted in Fig. 3.2.

More formally, we extract the features from the target samples $z_t = \varphi_s(x_t),\ \forall x_t \in \mathcal{X}_t$ using the pretrained source feature extractor φ_s. At this point, we perform an

[3]In the classic example of *cat vs. dog* classification, there is no guarantee that a clustering algorithm would assign cluster 1 to the class cat and cluster 2 to the class dog. As they can be easily inverted, such class misalignment becomes more exasperate in the presence of more than 2 classes.

unsupervised clustering over all the extracted features z_t to obtain K centroids $c_k = \frac{1}{|\mathcal{S}_k|}\sum_{a\in\mathcal{S}_k} z_t^a$, where $k \in \{1, 2, \ldots, K\}$, $K = |C_S| + |\widehat{C}_P|$, and $\mathcal{S}_k$ is the set of target samples assigned to the k-th cluster. Note that while $|\widehat{C}_S|$ is known, $|\widehat{C}_P|$ is arbitrarily chosen, since the real cardinality of the target private classes is unavailable under the open-set setting. In Sec. 3.3.2, we analyse the performance with different values for $|\widehat{C}_P|$ (which implies in different K).

To account for the class misalignment, we align the shared classes with a subset of the obtained clusters by computing the cosine similarity between the shared class prototypes $W_S[i]$ and the cluster centroids c_k, and we match each column in W_S with the centroid with the highest similarity. This strategy solves the class misalignment by determining a permutation that reorders the centroids c_k according to their match with shared class prototypes. Finally, centroids that have not been assigned to any shared classes prototypes will be considered as discovered prototypes of the target-private classes, and used to fill the columns of W_P (here the order of the clusters does not matter). The refined model $W_T = [W_S|W_P]$ is then used to produce an initial pseudo-labels assignment, as detailed in the next section.

3.2.3 Pseudo-Label Refinement with Neighbours Consensus

Similar to [30, 29], the refinement of the pseudo-labels is accomplished by aggregating knowledge from nearest neighbour samples. The underlying idea is that similar samples are likely to belong to the same class. We thus assume that features extracted from semantically similar samples lie close to each other in the feature space and we further enforce this assumption by means of a contrastive learning loss, as described in Sec. 3.2.6.

More formally, given a target sample x_t and a random weak augmentation t_{wa} drawn from a distribution $\mathcal{T}_{wa}$, we obtain a feature vector $z_t = \varphi_t(t_{wa}(x_t))$ from the augmented image, which is used to search the neighbours of x_t in the target features space [30]. The pseudo-label of x_t is refined by aggregating predictions from the neighbours by soft-voting [129] as follows:

$$\bar{p}_t^{(c)} = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} p_i'^{(c)}, \tag{3.1}$$

where $\mathcal{I}$ is the set of indices of the selected neighbours, p' is the softmax output, and the superscript c indicates the class index. To obtain a refined pseudo-label, we find the class c the maximum probability in Eq. (3.1), *i.e.* $\bar{y}_t = \arg\max_c \bar{p}_t^{(c)}$. The refined pseudo-label is assigned to the sample x_t and used as self-supervision signal (see Sec. 3.2.7).

This refining process relies on a bank $\mathcal{B}$ of length $M < |\mathcal{X}_t|$, storing pairs $\{z'_j, p'_j\}_{j=1}^{M}$ of features and softmax predictions. The neighbours of a sample x_t are then selected by computing the cosine distances between the features stored in the bank. The $|\mathcal{I}|$ samples with the lowest distance are selected as neighbours. Following [59, 30], we use a slowly changing momentum model $g'_t(\cdot) := \text{EMA}[h_t(\varphi_t(\cdot))]$ to update features z' and predictions p', in order to maintain the information stored in the bank stable during adaptation. A bank composed of M randomly selected samples reduces the number of pairwise comparisons needed for finding the neighbours.

Bank initialisation with discovered classes prototypes. As described above, the refinement of the pseudo-labels is performed by averaging neighbours probabilities stored in $\mathcal{B}$. In order to generate the initial pseudo-labels that reflect the clusters obtained in Sec. 3.2.2, we need to initialise the bank accordingly. However, the clustering strategy we use in Sec. 3.2.2 produces class assignments for all samples x_t rather than probabilities. To overcome this issue, after the clustering initialisation, we associate to each target sample a vector of probabilities proportional to its similarity to the centroids c_k. Given target features z_t and centroids c_k, we determined the probabilities vectors p' as the per-class similarity between z_t and each centroid in c_k. Formally, for each target sample, we obtain the probabilities vector p' to initialise the bank as follows:

$$p' = 1 - \frac{CosDist(z_t, c_k)}{\max\limits_{j \in \{1,\dots,|\hat{C}_P|\}} CosDist(z_t, c_k^j)} \tag{3.2}$$

where $CosDist(z_t, c_k) = 1 - \frac{z_t \cdot c_k}{||z_t||\,||c_k||}$ is the cosine distance and i is the class index. Finally, to obtain probabilities p'', the vector p' is provided to the softmax function σ with a temperature parameter τ_2, $p'' = \sigma(p'/\tau_2)$. Consequently, for all samples x_t the probability value for class c will be proportional to the similarity between the features $z_t = \varphi_t(x_t)$ and the centroid of c. Also, the class with the highest probability will be exactly the one assigned to the sample by the clustering strategy

in Sec. 3.2.2.

3.2.4 Pseudo-labels Uncertainty in the Open-set Scenarios

The refined pseudo-labels $\bar{y}_t$ are used as a self-supervision signal for a classification loss on target data, as detailed in Sec. 3.2.7. However, such pseudo-labels are not perfect, especially at the beginning of the iterative refining process, and thus still contain noise. Therefore, using all of them in the classification loss will disrupt adaptation, since the model would be trained with incorrect information. To mitigate this, we perform a sample selection, in order to select samples with only low uncertainty: these are selected stochastically via Bernoulli sampling, based on the uncertainty of their pseudo-label, estimated in a twofold manner, as detailed in subsequent paragraphs.

Uncertainty estimation via Neighbours Consensus. This first uncertainty estimation method is taken from our previous work [30]. It estimates the uncertainty of pseudo-labels (after their refinement) by measuring the consensus among neighbours' predictions. This strategy builds upon the hypothesis that if the neighbourhood agrees on a predicted class, the derived pseudo-labels should be considered reliable (low uncertainty). Otherwise, if the network predicts different classes for neighbour samples, the obtained pseudo-labels should be considered unreliable (high uncertainty). To this end, we compute the entropy $\mathcal{H}$ of $\bar{p}_t$ (Eq. (3.1)) as an estimator of pseudo-label uncertainty, noting that $\mathcal{H}$ yields low values in low uncertainty cases, and high values otherwise. We define the uncertainty coefficient u_t^{nc} as follows:

$$u_t^{nc} = \frac{\mathcal{H}(\bar{p}_t)}{\log_2 |C|}. \tag{3.3}$$

While the uncertainty coefficient u_t^{nc} reduces the impact of the noisy pseudo-labels in a closed-set scenario [30], we noted that it can fail to model uncertainty in open-set scenario. In fact, during the initial clustering assignment, a cluster made of target-private samples can be incorrectly assigned to any of the shared classes. This is mainly due to the fact that the model has not been optimised to separate shared and target-private classes during the pretraining on the source, implying that shared and target-private clusters can lie very close in the feature space. In this scenario,

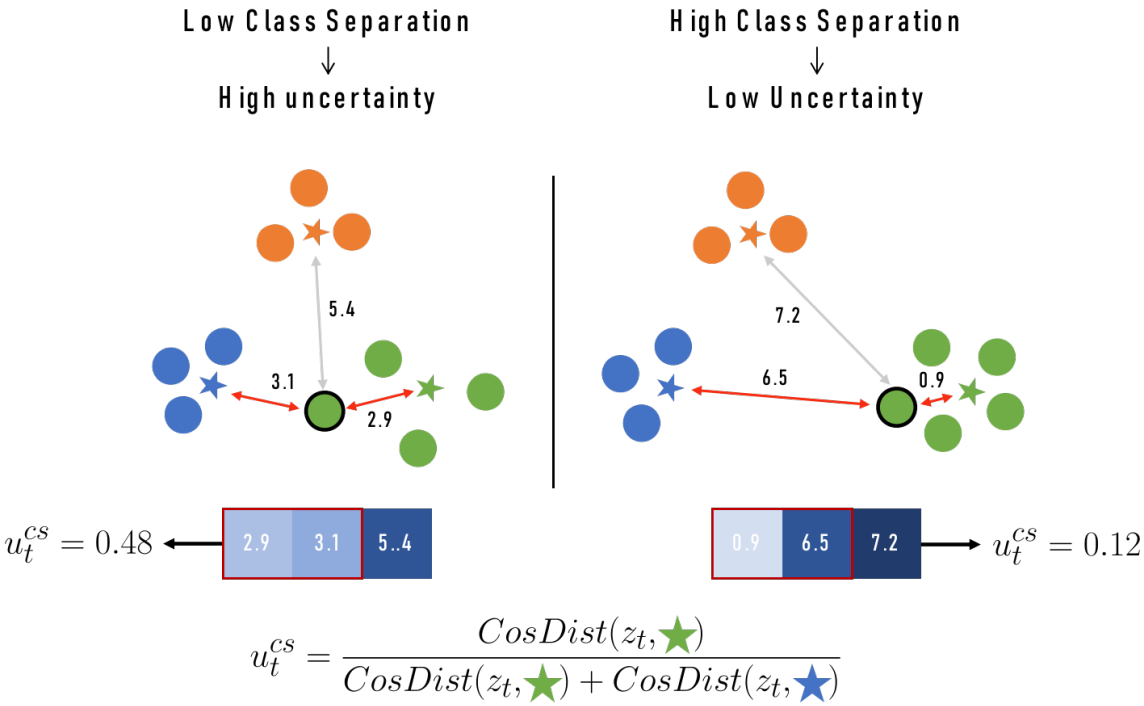

Figure 3.3: Pseudo-label uncertainty via Class Separation. The uncertainty of the pseudo-labels is estimated by computing the distance of samples from the two closest class prototypes. If a sample is similarly distant in the features space from two classes, its pseudo-label has high uncertainty. Contrarily, if a sample is close to one class and far from the other, its pseudo-label has low uncertainty and we consider it as reliable.

all the samples of the clusters will share the same incorrect pseudo-label, making the uncertainty estimation unsuccessful.

Uncertainty estimation via Class Separation. To tackle this problem, we propose a novel strategy to estimate the uncertainty of pseudo-labels that is well-suited in a open-set scenario, by analysing the distance of samples to class prototypes, as depicted in Figure 3.3. This new strategy is used in conjunction with the neighbour consensus. We build upon the hypothesis that when two classes are close in the features space, the distance of the samples from the two classes prototypes will be similar. Contrarily, when those classes are clearly separated, the samples will be close to one prototype and far from the other. Following this hypothesis, we consider the former and the latter scenarios of high and low uncertainty, respectively. In a situation where a cluster of target-private samples is so close to a shared class that can be switched, the distances of samples to prototypes of the two classes will

be similar, resulting in high uncertainty. More formally, for a target sample x_t we select from $\bar{p}_t$ the indexes $\{i, j\}$ of the two classes with the largest values as the most probable pseudo-labels for the sample. Then we compute the cosine distance between the features of the target sample $z_t = \varphi_t(x_t)$ and the prototypes $W[i]$ and $W[j]$. Finally, we obtain the uncertainty value u_t^{cs} as follows:

$$u_t^{cs} = \frac{CosDist(z_t, W[i])}{CosDist(z_t, W[i]) + CosDist(z_t, W[j])} \tag{3.4}$$

Bernoulli Sampling. To improve the model robustness to the noisy pseudo-labels for samples of both shared and target-private classes, we combine these two uncertainty estimation strategies for selecting samples with the lowest uncertainty to be used in the classification loss, as detailed in Sec. 3.2.7. Specifically, we compute the probabilities w_t^{nc} and w_t^{cs} for a target sample to have a correct pseudo-label with respect to both u_t^{nc} and u_t^{cs}, respectively:

$$w_t^{nc} = \mathcal{F}(u_t^{nc}) \tag{3.5}$$

$$w_t^{cs} = \mathcal{F}(u_t^{cs}) \tag{3.6}$$

where $\mathcal{F}$ is a monotonically decreasing function to assign high probabilities to samples with low uncertainty and viceversa. In Sec. 3.3.2, we analyse two choices for $\mathcal{F}$.

We select low uncertainty samples via Bernoulli sampling with probabilities w_t^{nc} and w_t^{cs}, which are combined as follows:

$$M_i = \mathrm{B}(w_t^{nc}) \oplus \mathrm{B}(w_t^{cs}) \tag{3.7}$$

where $\mathrm{B}(p) \in \{0, 1\}$ is the Bernoulli sampling with probability of success $0 \leq p \leq 1$ and $\oplus$ is a logical operator. In Sec. 3.3.2 we compare two choices for the operator $\oplus$.

Finally, we define $\mathcal{U} = \{x_t^i \in \mathcal{X}_t | M_i = 1\}$ as the set of uncertainty-based selected target samples. This set of samples is used during the adaptation process in the classification loss, as detailed in the next section.

3.2.5 Adaptation with Refined Pseudo-labels

The refined pseudo-labels obtained with neighbours' knowledge aggregation are used to compute a classification loss on target data to adapt the model to the new domain. We use the refined pseudo-labels $\bar{y}_t$ obtained from a weakly-augmented image $t_{\text{wa}}(x_t)$ as self-supervision for a strongly-augmented version $t_{\text{sa}}(x_t)$. The refining of the pseudo-labels is an iterative process, progressively improving the pseudo-labels accuracy during training through several steps.

As in our previous work [30], in addition to the proposed sample selection and exclusion strategies, to mitigate the effect of the noisy pseudo-labels, we use the negative learning loss [72, 126] as classification loss. Differently from [72, 126], which use the negative learning loss concurrently or alternating with a positive loss, here we do not use a positive loss in the entire training. The used classification loss is the following:

$$L_t^{cls} = -\,\mathbb{E}_{x_t \in \mathcal{U}} \Big[\sum_{c=1}^{C} \tilde{y}^c \, log\,(1 - p_{\text{sa}}^c) \Big], \tag{3.8}$$

where $\tilde{y}^c$ is a complementary label $\tilde{y} \in \{1, \dots, C\} \setminus \{\bar{y}_t\}$ chosen randomly from the set of labels and without the refined pseudo-labels, $p_{\text{sa}} = \sigma(h_t(\varphi_t(t_{\text{sa}}(x_t))))$ is the probabilistic output for the strongly-augmented image $t_{\text{sa}}(x_t)$, and $\mathcal{U}$ is the set of samples with low uncertainty, as detailed in Sec. 3.2.4. The random selection of the complementary labels $\tilde{y}$ is coherent with the negative learning framework [72].

3.2.6 Contrastive Loss for Open-set Scenarios

As described in Sec. 3.2.3, the process of pseudo-labels refinement via neighbours samples assumes that features extracted from same-class samples lie closer in the features space, compared to samples from other classes. We promote this assumption during adaptation via self-supervised contrastive learning. We build on MoCo [59] to aggregate features from different augmentations of the same image and separate representations originated from different samples (negative pairs). Specifically, for each sample x_t, we generate two strongly-augmented samples $t_{sa}(x_t)$ and $t'_{sa}(x_t)$, where $t_{sa}, t'_{sa} \in \mathcal{T}_{\text{sa}}$ are two randomly sampled transformations from the space of strong augmentations $\mathcal{T}_{sa}$. To build the constrastive pairs in the feature space, the feature extractor embeds augmented samples as query $q = \varphi_t(t_{sa}(x_t))$ and key

$k = \varphi_t(t'_{sa}(x_t))$. While we employ representations q and k associated to the same sample x_t as positive pairs, we build negative pairs by maintaining a queue Q_e that stores key features $\{k^{(i)}\}_{i=1}^{N}$ as detailed below.

MoCo [59] uses the pairs (q, k) as positive and all the pairs $\{(q, k^{(i)})\}_{i=1}^{N}$ as negative pairs by minimising and maximising their cosine distance for positive and negative pairs, respectively. Since features are stored in Q_e independently from the class, even features from samples belonging to the same class will be pushed away, which is in contrast with our objective to better aggregate same-class samples in the feature space. In [29], authors propose a strategy to exclude some negative pairs from the contrastive loss. For every negative pair, they just compare the pseudo-labels of the two samples. If they share the same pseudo-label, then the negative pair is excluded; otherwise, the pair is included in the negative pairs list. However, the exclusion strategy in [29] does not take into account the noise that inevitably affects the pseudo-labels during adaptation. Indeed, such noise may still lead to wrong negative pairs, *i.e.*, pairs classified differently but having the same ground-truth label. This causes instability of standard contrastive frameworks, highly affecting the learned features space. Hence, we follow the temporal exclusion strategy in [30], leveraging past predictions to identify and exclude pairs composed of same class samples, even with noisy pseudo-labels. This exclusion strategy is based on the intuition that, rather than only looking at the current pseudo-label, it is more reliable to look at its past history to have a higher probability of observing, at least once, the correct label: the history will probably reveal the correct one, improving the exclusion process. To this end, we build a temporal queue Q_e by storing, for each sample, also the refined pseudo-labels $\{\bar{y}^{(j)}\}_{j=1}^{\tau}$ of the τ past epochs, *i.e.*, $\{e-\tau, \ldots, e-1\}$, and we filter out all the pairs that shared the same pseudo-labels at least once in the past τ epochs from the set of available negative pairs, as illustrated in Figure 3.4. Although this process will likely filter out a lot of true negatives, it ensures most of the false negatives are excluded.

NL-InfoNCELoss. The refining of the pseudo-labels is an iterative process, which implies that an amount of noise is still present in the pseudo-labels. Moreover, the presence of target-private classes in the open-set setting, which are hard to be correctly identified, increases such noise with respect to the closed-set case. Inspired by [72] and differently from our previous work [30], we propose a novel contrastive

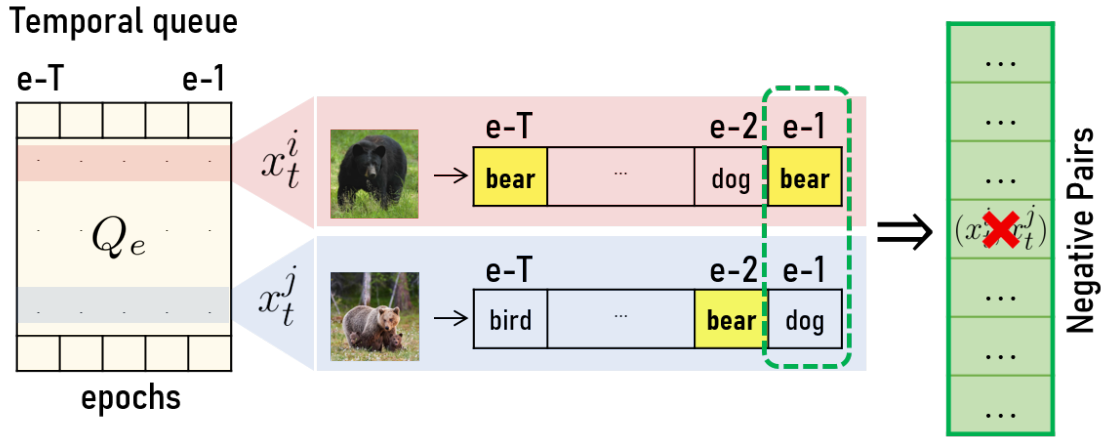

Figure 3.4: A couple of target images would be wrongly considered as a negative pair if only comparing the latest predictions (green box). Instead, since x_t^i and x_t^j share the same pseudo-labels (at least once) in the past T epochs, i.e. $\{e-T, ..., e-1\}$, we exclude them from the list of negative pairs. Figure from [30], reproduced with the permission of the authors.

loss, called NL-InfoNCELoss, that inserts the principle of negative learning in the standard InfoNCELoss. Although the strategies proposed in Sec. 3.2.4 and Sec. 3.2.6 aim at reducing the impact of such noise, we posit that using the negative learning paradigm can be beneficial for a contrastive loss to reduce the impact of false negative samples.

Accordingly, the formulation of the proposed NL-InfoNCELoss is the following:

$$L_{\text{NL-InfoNCE}} = -\,log\left(1 - \frac{exp(q \cdot k_-/\tau)}{\sum_{j \in \mathcal{N}_q} exp(q \cdot k_j/\tau)}\right), \tag{3.9}$$

$$\mathcal{N}_q = \{j | \bar{y}_j^{(i)} \neq \bar{y}^{(i)},\ \forall j \in \{1, \ldots, N\}, \forall i \in \{1, \ldots, \tau\}\},$$

where k_- is one negative sample chosen randomly from the set of negative samples, $\mathcal{N}_q$ is the set of indices of samples in Q_e that never shared with the query sample the same pseudo-labels in the past τ epochs. By optimising the proposed NL-InfoNCELoss, the model is trained to increase the feature distance between the query and a randomly selected negative sample, rather than to all the negative samples, as in the standard InfoNCELoss. Albeit this formulation slackens the separation effect of the contrastive loss, it reduces the impact of noisy pseudo-labels in the negative pairs. In Sec. 3.3.2, we analyse the contribution of the proposed loss

compared to the standard InfoNCELoss.

3.2.7 Overall Framework

In addition to the classification and the contrastive losses detailed in the previous sections, following the standard state-of-the-art protocol [30], we add the following regularisation term to prevent the posterior collapse:

$$L_t^{div} = \mathbb{E}_{x_t \in \mathcal{X}_t}\Big[\sum_{c=1}^{C} \bar{p}_q^c \, log \, \bar{p}_q^c\Big],$$

$$\bar{p}_q = \mathbb{E}_{x_t \in \mathcal{X}_t}\left[\sigma\left(h_t\left(\varphi_t\left(t_{sa}\left(x_t\right)\right)\right)\right)\right].$$

When trained with noisy labels, the model may learn a degenerate latent representation, leading to predict all the samples in a single class, especially if the noise is very skewed towards a single category. By encouraging diversity in model predictions, this term mitigates the occurrence of such phenomenon.

The overall objective function used to adapt the model to the target data is the following:

$$L_t = \gamma_1 L_t^{cls} + \gamma_2 L_t^{ctr} + \gamma_3 L_t^{div}, \tag{3.10}$$

where $\gamma_1 = \gamma_2 = \gamma_3 = 1$ are not-tuned hyperparameters.

3.3 Experiments

Datasets. We evaluated our approach performing experiments on the following benchmark datasets for the open-set source-free domain adaptation setting: Office31 [127], Office-Home [128], and VisDA-C [130].

– ***Office31*** contains three domains, Amazon (A), Dslr (D), and Webcam (W), with 31 classes, where the first 10 classes are shared and the last 11 private classes. The remaining 10 classes are not used in this setting.

– ***Office-Home*** consists of 65 classes of labelled images deriving from four specific domains, Art (Ar), Clipart (Cl), Product (Pr), and Real World (Rw). The first 25 categories in alphabetic order are shared classes and the remaining 40 classes are private.

– ***VisDA-C*** is a large-scale dataset containing images from two distinct domains: synthetic (S) and real (R). It contains 12 classes with a total of 280,000 images, where the first 6 classes represent shared categories, and the remaining 6 classes are private.

Evaluation metrics. Following [131, 41, 43], we use three evaluation metrics to compare the performance, including the average class accuracy over known classes OS^*, the accuracy of unknown class UNK, and $\text{HOS} = 2\frac{\text{OS}^* \times \text{UKN}}{\text{OS}^* + \text{UNK}}$ which is the harmonic mean between OS^* and UNK. HOS is the core metric in the latest open-set domain adaptation literature, since it requires a good balance on both shared and private class accuracy in an unbiased manner.

Implementation details. We use standard classification architectures comprising a feature extractor followed by a classifier. For fair comparison with competing methods, we used ResNet-50 [132] as a backbone in the experiments.

For source training, we initialise the ResNet backbone with ImageNet-1K [133] pre-trained weights available in the Pytorch model zoo. We train the source model with *only a standard cross-entropy loss*. For the adaptation stage, the target model is initialised with the source model's parameters. For open-set experiments, before the adaptation, we increase the classifier size to $2 \cdot |C_S|$ to account for the unknown classes. The initial clustering is performed with the K-means algorithm of Scikit-learn [134].

3.3.1 Results

This section presents results obtained by our method on Office-Home, Office31 and VisDA-C under SF-OSDA

Office-Home. Table 3.1 shows the performance of our method w.r.t. state-of-the-art UDA and SF-UDA methods on Office-Home under the open-set setting. Although our method does not use source data during the adaptation, we outperform several UDA methods, such as STA [43], OSBP [44], DAOD [136], and DANCE [62]. While the best average performance is achieved by GLC [51], our method surpasses some recent source-free approaches such as AaD [49] and USD [50] in both HOS

Table 3.1: HOS score (%) on Office-Home. All methods use the ResNet-50 backbone. The proposed approach achieves the highest accuracy on average (Avg.) compared with SF-UDA baselines and outperforms several UDA methods. The underlined results are the best within standard UDA methods, while bold ones are the best SF-UDA approaches.

Office-Home																				
Method	Source-free	Ar → Cl			Ar → Pr			Ar → Rw			Cl → Ar			Cl → Pr			Cl → Rw		
		OS*	UNK	**HOS**	OS*	UNK	**HOS**	OS*	UNK	**HOS**	OS*	UNK	**HOS**	OS*	UNK	**HOS**	OS*	UNK	**HOS**
PGL [135]	✗	63.3	19.1	29.3	78.9	32.1	45.6	87.7	40.9	55.8	85.9	5.3	10.0	73.9	24.5	36.8	70.2	33.8	45.6
STA [43]	✗	50.8	63.4	56.3	68.7	59.7	63.7	81.1	50.5	62.1	53.0	63.9	57.9	61.4	63.5	62.5	69.8	63.2	66.3
OSBP [44]	✗	50.2	61.1	55.1	71.8	59.8	65.2	79.3	67.5	72.9	59.4	70.3	64.3	67.0	62.7	64.7	72.0	69.2	70.6
DAOD [136]	✗	72.6	51.8	60.5	55.3	57.9	56.6	78.2	62.6	69.5	59.1	61.7	60.4	70.8	52.6	60.4	77.8	57.0	65.8
OSLPP [45]	✗	55.9	67.1	61.0	72.5	73.1	72.8	80.1	69.4	74.3	49.6	79.0	60.9	61.6	73.3	66.9	67.2	73.9	70.4
ROS [41]	✗	50.6	74.1	60.1	68.4	70.3	69.3	75.8	77.2	76.5	53.6	65.5	58.9	59.8	71.6	65.2	65.3	72.2	68.6
DANCE [62]	✗	54.4	53.7	53.1	82.2	35.4	49.8	87.4	25.3	39.4	71.2	28.4	40.9	74.6	32.8	45.9	81.3	18.4	30.2
cUADAL [42]	✗	55.8	75.6	63.6	69.6	73.9	71.6	81.8	73.3	77.5	54.9	82.0	65.0	61.8	77.4	68.3	69.5	76.3	72.6
ANNA [131]	✗	61.4	78.7	69.0	68.3	79.9	73.7	74.1	79.7	76.8	58.0	73.1	64.7	64.2	73.6	68.6	66.9	80.2	73.0
GATE [137]	✓	-	-	63.8	-	-	70.5	-	-	75.8	-	-	66.4	-	-	67.9	-	-	71.7
SHOT [36]	✓	**67.0**	28.0	39.5	**81.8**	26.3	39.8	**87.5**	32.1	47.0	**66.8**	46.2	54.6	**77.5**	27.2	40.2	**80.0**	25.9	39.1
AaD [49]	✓	50.7	66.4	57.6	64.6	69.4	66.9	73.1	66.9	69.9	48.2	**81.1**	60.5	59.5	63.5	61.4	67.4	68.3	67.8
USD [50]	✓	53.3	71.5	61.1	65.7	74.9	70.0	73.3	79.5	76.3	52.2	70.8	60.1	62.4	68.4	65.2	69.3	68.6	68.9
UMAD [138]	✓	-	-	59.2	-	-	71.8	-	-	76.6	-	-	**63.5**	-	-	69.0	-	-	71.9
GLC [51]	✓	-	-	**65.3**	-	-	**74.2**	-	-	**79.0**	-	-	60.4	-	-	**71.6**	-	-	**74.7**
LEAD [52]	✓	-	-	60.7	-	-	70.8	-	-	76.5	-	-	61.0	-	-	68.6	-	-	70.8
Ours	✓	46.6	**72.8**	56.9	61.1	**80.0**	69.3	74.1	**84.6**	**79.0**	44.8	62.4	52.7	62.7	**69.6**	66.0	69.5	**76.9**	73.0

Method	Pr → Ar			Pr → Cl			Pr → Rw			Rw → Ar			Rw → Cl			Rw → Pr			**Average**		
	OS*	UNK	**HOS**	OS*	UNK	**HOS**	OS*	UNK	**HOS**	OS*	UNK	**HOS**	OS*	UNK	**HOS**	OS*	UNK	**HOS**	OS*	UNK	**HOS**
PGL [135]	73.7	34.7	47.2	59.2	38.4	46.6	84.8	27.6	41.6	81.5	6.1	11.4	68.8	0.0	0.0	84.8	38.0	52.5	76.1	25.0	35.2
STA [43]	54.2	72.4	61.9	44.2	67.1	53.2	76.2	64.3	69.5	67.5	66.7	67.1	49.9	61.1	54.5	77.1	55.4	64.5	61.8	63.3	61.1
OSBP [44]	59.1	68.1	63.2	44.5	66.3	53.2	76.2	71.7	73.9	66.1	67.3	66.7	48.0	63.0	54.5	76.3	68.6	72.3	64.1	66.3	64.7
DAOD [136]	71.3	50.5	59.1	58.4	42.8	49.4	81.8	50.6	62.5	66.7	43.3	52.5	60.0	36.6	45.5	84.1	34.7	49.1	69.6	50.2	57.6
OSLPP [45]	54.6	76.2	63.6	53.1	67.1	59.3	77.0	71.2	74.0	60.8	75.0	67.2	54.4	64.3	59.0	78.4	70.8	74.4	63.8	71.7	67.0
ROS [41]	57.3	64.3	60.6	46.5	71.2	56.3	70.8	78.4	74.4	67.0	70.8	68.8	51.5	73.0	60.4	72.0	80.0	75.7	61.6	72.4	66.2
DANCE [62]	69.7	43.9	54.2	48.9	67.4	55.7	84.2	27.1	41.2	76.8	16.7	27.5	59.4	41.3	48.3	84.1	29.6	44.0	72.8	35.0	44.2
cUADAL [42]	52.1	82.4	62.9	42.7	80.7	54.6	71.7	83.4	76.8	67.3	79.6	72.6	52.5	71.1	59.9	77.7	75.6	76.7	63.1	77.6	68.5
ANNA [131]	63.0	70.3	66.5	54.6	74.8	63.1	74.3	78.9	76.6	66.1	77.3	71.3	59.7	73.1	65.7	76.4	81.0	78.7	65.6	76.7	70.7
GATE [137]	-	-	67.3	-	-	61.5	-	-	76.0	-	-	70.4	-	-	61.8	-	-	75.1	-	-	69.0
SHOT [36]	**66.3**	51.1	57.7	59.3	31.0	40.8	**85.8**	31.6	46.2	**73.5**	50.6	59.9	**65.3**	28.9	40.1	**84.4**	28.2	42.3	**74.6**	33.9	45.6
AaD [49]	47.3	**82.4**	60.1	45.4	72.8	55.9	68.4	72.8	70.6	54.5	**79.0**	64.6	49.0	69.6	57.5	69.7	70.6	70.1	58.2	71.9	63.6
USD [50]	54.3	73.8	62.6	47.3	69.6	56.3	70.0	74.5	72.2	64.6	71.3	**67.8**	53.8	65.5	59.1	73.3	69.1	71.1	61.6	71.5	65.9
UMAD [138]	-	-	62.5	-	-	54.6	-	-	72.8	-	-	66.5	-	-	57.9	-	-	70.7	-	-	66.4
GLC [51]	-	-	63.7	-	-	63.2	-	-	**75.8**	-	-	67.1	-	-	**64.3**	-	-	**77.8**	-	-	**69.8**
LEAD [52]	-	-	**65.5**	-	-	59.8	-	-	74.2	-	-	64.8	-	-	57.7	-	-	75.8	-	-	67.2
Ours	52.3	63.7	57.4	**65.3**	**80.7**	**72.2**	66.8	**78.6**	72.3	51.4	77.4	61.8	50.3	**74.6**	60.1	69.7	**73.1**	71.4	59.5	**74.5**	66.0

and UNK scores. Moreover, we achieve the best performance in UNK on multiple settings, showing ability in identifying private samples at the risk of including also some shared samples. This also shows a limitation of our method in small-scale datasets.

Office31. Table 3.2 compares our method with state-of-the-art UDA and SFDA methods on the Office31 dataset under the open-set setting. For the UDA setting, even though our method does not use source data at all during adaptation, it achieves comparable performance with other methods. Note that the availability of source data in the UDA setting is a considerable advantage, given that they can be used for aligning domains. For the more challenging SF-UDA setting, we achieve a competitive HOS score, outperforming recent baselines, such as USD [50] and

Table 3.2: HOS score (%) on Office31 and VisDA-C. All methods use the ResNet-50 backbone. The proposed approach outperforms the previous SF-UDA state-of-the-art by 2.1% on average (Avg.), while being competitive with UDA baselines. The underlined results are the best within UDA methods, while the bold one are the best SF-UDA approaches.

		Office31																						VisDA-C
Method	Source-free	A → D			A → W			D → A			D → W			W → A			W → D			**Average**			**Average**
		OS*	UNK	**HOS**	OS*	UNK	**HOS**	OS*	UNK	**HOS**	OS*	UNK	**HOS**	OS*	UNK	**HOS**	OS*	UNK	**HOS**	OS*	UNK	**HOS**	**HOS**
STA [43]	✗	91.0	63.9	75.0	86.7	67.6	75.9	83.1	65.9	73.2	94.1	55.5	69.8	66.2	68.0	66.1	84.9	67.8	75.2	84.3	64.8	72.6	71.0
OSBP [44]	✗	90.5	75.5	82.4	86.8	79.2	82.7	76.1	72.3	75.1	97.7	96.7	97.2	73.0	74.4	73.7	99.1	84.2	91.1	87.2	80.4	83.7	52.3
UAN [139]	✗	95.6	24.4	38.9	95.5	31.0	46.8	93.5	53.4	68.0	99.8	52.5	68.8	94.1	38.8	54.9	81.5	41.4	53.0	93.4	40.3	55.1	-
OSLPP [45]	✗	92.6	90.4	91.5	89.5	88.4	89.0	82.1	76.6	79.3	96.9	88.0	92.3	78.9	78.5	78.7	95.8	91.5	93.6	89.3	85.6	87.4	-
ROS [41]	✗	87.5	77.8	82.4	88.4	76.7	82.1	74.8	81.2	77.9	99.3	93.0	96.0	69.7	86.6	77.2	100.0	99.4	99.7	86.6	85.8	85.9	66.5
DANCE [62]	✗	94.3	50.7	66.9	92.8	55.9	70.7	97.0	66.8	80.0	97.6	73.7	84.8	82.4	85.3	65.8	81.6	60.6	70.2	91.0	60.2	73.1	67.5
cUADAL [42]	✗	86.4	95.1	90.1	86.0	90.4	87.9	98.6	97.7	98.2	99.3	99.4	99.4	75.4	87.8	80.5	67.6	87.8	75.1	85.6	93.0	88.5	-
ANNA [131]	✗	93.2	76.1	83.8	82.8	88.4	85.5	75.4	91.1	82.5	99.4	99.6	99.5	76.0	87.9	81.6	100.0	96.8	98.4	87.8	90.0	88.6	-
GATE [137]	✗	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	89.5	70.8
SHOT [36]	✓	**94.0**	46.3	62.0	**95.6**	42.3	58.7	**83.3**	39.1	53.3	**100.0**	75.7	86.1	**82.7**	46.6	59.6	**100.0**	69.7	82.1	**92.6**	53.3	67.0	28.1
AaD [49]	✓	73.0	84.6	78.3	63.5	**89.5**	74.3	63.6	**88.9**	74.2	78.0	**98.5**	87.0	61.9	**88.9**	73.0	94.6	**96.8**	95.7	72.4	**91.2**	80.4	42.4
USD [50]	✓	90.7	73.4	81.2	82.8	72.7	77.9	65.7	84.4	73.9	97.9	96.6	**97.3**	64.6	86.7	74.0	98.0	92.6	95.2	83.3	84.4	83.3	69.4
UMAD [138]	✓	-	-	88.5	-	-	84.4	-	-	86.8	-	-	95.0	-	-	88.2	-	-	95.9	-	-	89.8	66.8
GLC [51]	✓	-	-	82.6	-	-	74.6	-	-	**92.6**	-	-	96.0	-	-	**91.8**	-	-	96.1	-	-	89.0	72.5
LEAD [52]	✓	-	-	84.9	-	-	**85.1**	-	-	90.2	-	-	94.8	-	-	90.3	-	-	96.5	-	-	**90.3**	74.2
Ours	✓	85.7	**93.0**	**89.2**	90.9	77.8	83.9	73.3	83.4	78.0	89.9	95.7	93.2	65.0	80.0	71.8	98.6	94.6	**96.6**	83.9	89.6	85.4	**75.1**

AaD [49], by a notable margin of +2.1%, +5.0%. Although LEAD [52] achieves the average best result, our method, similar to GLC [51] and LEAD [52], achieves the highest HOS score in two source-target configurations, demonstrating to obtain a better balance between OS∗ and UNK. On the contrary, previous methods perform well on one but highly degrade on the other. For example, SHOT often achieves the best result on OS∗, but performs poorly on UNK, yielding a lower HOS score. Similarly, AaD [49] obtains higher result on UNK but lower on OS∗.

VisDA-C. Table 3.2 compares our method with state-of-the-art UDA and SFDA approaches on the VisDA-C dataset under the open-set setting. Our method consistently outperforms several recent UDA methods, such as STA [43], OSBP [44], and DANCE [62]. More importantly, in the source-free setting, our approach achieves the best HOS score, surpassing the top-performing baselines, LEAD [52], GLC [51], and USD [50], by significant margins of +0.9%, +2.6%, and +5.7%, respectively. These results demonstrate the effectiveness of our approach on large-scale datasets, highly fitting for real-world applications.

3.3.2 Analysis

Ablation study. Our method comprises several components to effectively adapt to a target domain in the context of source-free adaptation. Table 3.3 shows the benefit

Table 3.3: Ablation study of the components of the proposed method measured by HOS score (%) on Office31 and VisDA-C. First row excludes all the components. Second row includes the clustering initialisation Sec. 3.2.2. Third row and fourth row include the two strategies in Sec. 3.2.4 for estimating the uncertainty of pseudo-labels, respectively. Last row introduces the proposed NL-InfoNCELoss Sec. 3.2.6.

Clustering init.	Neighbours Consensus Uncertainty Estimation	Class Separation Uncertainty Estimation	NL-InfoNCELoss	Office31	VisDA-C
✗	✗	✗	✗	35.3	40.1
✓	✗	✗	✗	77.9	70.0
✓	✓	✗	✗	80.6	72.7
✓	✓	✓	✗	82.4	73.9
✓	✓	✓	✓	**85.4**	**75.1**

of these components: clustering initialisation (*c.f.* Sec. 3.2.2), neighbours consensus and class separation uncertainty estimations (*c.f.* Sec. 3.2.4), and the proposed NL-InfoNCELoss (*c.f.* Sec. 3.2.6). Clustering initialisation brings a major benefit to the training performance, doubling the average accuracy. The benefit in performance of estimating uncertainties is approx. +2% each. The best performance is achieved when also including the NL-InfoNCELoss, with a performance gain of approx. +3%.

Cardinality of target-private classes. As detailed in Sec. 3.2.2, the number of the target-private classes $|C_P|$ is unknown during adaptation. We set the expected number of target-private $|\widehat{C}_P|$ arbitrarily and we use this value to both extend the size of the classifier output and perform the initial clustering. Here, we study the performance of our method with different values of $|\widehat{C}_P|$. Results of this experimentation are shown in Figure 3.5. Particularly, the best performance is obtained when the number of target private classes is equal to the cardinality of source classes, *i.e.* $|C_P| = |C_S|$. Nonetheless, using even higher values for $|C_P|$, our method achieves good performance obtaining an HOS score close to the one achieved with the optimal value of C_P. Contrarily, the worst performance is obtained using $C_P = 1$. This results strengthen our hypothesis that predicting all the samples from target-private classes in a unique unknown class is suboptimal, since it requires aggregating semantically different samples together. Following this, we design our method to segregate these samples into multiple unknown classes. In the appendix (Section 3.3.3), we study the opposite scenario where $|\widehat{C}_P|$ is fixed while $|C_P|$ changes.

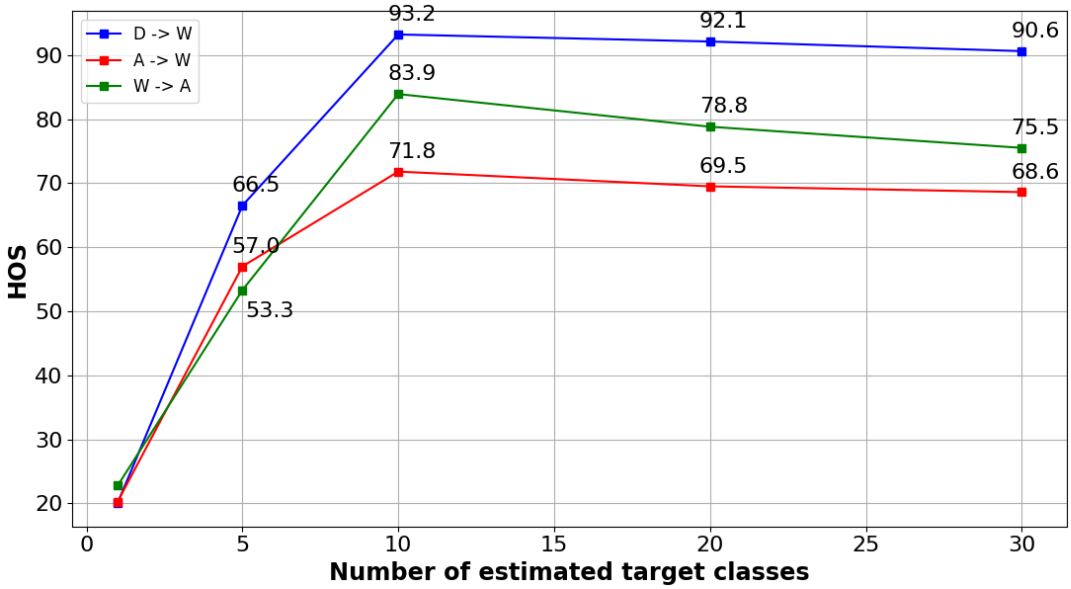

Figure 3.5: *HOS* score on three sub tasks of Office31 (%) with a various number of estimated target classes $|C_P|$. The best performance is obtained with $|C_P| = |C_S|$ target classes.

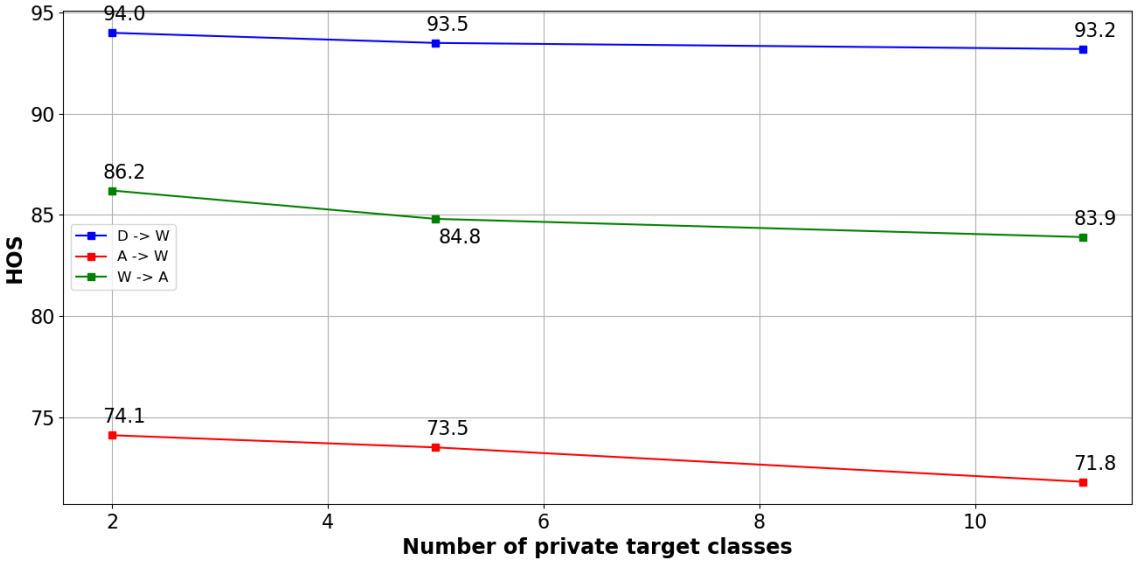

Figure 3.6: HOS score on three sub tasks of Office31 (%) with a various number of target-private classes $|C_P|$, maintaining fixed the number of estimated target-private classes $|\widehat{C}_P| = |C_S| = 10$.

Table 3.4: HOS score (%) on three sub-tasks of Office31 and VisDA-C using different choices of the function $\mathcal{F}$ in Eqs. (3.5) and (3.6).

		Office31				VisDA-C
Eq. (3.5)	Eq. (3.6)	D → W	A → W	D → A	Avg.	Avg.
$\mathcal{F}_{\text{lin}}$	$\mathcal{F}_{\text{lin}}$	91.5	79.0	76.2	82.2	73.3
$\mathcal{F}_{\text{lin}}$	$\mathcal{F}_{\text{exp}}$	90.5	77.4	77.8	81.9	73.1
$\mathcal{F}_{\text{exp}}$	$\mathcal{F}_{\text{exp}}$	93.1	83.6	78.0	84.9	74.6
$\mathcal{F}_{\text{exp}}$	$\mathcal{F}_{\text{lin}}$	93.2	83.9	78.0	**85.0**	**75.1**

3.3.3 Varying the Number of Target-private Classes

Here we study the performance of our method with different values of target-private classes $|C_P|$, while maintaining fixed the number of estimated target-private classes $|\widehat{C}_P| = |C_S| = 10$. As shown in Fig. 3.6, when the number of target-private classes is reduced, our model increases performance, as the setting become much closer to closed-set as $|C_P|$ is lower. When $|C_P|$ is increased, a drop in performance is observed but this behavior aligns with expectations, as more unseen classes are introduced on the target domain.

From uncertainty to selection probabilities. In Eqs. (3.5) and (3.6), the uncertainty coefficient w_t^{nc} and w_t^{cs} are estimated with the use of a monotonically decreasing function $\mathcal{F}$. We made two choices for this function: (i) a linear function $\mathcal{F}_{\text{lin}}(x) = 1 - x$; and (ii) an exponential function $\mathcal{F}_{\text{exp}}(x) = e^{-x}$.

Table 3.4 shows the effect of using either or both functions to calculate the uncertainty coefficients. Our model achieves comparable performance with all the combinations of these functions. As reported in [30], the exponential function works better for Eq. (3.5), as it avoids to penalise samples near the boundaries. For Eq. (3.6), the effect of the choice of $\mathcal{F}$ is marginal, as the performance drop is approx. 0.1% and 0.5%, respectively. A motivation of this behaviour is that the minimum value for the coefficient u_t^{cs} is 0.5, which avoids over-penalising samples when $\mathcal{F}_{\text{lin}}$ is used.

Combining uncertainty estimation strategies. In Sec. 3.2.4, we estimate the uncertainty coefficients w_t^{nc} and w_t^{cs}, which are used as the success probability to sample values from Bernoulli distributions. Sampled values are combined together via a logical operator $\oplus$ (Eq. (3.7)). We opted for two choices for the $\oplus$ operator: AND and OR operators. We compare the benefits of these operators in Table 3.5.

Table 3.5: *HOS* score on three sub tasks of Office31 (%) comparing different choices of the logical operator $\oplus$ in Eq. (3.7).

Method	D → W	A → W	D → A	Avg.
Ours w/ OR sampling	91.5	80.1	77.7	83.1
Ours w/ AND sampling	93.2	83.9	78.0	85.0

Table 3.6: HOS score (%) on Office31 and VisDA-C using positive vs. negative classification loss.

Method	Office31	VisDA-C
Ours w/positive	80.5	72.1
Ours w/positive+negative	82.7	74.0
Ours w/negative	**85.4**	**75.1**

Although either operator surpasses the state-of-the-art, our model achieves the best performance by combining the uncertainty probabilities with the AND operator. The AND operator requires that samples have low uncertainty in both measures, resulting in a more fine-grained sample selection with respect to the OR operator.

Positive vs. Negative Classification Loss. Table 3.6 shows results obtained using, as classification loss, only a standard positive loss (first row), a linear combination of a positive and a negative loss (second row), and only a negative learning loss. Although a positive classification loss is more affected by the noise in the pseudo-labels introduced by both the domain shift and the presence of private samples, our method overcomes multiple baselines even using a positive loss. This highlight the effectiveness of our uncertainty estimation strategies in dealing with both noises. While using a combination of positive and negative losses results improve, the best performance is obtained using only the negative loss as it guarantees more robustness to the introduced noise.

NL-InfoNCELoss vs. InfoNCELoss. Table 3.7 shows the effectiveness of using the proposed NL-InfoNCELoss against the standard InfoNCELoss in SF-OSDA. While using the standard InfoNCELoss our method still achieves good performance, introducing the proposed NL-InfoNCELoss we obtain a gain of +2.5% and +1.1% on Office31 and VisDA-C, respectively. This is probably due to the robustness to the noise introduced by negative learning.

Table 3.7: HOS score (%) on Office31 and VisDA-C using InfoNCELoss vs. NL-InfoNCELoss.

Method	Office31	VisDA-C
Ours w/InfoNCELoss	82.9	74.0
Ours w/NL-InfoNCELoss	**85.4**	**75.1**

Table 3.8: Clustering accuracy (%) on three sub tasks of Office31 evaluating the ability of our model to perform novel class discovery. Note that accuracy is calculated only for samples belonging to target-private classes.

D → W	A → W	D → A	Avg.
71.3	54.8	66.7	64.2

Discovery of the underlying semantics of novel classes. We present here additional experiments to asses whether our model is able, as a byproduct, to aggregate samples from target-private classes into different clusters based on the semantics. To achieve this objective, the model has to learn the semantics of target-private classes, even if it cannot be directly optimised for this scope, due to the lack of labels in the target domain. We follow the same protocol as in state-of-the-art novel class discovery works [140, 141]. We train the model by setting $|C_P|$ as the actual number of private classes: this is necessary to evaluate the results of these experiments. Since we do not use labels for target-private classes during adaptation, a class misalignment is present between the predicted and the ground-truth target-private classes. To solve this problem, at inference time, we match the predicted target-private classes against the ground-truth classes. Specifically, we compute target-private class prototypes by averaging features of samples belonging to each class, according to their ground truth labels. The same process is repeated to compute the estimated prototypes using predictions rather than labels. We then run the Hungarian algorithm to perform the matching between the ground-truth and the estimated prototypes, and adjust the predicted labels accordingly. Once the predicted and ground-truth target-private classes are aligned, we quantitatively and qualitatively assess the ability of our model to correctly classify samples from target-private classes. As in previous works[140, 141], we quantitatively evaluate the ability of our method to cluster samples from target-private classes by computing the clustering

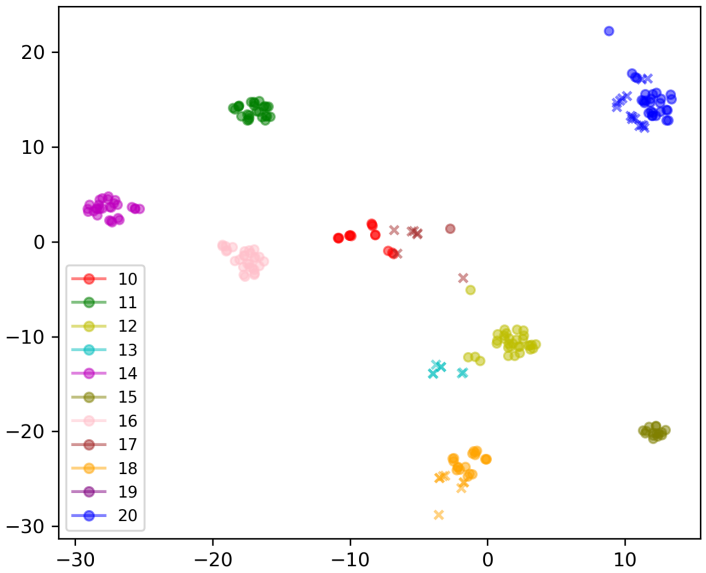

Figure 3.7: Feature space visualization for target-private classes. The colour of samples represents the predicted class. The • symbol represents correct predictions, whereas the × represents incorrect predictions. Our model produces a well-structured features space for target-private classes, where samples belonging to the same novel classes are clustered together. This opens the possibility of performing novel class discovery.

accuracy:

$$\text{ClusterAcc} = \max_{t \in \mathcal{P}(C^P)} \frac{1}{|C^P|} \sum_{j=1}^{|C^P|} \mathbb{1}(y_j = t(\hat{y}_j)) \tag{3.11}$$

where y_j and $\hat{y}_j$ are ground-truth labels and predictions, respectively. For each target-private sample, $\mathcal{P}(C^P)$ is the set of all possible permutations of C^P classes and t is a permutation.

Table 3.8 and Fig. 3.7 show that our model is able to achieve satisfying results in the classification of target-private samples, even if it has been neither optimised for this task nor trained with labels from these classes. This means the model has learned the underlying semantics of novel classes, enabling the possibility to potentially perform novel category discovery. Note that this behaviour is not possible with previous works that classify all the private samples in a single unknown class. Finally, Fig. 3.7 shows the ability of our model to produce a well-structured features space, where samples belonging to the same target-private class lie close in the space, allowing them to be clustered in the same novel class.

3.4 Summary

In this work, we introduced a novel approach to address the challenges of Source-Free Open-Set Domain Adaptation (SF-OSDA) for image classification. Unlike existing approaches that group samples from unknown classes into a single "unknown" class, our method leverages the granularity of target-private classes to segregate their samples into multiple clusters, resulting in a more effective adaptation that enables the learning of the underlying semantics of novel classes. To that end, we introduce a novel uncertainty estimation technique to handle noisy pseudo-labels produced during the adaptation. We also propose a novel *NL-InfoNCELoss* that integrates negative learning into self-supervised contrastive learning further increasing the robustness of the contrastive framework to the noise in the pseudo-labels. Our approach obtains competitive performance on small-scale datasets, such as Office31 and Office-Home, and it outperforms the most recent baselines on the large-scale benchmark dataset VisDA-C, demonstrating its effectiveness for real-world applcations. Further analyses show that our model can aggregate samples from target-private classes into different clusters as a byproduct of the adaptation, as it has learned the underlying semantics of novel classes. This behaviour also opens the possibility of performing novel class discovery.

While our method demonstrates significant advancements in the SF-OSDA setting, some aspects remain open for future exploration. For instance, the clustering-based initialisation, although effective in our experiments, could be further refined to ensure robust performance under extreme domain shifts or in highly imbalanced datasets. Additionally, while our uncertainty-guided sample selection effectively mitigates noise, future work could extend these strategies to enhance scalability and efficiency in scenarios with extensive numbers of target-private classes.

Chapter 4

Leveraging Text Robustness for Improving Unsupervised Domain Adaptation under Complex Shifts

Given the advent of novel Large Language Models (LLMs) and Vision-Language Models (VLMs), this chapter explore the possibility to integrate the language modality as additional information for the unsupervised adaptation of vision models. Indeed, language models have been demonstrated to be more robust to variations in domains and contexts than vision models. Hence, this chapter focuses on leveraging the language modality to improve the robustness of vision models when adapting across diverse geographical domains. Specifically, this chapter introduces a multimodal adaptation strategy that leverages textual information to guide a vision model into a more robust learning, in presence of complex shifts.

The findings and contributions of this chapter have been accepted to an international conference. [1]

4.1 Introduction

Although deep learning models have achieved remarkable performance in lots of computer vision tasks, they still fall short in their ability to generalise to different domains. Retraining deep learning models on new data requires a big effort to acquire and manually label images and should ideally be avoided. To address these

[1] Mattia Litrico, Mario Valerio Giuffrida, Sebastiano Battiato, Devis Tuia, "TRUST: Leveraging Text Robustness for Unsupervised Domain Adaptation" [17]. AAAI Conference on Artificial Intelligenc, 2026.

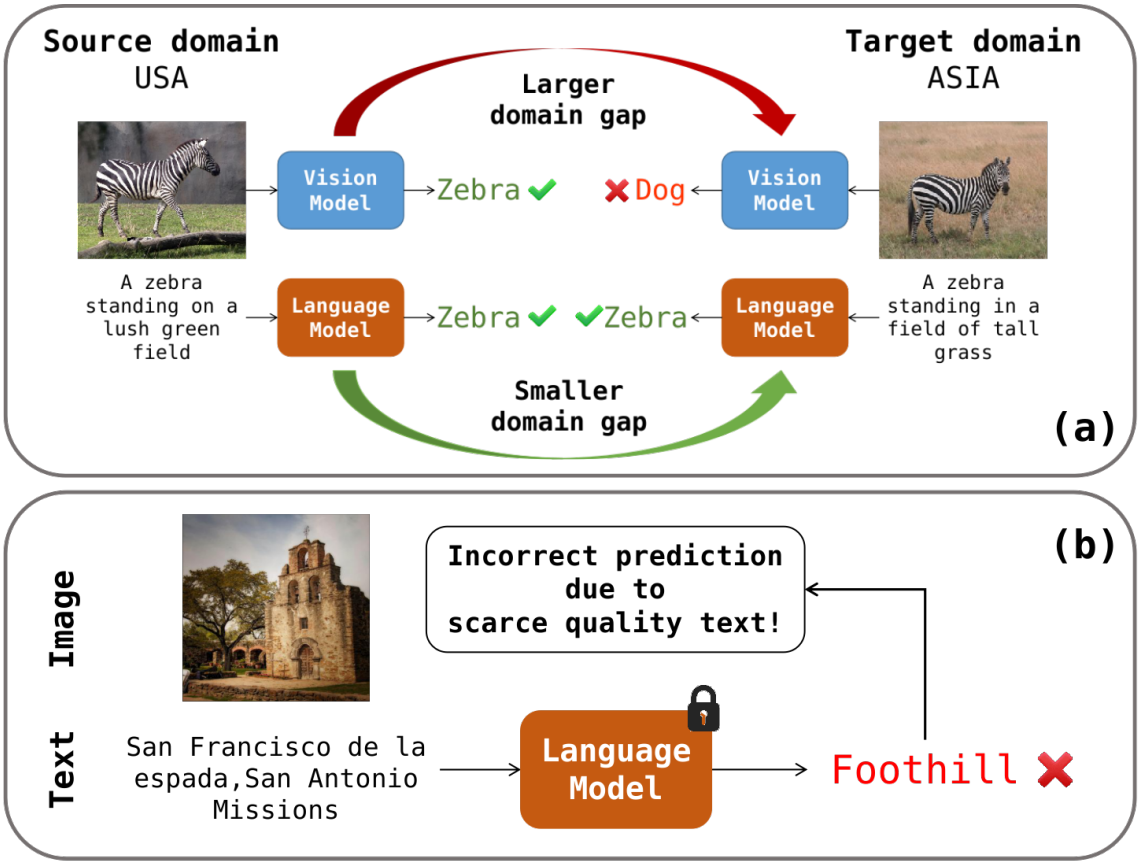

Figure 4.1: (a) geographical shifts strongly impact the appearance of both foreground objects and background, as visible in the two images from the USA and Asia. On the contrary, the captions contain valuable information about the semantic class and the text is minimally affected by the geographic shift, as stated in [87]. (b) Despite such apparent robustness, low quality captions may still lead to low classification accuracy.

limitations, Unsupervised Domain Adaptation (UDA) approaches for image classification have been proposed. UDA aims at transferring knowledge acquired on a labelled source domain to an unlabelled target domain, bridging the gap between them [21, 142, 117, 118, 25, 119, 26]. Recent UDA methodologies have been successful on classical domain shifts (*e.g.* synthetic-to-real), but they suffer a lot on more complex shifts (*e.g.* geographical), where both the background and objects' appearance change between domains [80].

Kalluri et al. [87] posited that, under complex shift, solely relying on images is challenging and they proposed LaGTran that integrates textual data for guiding the adaptation. Indeed, authors showed that textual data are more robust to complex

shifts, as they semantically describe the image content instead of focusing on appearance details, which is essential for bridging the domain gap (*c.f.* Figure 4.1(a)). However, LaGTran [87] used textual data only for generating pseudo-labels for the target domain, highly underusing the potential of the language modality to reduce the domain shift. Moreover, they blindly rely on the pseudo-labels generated from the text, which may be incorrect due to two main factors: (a) the scarce quality of text descriptions, especially for crowd-sourced texts (*e.g.* Figure 4.1 (b)); (b) the impact of domain shifts on language models, which is lower than image-based models but still present. For those reasons, solely training the vision model with pseudo-labels generated from the text is suboptimal for transferring the shift robustness from the language to the vision model.

To overcome these limitations, we propose TRUST(**T**ext **R**ob**UST**ness for unsupervised domain adaptation) a novel approach for UDA in image classification that exploits the potential of language guidance for the adaptation of a vision model. Similarly to LaGTran [87], we use a language model to generate pseudo-labels on target samples from text descriptions, but we also propose two novel components to cope with the shortcomings discussed above.

First, to reduce the impact of incorrect pseudo-labels generated from low-quality captions, we propose a novel multimodal uncertainty estimation strategy that reweights the classification loss for target samples, by evaluating how well the captions semantically describe the target images.

To this aim, we use a pretrained vision-language model (*i.e.* CLIP [88]) to evaluate the semantic correlation between target images and their captions, which serves as a measure of the uncertainty/reliability of generated pseudo-labels. The estimated uncertainty is then used to reweight the contribution of the pseudo-labels in the classification loss, reducing the negative impact of wrong pseudo-labels obtained from low-quality captions.

Second, to enhance the generalisation of the vision model, we aim to transfer the robustness to complex shifts from the language to the vision model, by promoting an effective alignment between their feature spaces. We propose to integrate a novel multimodal soft-contrastive learning loss, which uses the language modality to guide the contrastive training of the vision model. Unlike previous works, our soft-contrastive framework does not require to identify positive and negative pairs,

as it assigns to each pair a score of "*positiveness*" and "*negativeness*" based on how likely they share the same semantic content. Then, a pair of images acts simultaneously as a positive and negative pair and attracts/repulses samples with a strength proportional to the similarity of their captions. The benefits of this strategy are multiple: (a) we encourage the vision model to match the language model's feature space by attracting representations of images with similar captions and repulsing those with dissimilar ones; (b) differently than [54, 29, 30], it avoids contrasting images of the same class without relying on the output of the vision model, therefore limiting the confirmation bias; (c) we achieve a smoother contrastive training, where representations of each pair of samples are both attracted and repulsed based on how likely they share the same semantic content.

We benchmark TRUST on two datasets representing classical (DomainNet) and complex (GeoNet) shifts. In all cases, TRUST outperforms the current state-of-the-art methods: on GeoNet, we obtain the best performance with 61.16% accuracy on average, with a gain of +0.92% to the best competitor; on DomainNet, we achieve the best performance of 73.09%, largely improving performance compared to several recent UDA approaches.

To summarise, our main contributions are:

- We introduce a novel uncertainty estimation strategy that leverages CLIP to evaluate the uncertainty of pseudo-labels, by measuring the semantic correlation between images and their captions.
- We propose a novel multimodal soft-contrastive learning loss that uses the language modality to guide the contrastive training of the vision model. Our solution removes the problem of identifying positive and negative samples in UDA, preventing the confirmation bias.
- We validate our method on classical and complex domain shifts outperforming the state-of-the-art on both settings.

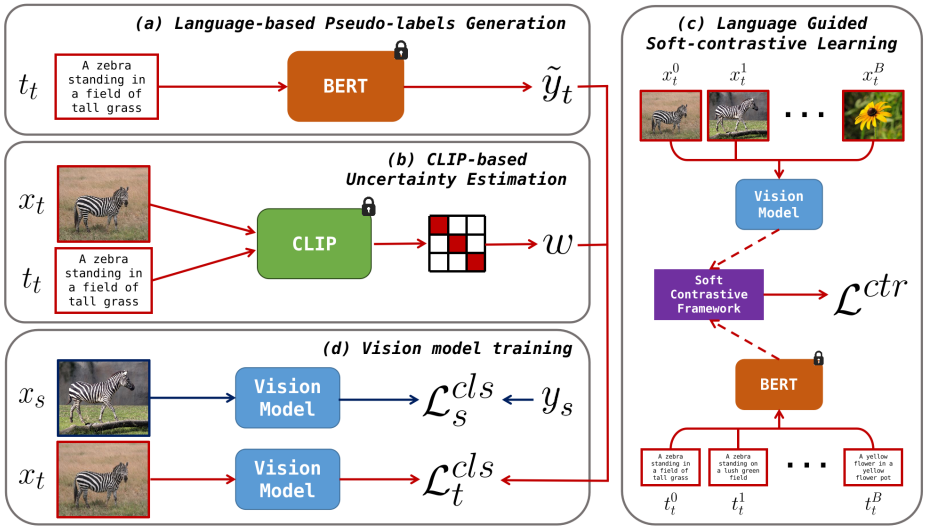

Figure 4.2: (a) We use a BERT model on target captions to generate pseudo-labels $\tilde{\mathbf{y}}_\mathbf{t}$ for target samples (Sec. 4.2.1). (b) We compute normalised CLIP similarity scores $\mathbf{w}$ from target images and texts as a measure of the reliability of the generated pseudo-labels $\tilde{\mathbf{y}}_\mathbf{t}$ (Sec. 4.2.2). (c) Feature representations of images go through a soft-contrastive framework, where they are attracted and repulsed to each other based on the similarity of their captions (dashed lines indicates features extraction). (d) A vision model is trained on both source and target images. On source images, we compute a classification loss $\mathcal{L}_s^{cls}$ using the ground truth labels $\mathbf{y}_s$. On target images we use the pseudo-labels $\tilde{\mathbf{y}}_\mathbf{t}$ as supervision and the classification loss $\mathcal{L}_t^{cls}$ is reweighted based on the estimated reliability score $\mathbf{w}$.

4.2 Proposed Method

An overview of TRUST is shown in Figure 4.2. Our aim is unsupervised domain adaptation (UDA), *i.e.* to learn a model for the target domain without having access to any ground-truth labels from it. In this context, we have access to a labelled source domain $\mathcal{D}_s : \{x_s^i, t_s^i, y_s^i\}_{i=1}^{N_s}$, where x_s and y_s are source images and ground-truth labels respectively, and t_s are source captions. Similarly, we have an unlabelled target domain $\mathcal{D}_t : \{x_t^i, t_t^i\}_{i=1}^{N_t}$. Captions are obtained from either associated metadata in web-collected images [143], or generated with image-to-text models [144] and they are used *only* at training time. TRUST trains a vision model composed of a feature extractor $f : \mathbb{R}^{H\times W\times 3} \rightarrow \mathbb{R}^P$ and a classifier $h : \mathbb{R}^P \rightarrow \mathbb{R}^C$, where P is the length of the feature vector, and C is the number of classes.

4.2.1 Language Guided Domain Adaptation

We address the problem of UDA leveraging pseudo-labels generated from captions. Similarly to [87], we fine-tune a BERT sentence classifier [145] in a supervised fashion using captions and labels from the source domain. The trained BERT model is then frozen and inputted with target captions to generate pseudo-labels for target samples. Then, source labels and the obtained target pseudo-labels are used as supervision to train a vision model on both source and target domains simultaneously, as detailed in Sec. 4.2.4.

More formally, we fine-tune a pretrained DistilBERT [146] model $\mathcal{B}$ on the source domain $\{t_s^i, y_s^i\}_{i=1}^{N_s}$ using source captions and labels. To do so, we optimise the following supervised objective:

$$\underset{\phi}{\operatorname{argmin}}\ \mathbb{E}_{\{t_s^i, y_s^i\}\sim\mathcal{D}_s}\ \mathcal{L}_{\text{CE}}(\mathcal{B}_\phi(t_s^i), y_s^i), \tag{4.1}$$

where ϕ denotes the parameters of the BERT model and $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss. Then, we use the fine-tuned model $\mathcal{B}$ to generate pseudo-labels $\tilde{\mathbf{y}}_\mathbf{t}$ on target captions $\mathbf{t}_\mathbf{t}$ by running inference passes as follows:

$$\tilde{y}_t^i = \underset{C}{\operatorname{argmax}}\ \mathcal{B}_\phi(t_t^i). \tag{4.2}$$

4.2.2 CLIP-based Pseudo-labels Uncertainty Estimation

Although previous studies [87] demonstrated that the language modality is more robust to complex domain shifts, blindly relying on the knowledge acquired from captions may lead to the generation of incorrect pseudo-labels $\tilde{\mathbf{y}}_\mathbf{t}$ due to: (a) the scarce quality of captions (*c.f.* Figure 4.1(b)), especially when they are crowd-sourced from the web; and (b) the domain shift that still exists in the language modality. Therefore, training an image classifier on such pseudo-labels (as in [87]) risks to disrupt the adaptation process.

To mitigate these issues, we propose a novel strategy to estimate the uncertainty/reliability of the pseudo-labels generated from target captions, based on the capacity of captions to semantically describe the corresponding images. Such uncertainty scores are then used to reweight the classification loss, accordingly. With

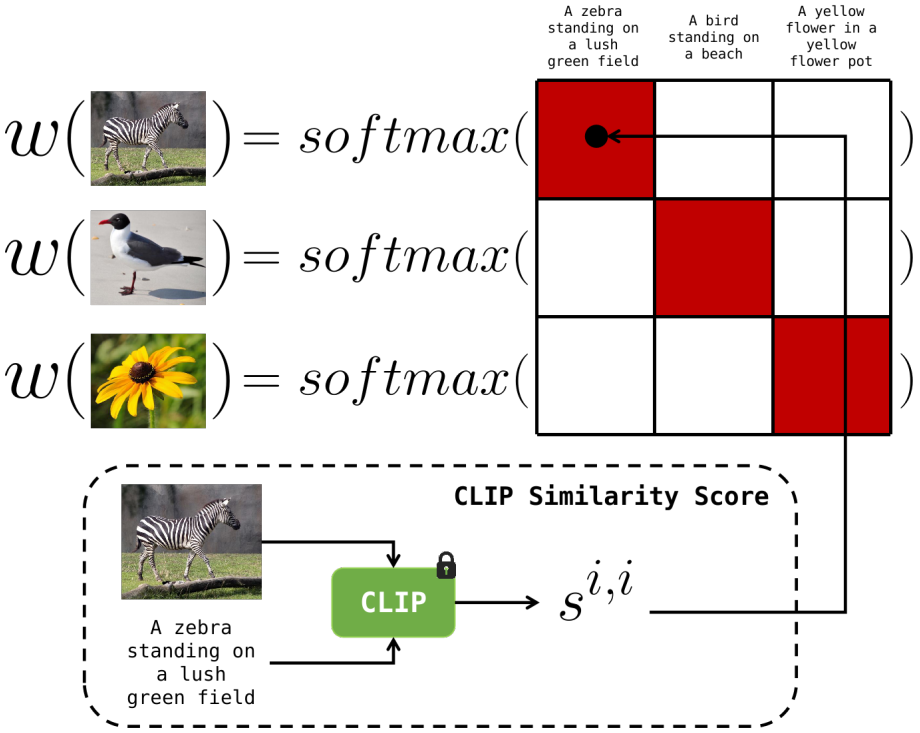

Figure 4.3: We compute CLIP similarity scores on each pair of target images and texts in a batch, obtaining the similarity matrix S. To calculate the reliability weight $\mathbf{w}$, we normalise the CLIP scores with a softmax on each row of S and select the values in the main diagonal, which measure the semantic correlation between much each image and its caption.

this aim, we use a pretrained CLIP model [88] to evaluate the correlation between images and corresponding captions. The underlying idea, shown in Figure 4.3, is that when the CLIP image/text similarity is high, the text accurately describes the image content. Consequently, we will assume that the pseudo-label generated with BERT (*c.f.* Sec. 4.2.1) is reliable. On the contrary, when the CLIP similarity is low, the text does not describe well the content of the image (e.g. Figure 4.1 (b)). This may occur when the text is of limited quality or because it does not capture the semantics of the image. In this case, the pseudo-label is considered unreliable. Differently from other uncertainty estimation strategies [30], which perform such estimation based on the output of the training model, our solution based on CLIP prevents the confirmation bias [78], since the CLIP estimation is not affected by the training of the TRUST's vision model, and CLIP's parameters are frozen during the adaptation process.

More formally, given a batch of image/caption pairs from the target domain, we compute the similarity matrix $S(E_{\text{im}}(\mathbf{x_t}), E_{\text{text}}(\mathbf{t_t}))$ as the cosine similarity between the embeddings of target images $\mathbf{x_t}$ and texts $\mathbf{t_t}$ obtained from the CLIP's image and text encoders E_{im} and E_{text}. The element $s^{i,j}$ of the similarity matrix indicates the semantic correlation between the i-th image and the j-th caption in the batch. Consequently, the diagonal elements $s^{i,i}$ indicates how much each target image is semantically related to its caption. However, this score is unbounded and needs to be normalised to measure the reliability of the pseudo-labels obtained from the captions. Therefore, we compute the softmax function for each row of the similarity matrix S to obtain a normalised score w_i, as follows:

$$w_i = \frac{\exp(s^{i,i})}{\sum_{j=1}^{B} \exp(s^{i,j})}. \tag{4.3}$$

The resulting score w_i is an estimation of the quality of the caption t_t^i to describe the image x_t^i. We use the softmax function for normalising $s^{i,i}$, to produce a score proportional to how much each image is semantically related to its caption (with respect to the other texts in the batch). The larger w_i (high reliability), the better the i-th image is described by its corresponding text, which leads to a higher confidence in the pseudo-label $\tilde{\mathbf{y}}_\mathbf{i}$. Conversely, a lower value of w_i (low reliability) means that the caption semantically describes the image as less as the other texts in the batch. Note that CLIP is pretrained and frozen during this process, to avoid adding an additional overhead for finetuning CLIP and preventing the confirmation bias.

4.2.3 Language Guided Soft-contrastive Learning

The language modality intrinsically benefits of a larger robustness to complex domain shifts compared to visual data, as demonstrated in [80]. We posit that combining the benefits of language and vision modalities improves the performance on the target domain. LaGTran [87] used a language model to generate pseudo-labels on target images for training a vision model. Despite its simplicity, this strategy alone does not encourage the vision model to inherit the robustness of the language model.

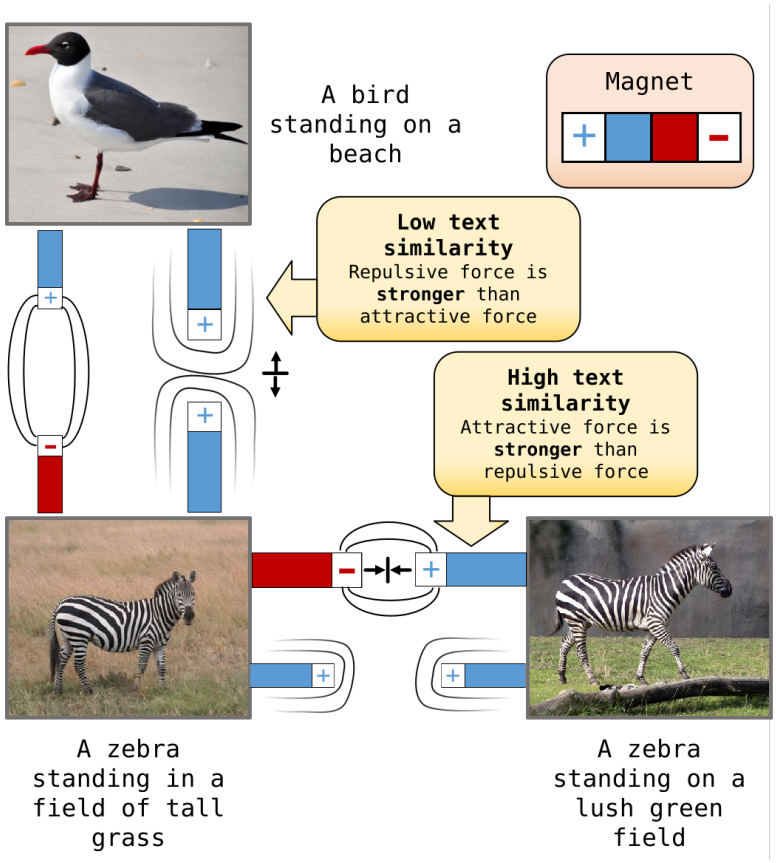

Figure 4.4: Representation of our soft-contrastive framework, where each pair of images is both attracted and repulsed based on their caption similarity: if two images have semantically similar captions, they are more attracted than repulsed and vice versa. Attraction is shown by magnets with opposite polarity (+/−) and repulsion by same-polarity magnets (+/+). The strength of each force is indicated by magnet size and field lines.

Therefore, we propose a novel language guided soft-contrastive learning framework to transfer the intrinsic robustness from the language to the vision model, by aligning their feature spaces. Differently than previous works [59, 54, 29, 30], our contrastive framework treats each pair of images as both a positive and a negative pair, with a score of "*positiveness*" and "*negativeness*" based on the semantic similarity between image captions. In this way, the feature representation of each pair will be both attracted and repulsed with a strength proportional to how likely they share the same semantic content. This strategy leads to multiple benefits. Firstly, it aligns the language and vision feature spaces, transferring the intrinsic robustness to complex domain shifts from the language to the vision model. Secondly, it does

not require to determine positive and negative pairs, since each pair plays simultaneously as a positive and a negative pair. This avoids using the vision model itself (as in [29, 30]) to discriminate between positive and negative pairs, reducing the effects of confirmation bias. Finally, our solution reduces the adverse effects of mistakenly assigning a pair as positive or negative. If two samples have incorrect pseudo-labels, the pair will not be strictly assigned to either the positive or negative class. Instead, their feature representations are both attracted and repulsed, limiting the impact of the incorrect assignment.

More formally, let $\mathcal{S} = \{a_w(x_t^i), a_s(x_t^i)\}_{i=1}^B$ be a batch composed of target samples, where we apply a weak $a_w \in \mathcal{A}_w$ and a strong augmentation $a_s \in \mathcal{A}_s$ drawn from distributions $\mathcal{A}_w$ and $\mathcal{A}_s$. Standard self-supervised contrastive methods [54, 59] optimise the following:

$$\mathcal{L}_{self}^{ctr} = -\sum_{i}^{B} \log \frac{exp(z_i \cdot \bar{z}_i/\tau)}{\sum_{j\in\mathcal{N}} exp(z_i \cdot z_j/\tau)}, \tag{4.4}$$

where $z_i = f(a_w(x_t^i))$ and $\bar{z}_i = f(a_s(x_t^i))$ are feature representations extracted from the weakly and strongly augmented i-th target sample, τ is a temperature parameter, and $\mathcal{N} : \{j | 1 \leq j \leq B, j \neq i\}$ is the set of indices of all the negative samples. In this formulation, only the augmented version of the same sample is treated as positive, while all the other samples are treated as negatives.

Inspired by [147], we generalise this formulation to account for multiple positive samples. Each pair of images in the batch is treated as positive, with a score of positiveness depending on how similar their captions are, leading to the following objective:

$$\mathcal{L}^{ctr} = -\sum_{i}^{B} \log \left\{ \frac{1}{|\mathcal{S}|} \sum_{p=1}^{|\mathcal{S}|} \frac{\text{sim}_{(i,p)} \cdot \exp(z_i \cdot \bar{z}_p/\tau)}{\sum_{j\in\mathcal{N}} (1 - \text{sim}_{(i,j)}) \cdot \exp(z_i \cdot z_j/\tau)} \right\}, \tag{4.5}$$

where $\text{sim}_{(a,b)} = \frac{f^{\mathcal{B}}(t_t^a) \cdot f^{\mathcal{B}}(t_t^b)}{||f^{\mathcal{B}}(t_t^a)|| \cdot ||f^{\mathcal{B}}(t_t^b)||}$ is the cosine similarity between the DistilBERT features vectors for the text descriptions of target samples x_t^a and x_t^b. Differently than Eq. (4.4), we treat each sample in the batch (e.g. $\bar{z}_p$) as positive of the i-th sample, requiring to include another summation over the cardinality of the multiviewed batch. This results in attracting the feature representations of a pair of images

(x_t^a, x_t^b) with a strength equal to $\text{sim}_{(a,b)}$, while repulsing them with a strength equal to $1 - \text{sim}_{(a,b)}$.

4.2.4 Overall Framework

To adapt the source to the target domain, we train a vision model on both the source and target domains, combining the source labels and the target pseudo-labels generated in Sec. 4.2.1. Differently than [87], we reweight the classification loss for target samples based on the reliability of the pseudo-labels estimated in Sec. 4.2.2. The higher the estimated reliability, the more it will contribute to the classification loss. Hence, we train TRUST with the following classification losses:

$$\mathcal{L}_s^{cls} = \frac{1}{|\mathcal{D}_s|} \sum_{i=1}^{|\mathcal{D}_s|} \mathcal{L}_{\text{CE}}(h(f(a_w(x_s^i))), y_s^i), \tag{4.6}$$

$$\mathcal{L}_t^{cls} = \frac{1}{|\mathcal{D}_t|} \sum_{i=1}^{|\mathcal{D}_t|} w_i \cdot \mathcal{L}_{\text{CE}}(h(f(a_w(x_t^i))), \tilde{y}_t^i) + (1 - w_i) \cdot \mathcal{L}_{\text{CE}}(h(f(a_w(x_t^i))), p_t^i), \tag{4.7}$$

where $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss, w_i is the reliability weight calculated from CLIP (*c.f.* Section 4.2.2), $p_t^i = h(f(a_s(x_t^i)))$ are predictions obtained by the vision model on strongly augmented samples, and a_w and a_s are weak and strong augmentations, respectively.

The overall loss function is the following:

$$\mathcal{L} = \mathcal{L}_s^{cls} + \mathcal{L}_t^{cls} + \mathcal{L}^{ctr}. \tag{4.8}$$

4.3 Experiments

Datasets. For the evaluation of TRUST, we use the GeoNet [80] and DomainNet [22] datasets. GeoNet is the largest dataset for geographical domain adaptation, with over 750,000 images, and has been recently introduced to study the effect of geographical shifts in images. GeoNet is composed of two subdatasets called GeoImnet, for image classification across 600 classes, and GeoPlaces, for scene recognition across 205 classes. DomainNet-345 is a dataset commonly used for UDA, consisting of 400,000 images across 345 classes. As previous works [80, 87], we present results

on 12 transfer scenarios involving 4 domains: Real (R), Clipart (C), Sketch (S), and Painting (P). For GeoNet, we use the metadata of the dataset and concatenate the tags, alt-text, and free-form captions provided for each image to create the captions. For DomainNet, we use the captions provided by [87], which have been generated with BLIP-2 [144].

Implementation details. Following [87], we use a ViT-base [2] and a Swin-base [3] backbone as the image encoder of the vision model on GeoNet and DomainNet, respectively. Both backbones are pre-trained on ImageNet-1k [133], and we add a 2-layer MLP as the classifier head. The vision model is trained using the SGD optimiser, with a learning rate of $3 \cdot 10^{-4}$. For the language model, we use a pre-trained Distill-BERT [146] model from HuggingFace and we fine-tune it on the source domain data using the AdamW optimiser with a learning rate of $5 \cdot 10^{-5}$. Both training employ a cosine decay scheduler. We use the same weak and strong augmentations as in [30]. The code will be available upon acceptance.

Baselines. We employ two text baselines (TextMatch and nGramMatch) as in [87] to evaluate the effect of using the source data for fine-tuning the language model, instead of using directly the target captions. We also compare with CLIP employing two baselines: CLIP zero-shot inference [82], performed by incorporating the domain information into the text prompt of CLIP (*e.g.*, *A clipart of a* `<class>`); CLIP model finetuned on the source to adapt it to the data distribution. Finally, we employ a training-free baseline by ensembling CLIP and TextMatch (CLIP+TextMatch) by averaging the CLIP's image-text similarity and the TextMatch's caption-label similarity, and then computing the argmax to obtain the ensemble predictions.

4.3.1 Results

GeoNet. Table 4.1 presents results of TRUST on the geographical shifts of the GeoNet datasets. Our method outperforms the source-only baseline by +14.63%. Compared to LaGTran [87], we improve performance by +0.92% and we achieve better performance on all the shift settings of both GeoImnet and GeoPlaces. We also compare our method with the zero-shot performance of CLIP and CLIP+TextMatch. Despite CLIP being trained on a significantly larger amount of data, our approach obtains a gain in performance of +10.80% and +9.28%, respectively. While in most of the settings the CLIP+TextMatch performs better than CLIP and TextMatch

	GeoImnet		Avg.	GeoPlaces		Avg.	Total Avg.
	U → A	A → U		U → A	A → U		
Baselines							
TextMatch [87]	49.68	54.82	52.25	53.06	50.11	51.58	51.92
nGramMatch [87]	49.53	51.02	50.27	51.70	49.87	50.78	50.93
CLIP+TextMatch	50.11	54.92	52.51	50.36	52.14	51.25	51.88
Zero-shot classification							
CLIP [88]	49.84	53.83	51.83	43.41	54.34	48.87	50.36
CLIP* [88]	57.79	59.12	58.45	48.91	55.89	52.40	55.42
Unsupervised Domain Adaptation							
Source only	52.46	51.91	52.18	44.90	36.85	40.87	46.53
CDAN [142]	54.48	53.87	54.17	42.88	36.21	39.54	46.86
MemSAC [148]	53.02	54.37	53.69	42.05	38.33	40.19	46.94
ToAlign [25]	55.67	55.92	55.79	42.32	38.40	40.36	48.08
MDD [149]	51.57	50.73	51.15	42.54	39.23	40.88	46.02
DALN [26]	55.36	55.77	55.56	41.06	40.41	40.73	48.15
PMTrans [32]	56.76	57.60	57.18	46.18	40.33	43.25	50.22
LaGTran [87]	<u>63.67</u>	<u>64.16</u>	<u>63.91</u>	<u>56.14</u>	<u>57.02</u>	<u>56.58</u>	<u>60.24</u>
TRUST	**64.81**	**65.12**	**64.96**	**57.56**	**57.17**	**57.36**	**61.16**

Table 4.1: Classification accuracy (%) under the geographical shifts of GeoNet. All methods use a ViT-base backbone. Best results are in **bold**, second best results are <u>underlined</u>. TRUST outperforms the state-of-the-art by +0.96% on GeoImnet (Avg.) and by +0.78% on GeoPlaces (Avg.). The symbol * indicates models finetuned on the source data.

used alone, results are still lower than TRUST . We hypothesise that CLIP and TextMatch make coherent mistakes, leading to a low improvement when ensembling the two models. Finally, compared to the CLIP model finetuned on source data (CLIP*), TRUST improves accuracy by +5.74%.

DomainNet-345. Table 4.2 presents results on DomainNet-345 that, exhibiting classical shifts, leads to generally higher scores than in GeoNet. For instance, CLIP and CLIP* achieve a notable accuracy of 70.38% and 71.26%. As for GeoNet, TRUST outperforms standard UDA approaches, CLIP and CLIP* by +3.46%, 2.71% and +1.83%, respectively. On non-real to real transfers, TRUST performs worse than CLIP and CLIP+TextMatch. A possible explanation is that CLIP is trained on a massive amount of real-world data, potentially eliminating any domain shifts between the train and test settings. Differently, TRUST is trained on non-real images (the source domain), which instead exhibit domain shifts with testing data. Nonetheless, TRUST achieves overall superior performance across diverse target domains, demonstrating its adaptability to different domain shifts. Compared to LaGTran, which also uses text during training, TRUST improves accuracy on

	Real →			Clipart →			Sketch →			Painting →			Average
	C	S	P	R	S	P	R	C	P	R	C	S	
Baselines													
TextMatch [87]	71.36	64.30	65.32	81.25	65.65	64.85	81.09	72.65	63.94	81.08	70.84	**64.17**	70.14
nGramMatch [87]	68.92	59.82	63.15	76.35	61.72	62.87	76.35	69.28	62.51	76.04	68.52	60.52	67.17
CLIP+TextMatch [87]	72.89	63.56	66.97	81.53	62.98	66.74	**81.89**	72.81	65.97	**81.87**	72.07	63.34	71.05
Zero-shot Classification													
CLIP [88]	72.39	60.90	66.81	81.37	60.90	66.81	81.37	72.39	66.81	81.37	72.39	60.90	70.38
CLIP* [88]	73.45	62.81	67.07	**81.71**	62.81	67.07	81.71	73.45	67.07	81.71	73.45	62.81	71.26
Unsupervised Domain Adaptation													
Source only	63.02	49.47	60.48	70.52	56.09	52.53	70.42	65.91	54.47	73.34	60.09	48.25	60.38
MCD [117]	39.40	25.20	41.20	44.60	31.20	25.50	34.50	37.30	27.20	48.10	31.10	22.80	34.01
MDD [149]	52.80	41.20	47.80	52.50	42.10	40.70	54.20	54.30	43.10	51.20	43.70	41.70	47.11
CGDM [150]	49.40	38.20	47.20	53.50	36.90	35.30	55.60	50.10	43.70	59.40	37.70	33.50	45.04
SCDA [28]	54.00	42.50	51.90	55.00	44.10	39.30	53.20	55.60	44.70	56.20	44.10	42.00	48.55
SSRT-B [151]	69.90	58.90	66.00	75.80	59.80	60.20	73.20	70.60	62.20	71.40	61.70	55.20	65.41
MemSAC [148]	63.49	42.14	60.32	72.33	54.92	46.14	73.46	68.04	52.75	74.42	57.79	43.57	59.11
CDTrans [33]	66.20	52.90	61.50	72.60	58.10	57.20	72.50	69.00	59.00	72.10	62.90	53.90	63.16
PMTrans [32]	74.10	61.10	**70.00**	79.30	63.70	62.70	77.50	73.80	62.60	79.80	69.70	61.20	69.63
LaGTran [87]	77.30	68.25	67.35	81.31	67.03	66.81	80.78	75.62	68.08	79.23	73.80	63.44	72.41
TRUST	**78.65**	**69.99**	68.69	80.19	**67.07**	**67.50**	79.98	**77.24**	**68.51**	80.62	**75.24**	63.49	**73.09**

Table 4.2: Classification accuracy (%) under the classical shifts of DomainNet-345. All methods use the Swin-base backbone. TRUST achieves the highest accuracy on 7 out of the 12 domain shifts and the highest accuracy on average. The symbol * indicates models finetuned on the source data.

Hard Contrastive	Soft Contrastive	CLIP Uncertainty	Avg. Acc.
✗	✗	✗	61.04
✓	✗	✗	61.50
✗	✓	✗	64.01
✗	✗	✓	62.11
✗	✓	✓	**64.96**

Table 4.3: Ablation studies of components of the proposed method measured by classification accuracy (%) on GeoImnet.

average by +0.68%, demonstrating the effectiveness of the introduced components. Moreover, TRUST performs best in almost all the transfer scenarios, achieving the best or second best results on 9 out of 12 scenarios.

4.3.2 Analysis

Ablation Studies. In Table 4.3, we report ablation studies for the TRUST components on GeoNet. When using only the pseudo-labels, TRUST achieves the lowest accuracy of 61.04%. When adding the standard contrastive loss [54] (second row)

Method	Avg. Acc.
TRUST w/ CLIP	59.87
TRUST w/ CLIP*	61.95
TRUST w/ CLIP-Text	62.72
TRUST w/ BERT*	**64.96**
TRUST w/ batch-max norm.	63.59
TRUST w/ bank-max norm.	63.87
TRUST w/ class-max norm.	63.75
TRUST w/ softmax (Eq. (4.3))	**64.96**

Table 4.4: Classification accuracy (%) on GeoImnet comparing CLIP and BERT for generating pseudo-labels (Sec. 4.2.1) and different strategies of normalising CLIP scores in Eq. (4.3). The symbol * indicates models finetuned on the source data.

that requires a hard selection of positive and negative pairs, the improvement in performance remains negligible. But when using our proposed language guided soft-contrastive loss (Sec. 4.2.3), we boost the performance by +2.51% (third row), showing the effectiveness of our proposed solution. The fourth row presents the results enabling only our CLIP-based uncertainty estimation (Sec. 4.2.2), which brings a gain in performance of +1.07%. Finally, in the last row, we show the gain obtained by the full TRUST model which, enabling both the two introduced components, further improves performance by +3.93%.

Generating pseudo-labels with CLIP. Table 4.4 (top part) presents results using CLIP [82] for generating pseudo-labels in Sec. 4.2.1, as an alternative to finetuning BERT. We compare three different CLIP-based approaches: the CLIP zero-shot inference (TRUST w/ CLIP), the CLIP model finetuned on source data (TRUST w/ CLIP*) and the generation of pseudo-labels by evaluating the similarity between captions and class names in the CLIP's text feature space (TRUST w/ CLIP-Text). Results show that using CLIP (instead of BERT) for obtaining the pseudo-labels leads to lower performance. Moreover, text-only based approaches, like CLIP-Text and BERT, achieve better results, justifying the assumption that language models benefit of a larger robustness to domain shift than vision models.

Normalisation of the CLIP similarity score. Table 4.4 (from second to last row) shows results using different choices for the normalisation of the CLIP similarity score $s^{i,i}$ (Sec. 4.2.2). Specifically, we compare three different solutions, where

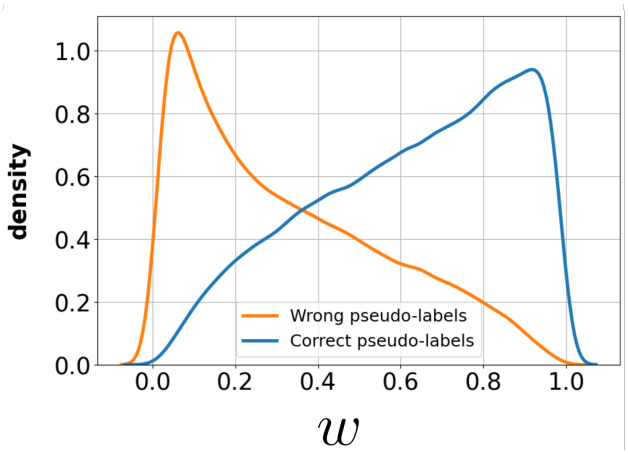

Figure 4.5: Probability density function of reliability weights estimated in Sec. 4.2.2 for target samples having correct and wrong pseudo-labels. We observe that our strategy is able to assign lower weight to wrong pseudo-labels, reducing their impact on the classification loss.

$s^{i,i}$ is divided by the maximum score in the batch (batch-max), the maximum score in the training set (bank-max), and the maximum score for the corresponding class in the training set (class-max). Despite these strategies overcome most of the baselines, using the softmax function (Eq. (4.3)) we achieve the best performance, with a gain of +1.10% over the second best normalisation strategy.

Effectiveness of CLIP-based uncertainty estimation. Figure 4.5 illustrates the effectiveness of the proposed CLIP-based uncertainty estimation strategy, showing the distribution of the reliability weights $\mathbf{w}$ (Sec. 4.2.2) for target samples with correct and wrong pseudo-labels $_{\mathbf{t}}$. Figure 4.5 shows that the two distributions are clearly separable and samples with correct pseudo-labels have high values of $\mathbf{w}$ (high reliability), while samples with wrong pseudo-labels have lower values of $\mathbf{w}$ (low reliability). These results validate our strategy to use CLIP to identify low-quality text descriptions, which is able to down-weight wrong pseudo-labels and to reduce their contribution in the classification loss.

Effects of the soft-contrastive loss. In Figure 4.6, we qualitatively analyse the effectiveness of our soft-contrastive loss in aggregating the representations of semantically similar samples that belong to the same class. Specifically, we show the top-3 nearest neighbour retrievals using features computed by the vision model, when trained with the standard hard contrastive loss [59, 54] (Hard contrastive)

Figure 4.6: Visualisation of nearest neighbors for two target images (grey borders) comparing the standard contrastive loss [59, 54] (top two rows), and our soft-contrastive loss (*c.f.* Sec. 4.2.3) (bottom two rows). Images with green borders represent correctly retrieved images; images with red borders represent retrieval mistakes. Our approach exhibits better "same-class" retrievals even in the presence of semantically similar classes.

or with the proposed soft-contrastive loss. Fig. 4.6 shows that our loss produces a more fine-grained retrieval even in presence of semantically similar classes (e.g. *streetcar*, *cable_car*, *shuttle_bus*). On the contrary, the hard contrastive loss leads to less accurate retrievals, likely because it results in a coarser-grained feature space, where visually similar classes are aggregated even though they represent different semantic concepts (e.g. *farmers_market* and *plaza*).

4.4 Summary

In this work, we introduced TRUST, a novel approach to UDA for image classification that leverages textual data to guide the adaptation to the target domain. TRUST generates pseudo-labels for target samples from captions and estimates their uncertainty using CLIP to mitigate the impact of wrong pseudo-labels in the classification loss. We also proposed a novel soft-contrastive learning framework that aligns vision and language feature space, to transfer the shift robustness from the

language to the vision model. Our extensive evaluations on DomainNet and GeoNet benchmarks demonstrated that TRUST outperforms the current state-of-the-art in both classical and complex domain shifts.

Regarding limitations and possible negative impact, TRUST introduces the limited overhead to require text data in both domains. Additionally, using text data in the training may introduce some bias in the model. Similar to other UDA methods, TRUST allows to deploy AI across domains, possibly leading to harmful applications in case of misuse.

Chapters 3 and 4 presented two approaches based on domain adaptation as a solution to reduce the burden of collecting annotated data. The next chapter will focus on a different paradigm, namely the weak-supervised learning, where the annotation cost is reduced by using coarse-grained labels that are easier to be obtained.

Chapter 5

Crowd Density Estimation without Location-level Annotations

Beyond domain adaptation, another way to reduce the burden of collecting fine-grained annotations is weak supervision, which relies on weaker forms of annotations that are easier and more economical to obtain. Hence, this chapter focuses on the task of crowd density estimation, which involves predicting spatial density maps that indicate how many objects are present in different regions of an image. This task requires fine-grained, location-level annotations, where each object is individually marked, thus making the annotation process very costly. To overcome this issue, this chapter proposes a weakly supervised pipeline for crowd density estimation that avoids the need for location-level annotations. By leveraging only count-level labels, the framework generates spatially meaningful density maps, significantly lowering annotation costs.

The findings and contributions of this chapter have been submitted to an international journal. [1]

5.1 Introduction

Crowd density estimation aims to predict a spatial map representing the density of people in images [152]. Despite considerable strides in this area, most current

[1] Mattia Litrico, Feng Chen, Michael Pound, Sotirios A Tsaftaris, Sebastiano Battiato, Mario Valerio Giuffrida, "Count2Density: Crowd Density Estimation without Location-level Annotations" [18]. Under Review, Pattern Recognition, 2025.

Approach	Annotations	Annotations Cost	Predictions	Subregion counting
Fully supervised [154, 155]	Location-level	High	Density maps	✓
Note: Entire training set is annotated.				
Semi-supervised [105, 106]	Location-level	Medium	Density maps	✓
Note: A portion of the training set is annotated.				
Cross-domain [109, 156]	Location-level	Medium	Density maps	✓
Note: Source domain is fully annotated, target domain is unlabelled.				
Regression-based [157, 110]	Count-level	Low	Crowd counts	✗
Note: They do not directly predict density maps, which are extracted from feature maps lacking quantitative density information.				
COUNT2DENSITY	Count-level	Low	Density maps	✓
Note: Count-level annotations are used to generate pseudo-density maps.				

Table 5.1: Differences with previous approaches. COUNT2DENSITY is trained to predict density maps, by generating pseudo-density maps from count-level annotations. This way of training reduces the effort of collecting location-level annotation while enabling subregion counting.

approaches require fine-grained location-level annotations to train a model, such as bounding boxes or points [104]. The need for large and fine-grained annotated datasets has become more pressing, as deep learning approaches for density estimation have gained popularity [98]. However, collecting such annotations is tedious, error-prone, and time-consuming, often requiring domain-specific expertise, especially in scenarios where objects are difficult to detect [153]. This significant annotation cost and effort act as a major bottleneck for the widespread deployment of crowd density estimation in practical applications and for the development of large, diverse datasets required to train state-of-the-art deep learning models.

To reduce the burden of collecting fine-grained location-level annotations, several approaches have been proposed, as summarised in Table 5.1. Semi-supervised methods predict density maps using datasets partially annotated with location-level data [103], reducing the number of annotated images but still requiring some location-level annotations. Cross-domain adaptation trains deep networks on fully annotated source datasets and adapts them to unlabelled target datasets, requiring location-level annotations for the source domain [156]. Unsupervised methods avoid this requirement but perform significantly worse than semi-supervised and cross-domain approaches [102]. While these methods offer advancements, they still fall short of fully eliminating the need for expensive location-level annotations or compromise on performance for practical applications. Another approach is regression-based methods, whereby the model is trained to predict the total number of people in images

[110]. Although this approach helps to reduce the burden of collecting fine-grained annotations, its predictions do not provide spatial information, preventing subregion counting as these models do not generate any density maps. Therefore, there remains a critical need for methods that can significantly reduce annotation costs by relying solely on readily available count-level data, while simultaneously retaining the ability to generate meaningful density maps that provide quantitative spatial information and enable fine-grained tasks such as subregion counting.

A recent paper stated, "*Developing techniques to automatically generate and refine labels using unsupervised or semi-supervised learning approaches could contribute to model generalization and improve accuracy in real-world scenarios where annotated data may be scarce or unreliable.*" [153]. COUNT2DENSITY aims to address this gap in the field.

To overcome the need for location-level annotations while still predicting spatial information for crowd density estimation, we propose COUNT2DENSITY, a novel pipeline designed to directly predict meaningful density maps leveraging only count-level annotations during training, which are easier to collect [104]. To retrieve quantitative spatial information from counting values, we generate pseudo-density maps from past predictions stored in a historical map bank, updated with an exponential moving average of predicted density maps. At each iteration, an averaged density map is fetched from the bank and converted into an attention map. Then, we leverage the count-level annotation to sample as many locations as the total number of people in the image, using the attention map values as a probability prior. The generated pseudo density-map is then used to train the model in a *self-supervised* fashion. To further improve the spatial awareness of the model, we incorporate a self-supervised contrastive spatial regulariser that encourages similar features within crowd regions, while promoting dissimilarity with background regions. This unique integration of a Historical Map Bank and a contrastive spatial regulariser allows COUNT2DENSITY to infer detailed spatial density information from count-level supervision, a challenge previously under-explored in the literature.

We benchmark our method on several major crowd counting datasets [98, 158, 159] outperforming the state-of-the-art by a large margin. When compared to cross-domain approaches, our method reduces the Mean Absolute Error (MAE) by 49.1 (from 198.3 to 149.2) on UCF-QNRF, and by 7.8 (from 99.3 to 91.5) on

ShanghaiTech-a. We also present results obtained by training our method in a semi-supervised manner, further outperforming related state-of-the-art methods. Qualitative results demonstrate that COUNT2DENSITY generates meaningful density maps from count-level annotations, enabling subregion counting. Additional analyses validate the effectiveness of each individual component in our pipeline.

The main contributions of our work are as follows:

- To reduce the cost of collecting fine-grained location-level annotations, we introduce COUNT2DENSITY, a novel pipeline that predicts meaningful density maps using only count-level annotations, overcoming the fundamental limitation of regression-based methods that lose spatial information.
- To capture spatial information, COUNT2DENSITY leverages count-level annotations to generate pseudo-density maps by sampling points from past predictions. Additionally, we add a self-supervised contrastive spatial regulariser to enhance the model's spatial awareness, guiding representations in both crowd and background areas. These two methodological contributions are central to the ability of our model to infer rich spatial information from count-level annotations.
- We evaluate COUNT2DENSITY on several benchmark datasets, achieving significant improvements over state-of-the-art cross-domain and semi-supervised approaches. Our analyses demonstrate the ability of COUNT2DENSITY to effectively retrieve spatial information from count-level annotations and establish a new paradigm for density estimation without the need for burdensome and costly annotations.

5.2 Proposed Method

Fig. 5.1 offers an overview of our approach: an input image x_i is provided into the model to predict a density map. The predicted density map $\hat{M}_i$ is used in the loss function, minimising the difference with a generated pseudo-density map $\tilde{M}_i$. The pseudo-density map is obtained by sampling as many locations as in the ground-truth count-level annotation, leveraging past predictions stored in a *Historical Map Bank* as probability prior. To mitigate the confirmation bias, the historical map

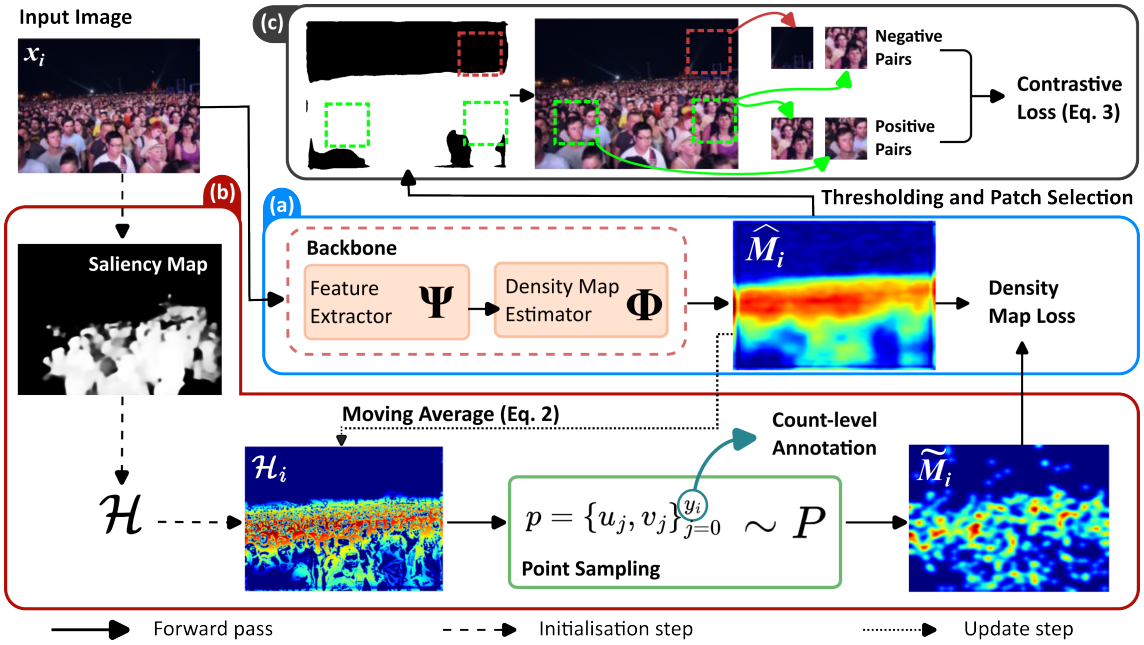

Figure 5.1: Overview of COUNT2DENSITY. (a) The input image x_i is provided to the backbone to predict a density map $\hat{M}_i$. (b) x_i is provided to an unsupervised saliency estimator to initialise the Historical Map Bank $\mathcal{H}$ (Section 5.2.2). The historical map bank is updated at each epoch, using an Exponential Moving Average (Equation (5.2)) of the predicted density map $\hat{M}_i$. For each image x_i, the information in $\mathcal{H}$ is retrieved and $\mathcal{H}_i$ is used to sample y_i (count-level annotation) points to generate a pseudo-density map $\tilde{M}_i$ that is used as source of self-supervision to update the weights of the feature extractor Ψ and the density map estimator Φ. (c) Using the threshold image obtained from the predicted density map $\tilde{M}_i$, we distinguish between crowded and background regions. Positive pairs are then formed by pairing crowded areas. Note that the patch selection is performed on the feature maps extracted by Ψ. (Best viewed in colour.)

bank is updated iteratively using an exponential moving average. Lastly, to improve the spatial awareness of the model, we adopt a self-supervised contrastive loss to encourage distinguishable features between crowd and background regions.

5.2.1 Problem Statement

Let $\mathcal{D} = \{x_i, y_i\}_{i=0}^{N}$ be a dataset containing N images $x_i \in \mathbb{R}^{3\times H\times W}$ and count-level labels $y_i \in \mathbb{N}$. Given a feature extractor $\Psi(x_i)$, which extracts a feature vector $z_i \in \mathbb{R}^D$ from an input image x_i, and a decoder Φ, we train the network to predict a density map $\hat{M}_i \in \mathbb{R}^{\hat{H}\times\hat{W}}$ obtained as $\hat{M}_i = \Phi(z_i)$. The predicted total count $\hat{y}$ is

calculated by integration from $\hat{M}_i$ as described in [152]:[2]

$$\hat{y}_i = \sum_{u=1}^{W} \sum_{v=1}^{H} \hat{M}_i(u, v). \tag{5.1}$$

5.2.2 Generating Pseudo-Density Maps

Count-level annotations lack quantitative spatial information. The purpose of COUNT2DENSITY is to retrieve spatial information from only count-level annotations, as count values are *the only source of supervision* used during training. Inspired by the paradigm of training from noisy labels, COUNT2DENSITY generates pseudo-density maps from count values by sampling locations that are likely to belong to crowd-dense regions. Any sampling errors are treated as noise within the pseudo-labels. The generated pseudo-density map retrieves quantitative spatial information from count-level annotations, and these are used as self-supervision for training the model.

Fig. 5.1(b) shows how the pseudo-density maps are generated during training. We build a *historical map bank* $\mathcal{H} \in \mathbb{R}^{N \times H \times W}$ that stores past predictions $\hat{M}_i$ for each input image x_i in the training set. Specifically, each entry $\mathcal{H}_i \in \mathbb{R}^{H \times W}$ in the historical map bank is updated at each epoch t by calculating an exponential moving average of the predicted density map $\hat{M}_i$ as follows:

$$\mathcal{H}_i^{(t)} = \alpha \hat{M}_i^{(t)} + (1 - \alpha)\mathcal{H}_i^{(t-1)}, \tag{5.2}$$

where $0 \leq \alpha \leq 1$ is a hyperparameter controlling the update of the historical map bank, and the subscript i indicates that the i-th map is obtained by averaging past predictions from the i-th training image. In this way, the averaged density maps in the historical map bank act as an ensemble of networks, alleviating confirmation bias [78]. Note that for each image in the training set, we store a corresponding map in the bank. This means that the output from the i-th image, contribute in updating only the i-th map in the bank.

The historical map bank calculated in Equation (5.2) is computed for each image in the training set and used as a probability prior to generate the pseudo-density maps at each training iteration. Specifically, we take the i-th map in the bank $\mathcal{H}_i^{(t)}$

[2]Note that Eq. (5.1) can be adopted for subregion counting.

at epoch t and we normalise[3] it to make its values represent the probability of a pixel being in a crowd region. We denote the normalised map as $\widehat{\mathcal{H}_i^{(t)}}$. Then, to generate the pseudo-density maps, we sample $p = \{u_j, v_j\}_{j=0}^{y_i}$ number of locations equal to the count label y_i, where u_j and v_j indicate the horizontal and vertical coordinates of the sampled point, respectively. The sampling is performed using a hypergeometric distribution $\mathcal{M}(y_i, \widehat{\mathcal{H}_i^{(t)}})$, where y_i is the number of sampled locations and $\widehat{\mathcal{H}_i^{(t)}}$ is the probability of each pixel being sampled. This process is equivalent to performing y_i Bernoulli samplings without replacement.

At the beginning of the training, the historical map bank contains arbitrary values, diminishing the spatial information in the generated pseudo-density maps. To address this issue, we opt to initialise the historical map bank using an unsupervised saliency estimator. This estimator predicts a binary map, where non-zero pixels indicate salient regions within an image. This initialisation provides a noisy prior for generating subsequent pseudo-density maps. An example of the saliency estimation is depicted in Fig. 5.1. Unsupervised saliency estimation methods [160] have been shown to perform on par with their supervised counterparts [161] across several scenarios. The current literature is rich in saliency estimation approaches, using hand-crafted priors or relying on videos to learn a salient object detector. Here, we use BAS-NET [162], as it is fully *unsupervised* and exhibits good performance in different scenarios. To initialise the Historical Map Bank, we run BAS-NET on each of the training images and we use the prediction to initialise the corresponding map in the bank. More formally, we initialise the Historical Map Bank as follow:

$$\mathcal{H}_i^0 = \text{BASNET}(x_i),$$

where x_i is a training image and $\mathcal{H}_i^0$ is the corresponding map in the Historical Map Bank. The superscript 0 means that this initialisation is performed before the training ($t = 0$). In Section 5.3.2, we present an analysis of the performance achieved by comparing various saliency estimators. We also show the performance of COUNT2DENSITY without saliency initialisation. Our analyses will show that initialising $\mathcal{H}$ with BAS-NET [162] improves performance, and it does not require any extra source of supervision, as it was designed to operate without ground-truth.

[3]The normalisation is performed by dividing each pixel by the maximum value in the map $\mathcal{H}_i^{(t)}$. After that, we apply a Gaussian filter to smooth out the result.

5.2.3 Learning Spatially Consistent Features

Crowd images can be divided into crowded and background areas. Specifically, crowd regions are typically characterised by distinct patterns, making them significantly different from background areas (*e.g.*, sky, buildings, etc.). Based on this observation, we leverage a self-supervised contrastive feature regulariser to promote spatial consistency for the feature extractor $\Psi(\cdot)$, as shown in Fig. 5.1(c). This regulariser encourages the feature distribution within crowded regions to be similar. The distance between representations extracted from crowded areas is minimised, while the distance between features extracted from background areas is maximised. Contrastive learning [163] requires positive pairs (f, f^+) and negative pairs $\{(f, f_0^-), (f, f_1^-), ..., (f, f_K^-)\}$, which we select using the predicted density map $\hat{M}$. Specifically, we normalise $\hat{M}$ and then apply a threshold to select non-zero pixels that corrispond to crowd regions. At this point, we build positive pairs by matching two crowded areas and negative pairs by matching a crowd and a background area. We then optimise the following contrastive loss:

$$\mathcal{L}^{ctr} = -\log \frac{\exp\left(\Psi(f)^T \cdot \Psi(f^+)/\tau\right)}{\sum_{k=0}^{K} \exp\left(\Psi(f)^T \cdot \Psi(f_k^-)/\tau\right)}, \tag{5.3}$$

where the temperature τ controls the scale of the dot product, and T indicates the transposition operation.

5.2.4 Total Objective Function

We train COUNT2DENSITY by optimising the following combined objective function:

$$\mathcal{L}(x_i; \Psi, \Phi) = \mathcal{L}^{map} + \mathcal{L}^{ctr}, \tag{5.4}$$

where $\mathcal{L}^{map}$ and $\mathcal{L}^{ctr}$ are equally weighted.

Our approach is agnostic to the choices of Ψ and Φ (the backbone), and we adapted several state-of-the-art architectures to demonstrate its effectiveness. As different backbones use different loss functions for the $\mathcal{L}^{map}$ term in Eq. (5.4), we modify $\mathcal{L}^{map}$ accordingly.

5.3 Experiments

Implementation Details. We used PyTorch as the deep learning framework. Since COUNT2DENSITY is agnostic to the backbone (*c.f.* Section 5.2.4), we integrated our training framework into the following state-of-the-art architectures: (i) [164], which models noisy annotations for crowd counting (referred to as NCC); (ii) [154], which adopts a Bayesian loss (BL); (iii) [165], which employs a generalised loss (GL); and lastly, (iv) the Multifaceted Attention Network [155] (MAN). To integrate our framework with these methodologies, we use their same backbones and losses ($\mathcal{L}^{map}$) Moreover, we also evaluate our framework using a standard U-Net architecture and a Mean-Squared Error (MSE) as $\mathcal{L}^{map}$. For fair comparisons, hyperparameters were set as in their respective implementations. We set $\alpha = 0.7$ in Eq. (5.2), and $\tau = 0.07$ in Eq. (5.3). The value of α was chosen following a grid search, while the value of τ was taken from [30].

Datasets. We adopted the following datasets: SHANGHAITECH [98], UCF-QNRF [158], JHU-CROWD++ [159], NWPU-CROWD [166]. SHANGHAITECH consists of two subsets: Part A (482/300 images for training/testing) and Part B (716/400). UCF-QNRF is a large-scale dataset with 1,535 high-resolution images (1,201/334 for training/testing). JHU-CROWD++ contains 4,317 images (2,722/500/1,600 for training/validation/testing). NWPU-CROWD [166] is a massive crowd counting dataset containing highly congested crowds and appearance variations.

Evaluation Metrics. As in [154, 155, 164, 165], we use the *Mean Absolute Error* (MAE) and the *Root Mean Squared Error* (MSE) as evaluation metrics:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \hat{y}_i\right|, \text{MSE} = \left(\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2\right)^{\frac{1}{2}}, \tag{5.5}$$

where N is the number of images, while y_i and $\hat{y}_i$ are the count-level annotations and predicted counts (*c.f.* Eq. (5.1)), respectively. For a qualitative evaluation of the predicted density maps, we use the *Structural Similarity Index Measure* (SSIM) and *Peak Signal-to-Noise Ratio* (PSNR).

	Annotation Type			UCF-QNRF[158]		ShT-A[98]		ShT-B[98]		JHU-CROWD++[159]		NWPU-CROWD [166]	
Method	**Locations**	**Counts**	**Self-labels**	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
Density estimation Methods													
Fully Supervised													
NCC [164]	✓			85.8	150.6	61.9	99.6	7.4	**11.3**	67.7	258.3	96.9	534.2
GL [165]	✓			84.3	147.5	61.3	95.4	**7.3**	11.7	59.9	259.5	79.3	346.1
BL [154]	✓			85.8	150.6	61.9	99.6	7.4	**11.3**	67.7	258.5	105.4	454.2
MAN [155]	✓			**77.3**	**131.5**	**56.8**	**90.3**	–	–	**53.4**	**209.9**	**76.5**	**323.0**
Semi-supervised													
L2R [103]	▲		✓	148.9	249.8	90.3	153.5	15.6	24.4	–	–	125.0	501.9
UDA [167]	▲			–	–	93.8	157.2	15.7	24.1	–	–	–	–
MT [168]	▲			–	–	94.5	156.1	15.6	24.5	–	–	129.8	515.0
ICT [169]	▲			–	–	92.5	156.8	15.4	23.8	–	–	–	–
MATT [104]	▲	✓		–	–	80.1	129.4	11.7	17.5	–	–	–	–
IRAST [105]	▲		✓	135.6	233.4	86.9	148.9	14.7	22.9	135.0 †	450.7 †	122.1 †	484.3 †
GP [106]	▲			160.0	275.0	102.0	172.0	15.7	27.9	–	–	–	–
PAL [108]	▲			128.1	218.0	72.7	111.6	12.0	18.7	129.6	400.4	178.7	1080.4
AC-AL [107]	▲			–	–	87.9	139.5	13.9	26.2	–	–	–	–
CU [170]	▲			104.0	164.2	70.7	116.6	**9.7**	**17.7**	74.8	281.6	108.7	458.0
COUNT2DENSITY (Semi-sup.) 5%	▲	✓		120.1	201.0	70.5	105.6	13.0	20.1	65.8	189.1	107.5	448.1
COUNT2DENSITY (Semi-sup.) 10%	▲	✓		**102.2**	**152.4**	**62.7**	**98.3**	10.2	16.7	**60.9**	**181.5**	**99.7**	**398.0**
Cross-Domain													
SE+FD [171]	*			–	–	129.3	187.6	16.9	24.7	–	–	–	–
CD-CC [156]	*			198.3	332.9	109.2	168.1	**11.4**	**17.3**	–	–	–	–
BLA [109]	*			198.9	316.1	99.3	145.0	11.9	18.9	–	–	–	–
IRAST [105] †		✓		437.1	498.4	218.5	289.9	64.3	69.9	248.4	683.5	287.3	824.1
Yang et al. [112]		✓		-	-	104.6	145.2	12.3	21.2	-	-	-	-
COUNT2DENSITY		✓		191.2	321.3	95.3	154.9	15.4	23.7	115.8	385.7	132.3	514.8
COUNT2DENSITY (ncc)		✓		164.6	270.5	100.3	165.2	35.1	47.3	102.2	298.2	122.9	489.9
COUNT2DENSITY (bl)		✓		162.4	282.2	97.5	149.8	50.8	79.4	92.4	269.4	125.7	495.4
COUNT2DENSITY (gl)		✓		159.5	268.4	91.9	141.4	18.6	29.9	105.5	382.5	123.1	492.0
COUNT2DENSITY (man)		✓		**149.2**	**258.2**	**91.5**	**135.2**	15.5	24.6	**78.3**	**220.8**	**122.2**	**485.3**

Table 5.2: Comparison of COUNT2DENSITY with several related methodologies. For each approach, we report the type of annotation used: locations, counts, or self-labels. The ✓ symbol indicates that the entire dataset is annotated (fully supervised); the ▲ symbol indicates that only 5-10% of the dataset is annotated (semi-supervised). The * symbol indicates that the entire source domain is annotated (cross-domain). Best results are marked in **bold**, second best results are underlined. Results are compared within each category (fully supervised, semi-supervised, cross-domain). Unless specified with a † symbol, results are taken from the related publications.

5.3.1 Results

Table 5.2 presents the results of COUNT2DENSITY compared with semi-supervised density estimation and cross-domain approaches. For completeness, we also report results obtained by fully supervised methodologies. Finally, Fig. 5.2 shows examples of estimated density maps.

We first compare COUNT2DENSITY against a modified version of IRAST [105], where we trained it using only count-level annotations to remove any contribution from location-level ones. In this scenario, the results in Table 5.2 show that IRAST degrades in performance, yielding higher counting errors compared to our proposed method.

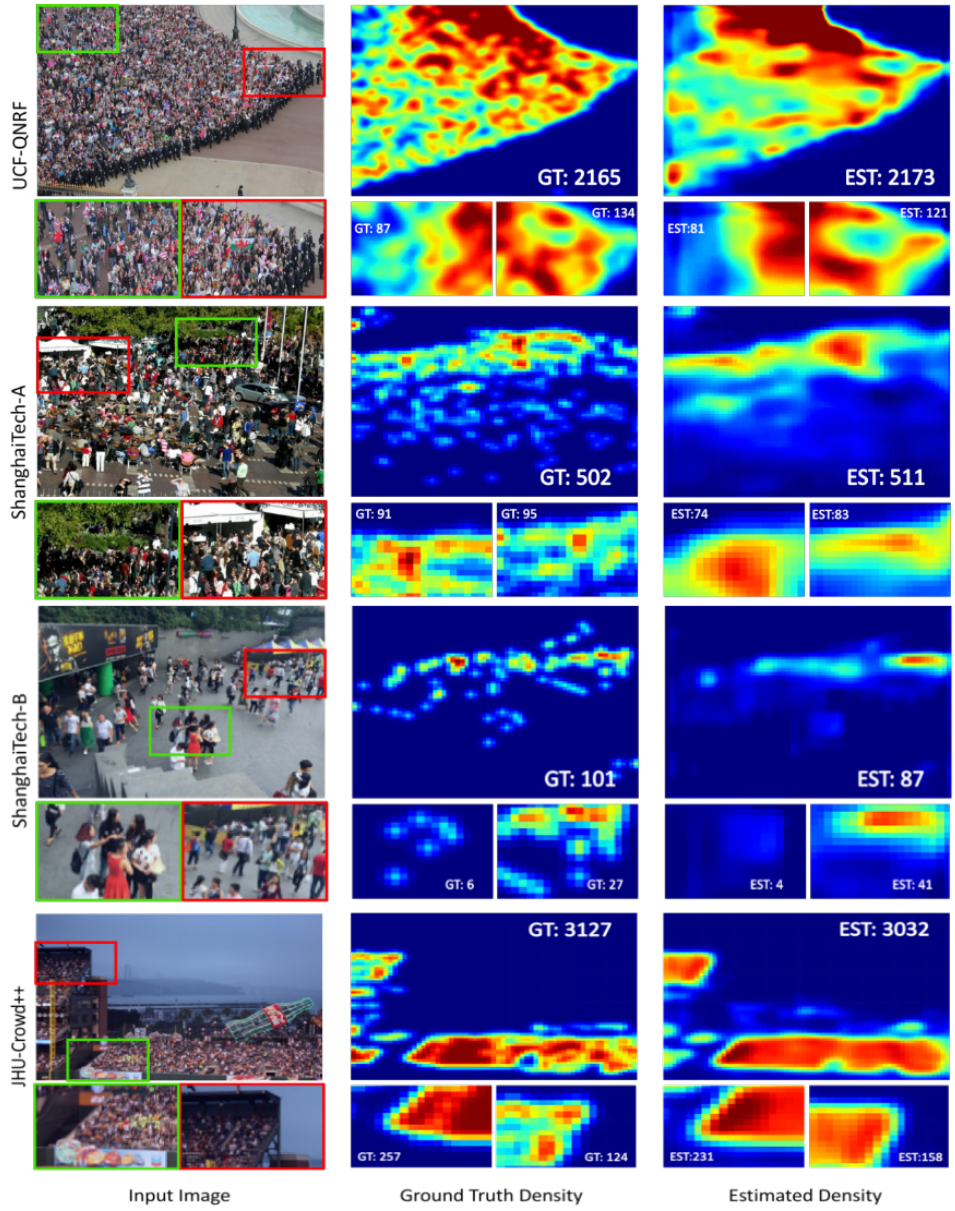

Figure 5.2: Density maps predicted by COUNT2DENSITY compared to ground-truth density, showing both global and subregion counting.

Comparisons with Cross-Domain Approaches. Table 5.2 also presents comparisons with cross-domain approaches performing synthetic-to-real adaptation. Although these approaches use location-level annotations in the source domain, they are unsupervised when transferring to the target domain. Nevertheless, this still constitutes an advantage over COUNT2DENSITY, which never sees location-level

Method	MAE ↓	MSE ↓	SSIM ↑	PSNR ↑
CD-CC [156]	<u>11.9</u>	<u>18.9</u>	<u>0.77</u>	<u>21.6</u>
BLA [109]	12.8	20.6	0.72	21.1
IRAST [105]	27.5	31.8	0.58	18.75
COUNT2DENSITY	**9.3**	**16.1**	**0.85**	**22.5**

Table 5.3: Subregion counting performance and quality of predicted density maps on the UCF-QNRF dataset.

ground-truth annotations during training. Despite this, COUNT2DENSITY drastically outperforms cross-domain approaches, especially on a large-scale dataset such as UCF-QNRF. Cross-domain approaches, however, yield better performance than ours on ShanghaiTech-B, as this dataset is characterised by less dense scenes. This occurs because the sampling strategy in Section 5.2.2 is better suited for highly dense scenes, as the sampled points are more likely to be situated within crowd regions.

Comparisons with Semi-Supervised Approaches. Table 5.2 shows comparisons with semi-supervised methods. Unlike COUNT2DENSITY, these methods require a subset of the training set labelled with location-level annotations. Additionally, some semi-supervised approaches [107] augment the dataset by cropping images into patches, which have been shown to improve counting performance [155, 164, 165]. However, this training strategy cannot be applied in our method because we solely rely on count-level annotations, which are calculated over the entire image and thus lack count information on local patches. Despite these major differences, COUNT2DENSITY achieves comparable performance across all the datasets. Specifically, our method outperforms IRAST and PAL on JHU-CROWD++, reducing the MAE by $\sim 40\%$. On the ShanghaiTech datasets, COUNT2DENSITY outperforms most of the semi-supervised approaches.

We also assess the performance of COUNT2DENSITY when equipped with a few ground-truth location-level annotations and trained in a semi-supervised fashion. This is achieved by slightly modifying our pipeline as follows: for those images in the training set with location-level annotations, we provide the ground-truth M_i instead of the pseudo-density map $\tilde{M}_i$ when evaluating $\mathcal{L}^{map}$ (*c.f.* Equation (5.4)). With this setup, COUNT2DENSITY outperforms the recent CU on ShanghaiTech-A and JHU-CROWD++, achieving the best MAE of 70.5 and 65.8, respectively.

Method	Precision ↑	Recall ↑	F1 ↑
STEERER (Full-sup.) [172]	78.6	72.7	75.5
ADASEEM [173]	**77.7**	10.1	17.8
IRAST [105]	55.4	42.8	48.3
COUNT2DENSITY	75.2	**67.9**	**71.3**

Table 5.4: Localisation performance on the UCF-QNRF dataset.

Quality of Density Maps. Fig. 5.2 shows qualitative examples of predictions made by COUNT2DENSITY compared to the ground-truth. In Table 5.3, we also quantitatively evaluate our predictions, assessing the ability to perform subregion counting. Specifically, for subregion counting, we divide the density maps into small tiles (subregions) and compare the counts between the predicted and ground-truth tiles. For quantitatively assess the quality of the density maps, we calculate the SSIM and PSNR between the predictions and ground truth. The results in Table 5.3 demonstrate that COUNT2DENSITY outperforms competing methods in subregion counting. Additionally, our method achieves higher SSIM and PSNR values, indicating better quality density maps compared to other methods.

To further evaluate the quality of predicted density maps, we follow the methodology in [172] to assess the localisation performance of COUNT2DENSITY. Briefly, we determine location-level points from the predicted density maps by identifying local minima. Table 5.4 shows that COUNT2DENSITY outperforms IRAST and ADASEEM with a higher F1-score, and yields comparable performance with respect to a fully-supervised baseline, such as STEERER [172].

5.3.2 Analysis

Ablation Study. In Table 5.5, we report the results of the ablation study, by removing each individual component from our pipeline (*c.f.* Fig. 5.1). Removing all components leads to the worst performance, with an MAE of 218.2. By adding the Historical Map Bank, we boost performance reducing the MAE by $\sim 30\%$. This result demonstrates that the Historical Map Bank can mitigate the effect of confirmation bias, as discussed in Section 5.2.2. Furthermore, the addition of either the saliency initialisation or the contrastive regularisation improves performance. Finally, it

Historical Map Bank	Saliency Init.	Contrastive Reg.	MAE	MSE
✗	✗	✗	218.2	385.4
✓	✗	✗	158.1	271.1
✗	✓	✓	206.4	354.2
✓	✓	✗	154.7	265.4
✓	✗	✓	152.5	259.0
✓	✓	✓	**149.2**	**258.2**

Table 5.5: COUNT2DENSITY ablation analysis on UCF-QNRF.

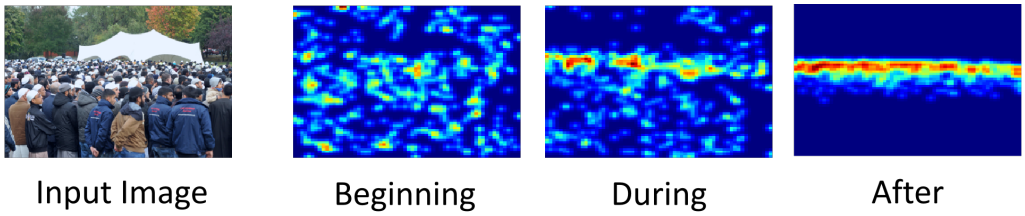

Figure 5.3: Pseudo-density maps generated during training. At the beginning, the pseudo-map is generated using the saliency estimator. During training, we apply Eq. (5.2) to refine the pseudo-map, refining the quantitative spatial information over time.

is clear that COUNT2DENSITY achieves the best performance when all of three components work simultaneously.

Hypergeometric *vs.* Naive Sampling. To demonstrate the effectiveness of sampling from a hypergeometric distribution, as described in Section 5.2.2, we assessed a naive sampling strategy by selecting the top-y_i points from $\mathcal{H}_i^{(t)}$. On UCF-QNRF, the naive sampling strategy yields an MAE of 200.1, while our approach reduces the error by $\sim 25\%$ (our best MAE on UCF-QNRF is 149.2, as reported in Table 5.2).

Evolution of the Pseudo-Density Maps During Training. Fig. 5.3 shows the generated pseudo-density maps at different stages of training. At the first epoch, the pseudo-density maps are generated from the output provided by the unsupervised saliency estimator. Although this provides a spatial prior, the generated maps do not contain much spatial information at the beginning. During the training, our method progressively refines the pseudo-density maps, leading to meaningful estimations toward the end of the training.

Historical Map Bank Initialisation. We assess how much COUNT2DENSITY

depends on a good initialisation of the historical map bank to extract spatial information from count-level annotations. The ablation study in Table 5.5 shows that COUNT2DENSITY performs well without initialising the historical map bank (MAE is $\sim 2\%$ higher than when BAS-NET is used). We took this analysis a step further, and the results are in Table 5.6. We initialised the historical map bank with a naive approach, by placing a blob with a fixed radius at the centre of each pseudo-density map. Although this strategy does not provide accurate spatial information, we observe a reduction in the error by $< 1\%$ when compared with no initialisation. We also compared unsupervised BAS-NET against another unsupervised saliency estimator, TGSD [174], showing that the unsupervised BAS-NET achieves the best performance. For completeness, we trained BAS-NET in a supervised fashion to establish a lower bound performance. Although it is expected that a supervised saliency estimator would improve performance,the such improvement is marginal, reducing the MAE by $< 0.5\%$, showing that COUNT2DENSITY performs well without relying on location-level annotations to initialise $\mathcal{H}$. Overall, it can be observed in Table 5.6 that COUNT2DENSITY is not overly reliant on the initialisation of the historical map bank and the ablation in Table 5.5. Despite the performance reduction observed with the historical map bank, best results are achieved when the moving average in Eq. (5.2) is used in tandem with the contrastive regulariser and saliency initialisation with the unsupervised BAS-NET [162].

Computational efficiency. We train COUNT2DENSITY for 18-24 hours on a single NVIDIA RTX A6000 depending on the backbone used. We observed a peak of about 10GB of GPU ram usage during training and 4GB during inference. For inference, COUNT2DENSITY takes about $0.3s$ to process a single image and provide the corresponding density map.

5.4 Summary

To reduce the cost of collecting fine-grained annotation for crowd density estimation, we introduced COUNT2DENSITY, a method that leverages only count-level annotations to predict meaningful density maps, also enabling subregion counting. Our approach is capable of extracting spatial information from count values to generate pseudo-density maps, which are then used to train the model in a self-supervised

Method	MAE	MSE
None	152.5	259.0
Centred blob	151.4	260.4
TGSD [174]	150.1	**257.9**
Unsup. BAS-NET [162]	**149.2**	258.2
Sup. BAS-NET [162]	148.6	257.4

Table 5.6: Comparison with different saliency estimators used to initialise the historical map bank.

fashion. We show that COUNT2DENSITY outperforms cross-domain approaches by a large margin across several datasets and shows significant improvements in performance when trained in a semi-supervised setting, surpassing the state-of-the-art.

While our method demonstrates strong performance, it is not without limitations. One notable challenge is the memory demands required to store the historical map bank, which may limit the scalability of our approach for larger datasets or real-time applications. Additionally, the training time and computational overhead associated with maintaining the map bank could be prohibitive in certain settings. Future work should focus on more efficient memory management strategies or alternative architectural designs to address these concerns.

Despite these limitations, our method presents a promising direction for crowd density estimation and other domains where spatial information is unavailable during training. COUNT2DENSITY opens up opportunities in challenging scenarios where location-level annotations are scarce or impractical to collect. Overall, our approach offers an effective solution in density estimation, with potential applications far beyond crowd counting.

Chapter 6

Conclusions

This thesis has investigated fundamental challenges and proposed novel methodologies to advance deep learning in scenarios characterized by limited labelled data. Each chapter focused on different aspects of reducing the reliance on large-scale, fully and fine-grained annotated datasets, while improving model adaptability and robustness in complex real-world environments.

Chapter 3 introduced a source-free open-set unsupervised domain adaptation framework designed to transfer knowledge from pretrained models without access to source data, while effectively handling unknown classes. By promoting a better segregation of unseen categories, this approach enhances the flexibility of deep learning systems to operate under realistic constraints where labelled data is unavailable and unseen classes are present.

Chapter 4 explored the integration of language modality within unsupervised domain adaptation, demonstrating that incorporating textual information significantly improves model robustness when adapting across geographically diverse data distributions. This contribution highlights the potential for leveraging multimodal data to overcome domain shifts in regions where annotated datasets are difficult to obtain, facilitating a wider use of AI technologies worldwide.

Chapter 5 proposed a weakly supervised crowd density estimation pipeline that learns to predict spatially meaningful density maps using only count-level annotations. This method reduces the annotation cost and effort compared to existing approaches that rely on fine-grained location-level annotations, enabling practical adoption in surveillance, medical imaging, and other domains with limited annotation resources.

Appendices A and B presents an example of the use of generative models for

data augmentation. Specifically, we investigated the training of diffusion models for modelling MRI trajectories associated with neurodegenerative diseases. This work demonstrates how generative modelling can synthesize realistic data to support the training of deep learning models in context where acquiring large annotated clinical datasets is challenging, thus addressing critical data scarcity issues in medical AI applications.

Collectively, these contributions offer a set of strategies for mitigating the annotation burden for the training of deep learning models, by combining advances in domain adaptation, weak supervision, multimodal integration, and generative data augmentation. Comprehensive experimental evaluation validates the effectiveness of our approaches across multiple datasets, including context such as medical or surveillance.

In summary, this thesis paves the way for more adaptable and robust deep learning frameworks in scenarios where labelled data is limited.

6.1 Limitations and Future Works

Despite the advance illustrated in this thesis, the proposed methodologies have some limitations. For instance, the use of language models depends on the availability and quality of aligned textual data. Weak supervised density estimation with count-level annotations may not fully match the accuracy obtained with fine-grained annotations. Data augmentation using generative models requires further validation for clinical use. Finally, model interpretability remain challenges for practical deployment.

Despite the advance illustrated in this thesis, the proposed methodologies have some limitations. First, the methods explored in this thesis still rely on assumptions about the nature of domain shift, the availability of auxiliary modalities or the structure of weak labels. In real-world settings—such as healthcare or autonomous systems, data variability can be more complex, noisy, or unpredictable than the benchmarks used for evaluation.

A second limitation concerns generalization. Model performance highly depends on the representativeness of the available data. When training data are biased, incomplete, or reflect only a subset of real-world conditions, models may struggle to

generalize to new scenarios. The integration of large-scale sources of data (e.g. web raw data) could be a promising direction.

Another limitation is related to computational cost. While the goal is to reduce annotation effort, the introduced methods remain costly to train and use. This can hinder deployment in settings with limited computing infrastructure or countries with constrained technological resources.

Future works could include a deep investigation of recent foundation models and large language models to further enhance the generalization of vision models in presence of limited annotations. Moreover, future works could broadly investigate technologies to reduce the computational requirements, thus allowing the spread of such technologies. These directions promise to continue pushing the boundaries of adaptable deep learning models that could be widely deployed in real-world scenarios, even in presence of safety critical applications.

Appendix A

TADM: Temporally-Aware Diffusion Model for Neurodegenerative Progression on Brain MRI

This chapter investigates a real-world domain in which the lack of annotated datasets is a major limitation, significantly slowing the development of deep learning technologies. Specifically, we explore the use of generative models to synthesize realistic and diverse medical imaging data, helping to mitigate the scarcity of annotated clinical images. To this end, we propose a novel methodology based on diffusion-based generative models to learn MRI trajectories related to neurodegenerative diseases. Such models can generate new MRIs at predetermined temporal intervals, being valuable not only for data augmentation but also for clinical applications, such as early diagnosis.

The findings and contributions of this chapter have been published in a international conference. [1]

[1] Mattia Litrico, Francesco Guarnera, Mario Valerio Giuffrida, Daniele Ravì, Sebastiano Battiato, "TADM: Temporally-Aware Diffusion Model for Neurodegenerative Progression on Brain MRI" [19], International Conference on Medical Image Computing and Computer-Assisted Intervention, 2024.

A.1 Introduction

The capability to forecast structural changes in brain MRIs over time is critical in medical imaging. The prediction of temporal brain trajectories has proven its usefulness in several applications, such as recovering missing images in longitudinal data, as a potential virtual placebo or for patient stratification [175, 176, 5, 177]; or for diagnosis and prognosis of Alzheimer's disease [178]. While Alzheimer's disease (AD) diagnosis commonly relies on neuropsychological and behavioural assessments, imaging data significantly aid in identifying characteristic disease effects on the brain, even in its early stages [179]. Recent advancements in Artificial Intelligence (AI) have driven the development of sophisticated spatial-temporal disease progression models [180], empowering accurate prediction of brain structural progression. In particular, generative models have been proposed to simulate future MRI scans starting from past MRI scans used as inputs. One of the AI solutions employed in this context involves the training of Generative Adversarial Networks (GANs). For example, 4D-DANINet [5] utilizes adversarial training alongside a series of biologically informed constraints to enhance the generation process. Another approach proposed in [175] uses a conditional 3D GAN with morphology constraints to predict deformations, instead of directly manipulating image pixels. More recently, approaches based on Denoising Diffusion Probabilistic Models (DDPMs) have demonstrated exceptional performance in this domain. For instance, in [181], a diffusion model is combined with a transformer network. The transformer generates a latent representation from a sequence of input MRIs, which is used to condition the generation process of the diffusion model. Another notable solution is Diffusion Deformable Model (DDM) [182], which introduces a methodology to combine a diffusion and a deformation module to generate images that interpolate between two MRI scans. Similarly, DiffuseMorph [183] trains a diffusion model to estimate a deformation field between two scans. This deformation field facilitates the translation of one input image into another, thereby enabling the generation of interpolated scans.

Due to the complexity of age-related changes in brain morphology during disease progression [184], all these approaches often fail to accurately capture the corresponding temporal evolution, facing the following limitations: (i) approaches that operate conditioning on patient's age, such as [175, 5, 185, 186], do not explicitly capture the relationship between structural changes in brain MRIs and the time

interval over which these changes occur; additionally, they require age-balanced datasets, which are often lacking in real-world applications; (ii) approaches based on interpolation, including DDM [182] and DiffuseMorph [183], have limitations in their capability to generate MRIs beyond the two input scans, reducing their relevance in clinical applications that require predicting future scans; and (iii) other approaches, like SADM [181], require longitudinal data at inference time, limiting again their application in real-world contexts where a series of scans for the patient are not available.

To address these issues, we propose TADM, a novel diffusion-based approach for brain progression modelling which operates directly on T1-weighted MRIs. Our model is trained to learn the distribution of brain changes within a specified time interval. To achieve this, we employ a three-fold strategy. Firstly, TADM learns to predict the intensity difference between baseline and follow-up MRIs. This avoids the need to generate entirely new scans, reduces the complexity of the problem, and mitigates generation errors. Secondly, we condition the model on the age gap between the input and output scans rather than directly on the output age, aiming to better learn the relationship between observed differences and the time interval. Given that the same age gap can arise between scans acquired at different ages, conditioning on age gap avoids the necessity of including samples from every age group in the training set. This is particularly beneficial when the dataset has limited samples in some age groups. Lastly, we propose to leverage a Brain-Age Estimator (BAE) to predict age differences between two scans. During training, these predicted age gaps are used in the loss function of our model, allowing the generation of images that accurately reflect the expected age gaps between the inputs and the predictions.

We evaluate our method on the OASIS-3 dataset [187]. Specifically, we compute similarity metrics and region size in three areas of the brain to estimate the difference between real and predicted follow-up brain images. TADM improves similarity metrics by 4% and obtains the best performance on estimating region size reducing the error by 24%. Additionally, our qualitative analysis shows visual improvements in our approach in terms of better mimicking the temporal progression of brains.

In conclusion, the contributions of this work are: (i) introducing TADM, a diffusion-based approach for modelling brain progression trained on T1-weighted MRIs; (ii) learning the distribution of intensity differences between MRI scans to

reduce the complexity of the generation process; (iii) conditioning on age gap to better capture the relationship between brain changes and time intervals; and (iv) proposing to leverage BAE to allow the generation of images that accurately reflect the expected age gap.

A.2 Background

Denoising Diffusion Probabilistic Models (DDPMs) [188] are generative models that learn a Markov chain process to convert a Gaussian distribution into data distribution. During the diffusion process, Gaussian noise is added in successive steps to a sample x_0 from the given data distribution $q(x_0)$, to convert it into a latent variable distribution $q(x_t)$, as follows:

$$q(x_1, x_2, \ldots, x_T|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}) \tag{A.1}$$

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}) \tag{A.2}$$

where $t \in 1, \ldots, T$ is the diffusion step, $\mathcal{N}$ represents the Gaussian distribution, β_t is a noise variance , and $\mathbf{I}$ is the identity matrix.

Then, during the reverse process, the latent variable distribution $p_\theta(x_t)$ is transformed progressively into the data distribution $p_\theta(x_0)$, which is parameterized by θ, by training the model to learn the following Gaussian transformations:

$$p_\theta(x_0, x_1, \ldots, x_{T-1}|x_T) = \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t) \tag{A.3}$$

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_0(x_t, t), \sigma_0(x_t, t)^2\mathbf{I}) \tag{A.4}$$

$$p(x_t) = \mathcal{N}(x_T; \mathbf{0}, \mathbf{I}) \tag{A.5}$$

where $\mu_0(x_t, t)$ represents the mean of the Gaussian distribution, and $\sigma_0(x_t, t)^2$ denotes the variance, at the t reverse step.

A.3 Proposed Method

Here we provide the details of TADM. Our pipeline is depicted in Figure A.1 and consists of the following blocks: (a) a DDPM; (b) an Encoder; and (c) BAE.

During training, we use pairs of MRI scans denoted as I_{T_a} and I_{T_b}, obtained from the same patient at two time points T_a and T_b. These scans are used to compute a residual image $I_{\Delta_{a,b}} = I_{T_b} - I_{T_a}$ that we leverage to train the DDPM aimed at predicting residuals $\widehat{I}_{\Delta_{a,b}}$. To generated the output scan $\widehat{I}_{T_b}$ at time T_b we then add the predicted residual $\widehat{I}_{\Delta_{a,b}}$ to the baseline scan I_{T_a} at time T_a. Additionally, to achieve patient individualization, the DDPM is conditioned with a latent representation extracted by the encoder ϕ from I_{T_a} and other patient-specific data (*i.e.* cognitive status and age). Finally, we leverage BAE to predict the time interval $\Delta_{a,b} = T_b - T_a$ between I_{T_a} and the estimated $\widehat{I}_{T_b}$ to aid the generation process of the DDPM. During inference, we use the scan I_{T_a} at time T_a together with the desired time interval $\Delta_{a,b'}$ as an input to generate a future unseen scan $I_{T_{b'}}$ at time $T_{b'}$ of the same patient.

A.3.1 Conditioning the Diffusion Model

We condition the DDPM to generate residual images using the following information: (i) the image representation $\phi(I_{T_a})$ extracted by the encoder ϕ on the baseline; (ii) the time interval $\Delta_{a,b}$; (iii) other patient's specific data.

Image Representation. To obtain individualization at the subject level, we condition the model using a latent representation z_a of the baseline scan I_{T_a}. In particular, the latent representation is obtained leveraging a pretrained encoder based on Residual-in-Residual Dense Blocks (RRDB) [189].

Time Interval (Age Gap). Conditioning the progression directly on age does not explicitly capture the relationship between structural changes in brain MRIs and the time interval over which these changes occur. Moreover, this strategy necessitates age-balanced datasets, which are difficult to observe in real-world scenarios. To tackle this limitation, we propose to condition the model using the age gap between scans $\Delta_{a,b}$. Since the same age gap can occur between scans acquired at different ages, conditioning on the age gap eliminates the need for including samples from every age group in the training set. This is particularly advantageous when the

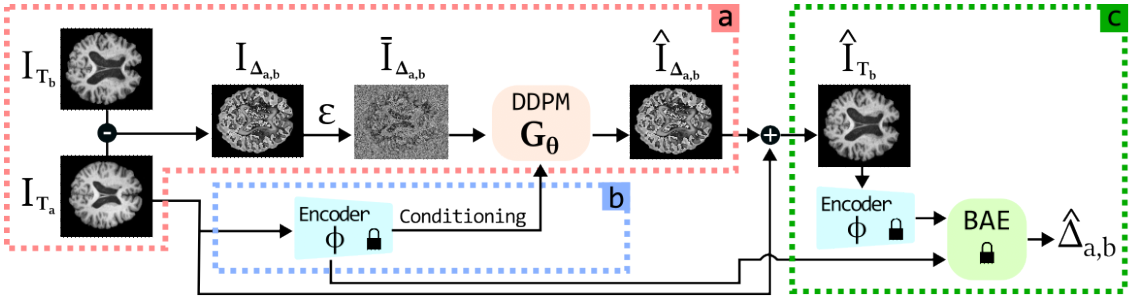

Figure A.1: TADM comprises three main parts (surrounded by coloured dashed boxes). RED: The DDPM takes a residual image, calculated as the difference between two scans acquired at two the time points T_a and T_b, on which random noise ϵ is applied. The DDPM is trained to denoise the residual image through several diffusion steps. The denoised residual image $\widehat{I}_{\Delta_{a,b}}$ and the scan I_{T_a} are summed together to estimate the scan $\widehat{I}_{T_b}$ at time T_b. BLUE: Here, we encode the scan I_{T_a} to extract a representation z_a used to condition the DDPM, in conjunction with other patient-specific data. GREEN: The estimated $\widehat{I}_{T_b}$ is provided to the encoder to extract the features z_b that, together with the previously extracted features z_a, are provided as inputs to a BAE to predict the time interval $\widehat{\Delta}_{a,b}$. The PADLOCK indicates a model with frozen parameters.

dataset has limited samples in certain age groups. In our implementation, we encode $\Delta_{a,b}$ using positional encoding [190] before incorporating it into the model.

Patient-Specific Data. We also condition the model using the patient's cognitive status (D) and age at baseline (A). Indeed, age at baseline is crucial information as diseases progress at different rates over the course of ageing. However, by only using $\Delta_{a,b}$, our model would not capture such age-related progression information. Note that this is different from previous work on age conditioning, as they use age at the prediction rather than age at baseline.

A.3.2 Leveraging BAE to improve the temporal awareness

To encourage the model to generate images that accurately reflect the expected age gap between the input and the prediction, we propose to leverage a BAE model [191] to predict the age gap between two MRI scans. BAE is trained offline on our training set and it is not further fine-tuned during the DDPM training. Specifically, given the baseline I_{T_a} and the generated $\widehat{I}_{T_b}$ scans, the predicted age gap is computed as $\widehat{\Delta}_{a,b} = \Psi(\Phi(\widehat{I}_{T_b})) - \Psi(\Phi(I_{T_a}))$, where Ψ is the BAE model. This information will be used later to train the DDPM and improve the generation. In particular, if the

DDPM generates a scan $\widehat{I}_{T_b}$ that closely approximates the ground truth I_{T_b}, the predicted age gap $\widehat{\Delta}_{a,b}$ should closely match the actual age gap $\Delta_{a,b}$. Any error in the estimation of the age gap will be corrected through backpropagation in the diffusion model to refine the generation process.

A.3.3 Overall Framework

In this section, we will provide a complete overview of the framework.

Training. During the diffusion process, the DDPM is trained to predict the noise ϵ added to the input $I_{\Delta_{a,b}}$. This is obtained by minimising the following objective:

$$\mathcal{L}^{DML} = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,1),\bar{\mathbf{I}}_{\Delta_{a,b}},t}\left[||G_\theta(\bar{\mathbf{I}}_{\Delta_{a,b}}, t; z_a, \Delta_{a,b}, A, D) - \epsilon||_2^2\right] \tag{A.6}$$

where G_θ is the DDPM parametrized by θ and t is the diffusion timestep.

Additionally, as mentioned in Appendix A.3.2, we incorporate the output from BAE as an additional term in the loss function of the DDPM. Specifically, we define the loss on the expected brain age gap as follows:

$$\mathcal{L}^{BAE} = (\widehat{\Delta}_{a,b} - \Delta_{a,b})^2. \tag{A.7}$$

Note that the gradient of this loss updates only the DDPM parameters θ.

Finally, the overall loss is obtained by combining Eqs. (A.6) and (A.7) as follows:

$$\mathcal{L}^{Tot} = \mathcal{L}^{DML} + \mathcal{L}^{BAE} \tag{A.8}$$

Inference. The model takes as input a baseline MRI $\mathbf{I}_{T_a}$ and predicts a follow-up MRI $\mathbf{I}_{T_{b'}}$ with an arbitrarily time interval $\Delta_{a,b'}$ with respect to the baseline. The reverse process starts from a Gaussian noise variable X_T that is progressively denoised through $G_\theta(X_t, t; z_a, \Delta_{a,b'}, A, D)$. The predicted residual image $\hat{\mathbf{I}}_{\Delta_{a,b'}}$ is then added to the baseline MRI to generate the predicted follow-up.

A.4 Experimental Results

Implementation Details. To evaluate our approach, we use 2,535 T1-weighted (T1w) brain MRIs from 634 subjects from the OASIS-3 dataset [187]. Scans were

Table A.1: Comparison study: Results showing the performance in terms of image-base and region size in comparison to other methods.

			Region Size Error (%) ↓			
Method	SSIM ↑	PSNR ↑	Gray Matter	White Matter	Cerebrospinal Fluid	Total Brain
DiffuseMorph [183]	0.68	19.67	10.40 ± 3.45	3.49 ± 2.58	4.65 ± 2.80	46.30 ± 7.51
4D-DaniNet [5]	0.65	16.99	2.21 ± 1.08	2.57 ± 1.98	3.12 ± 3.65	9.31 ± 8.72
DDM [182]	0.69	19.59	2.44 ± 1.35	3.05 ± 2.74	4.37 ± 3.12	10.85 ± 11.64
TADM (Proposed)	**0.72**	**20.51**	**1.69 ± 1.54**	**1.85 ± 2.20**	**2.70 ± 2.29**	**6.84 ± 5.00**

acquired over a period of 15 years and subjects are between 42 and 95 years old, classified as cognitively normal, Mild Cognitive Impairment (MCI), and AD. We apply linear registration through the MNI152 template and skull removal using the FSL library [192] to all the MRIs. The dataset is divided into training set (70%), validation set (10%) and test set (20%). We used the validation set to optimize the hyperparameters of BAE.

Following [193], we adopt the U-Net as architecture of the diffusion model G_θ and the same hyperparameters. Results of all baseline methods are obtained using their publicly available codes.

Evaluation Metrics. We compute image-based metrics and region size in relevant brain areas to assess the performance of our method. Specifically, for the image-based metrics we use the Structural Similarity Index Measure (SSIM) and Peak Signal-to-Noise Ratio (PSNR) between the generated and the actual MRIs. On the other side, region sizes are used to evaluate the accuracy of disease progression. The regions considered in our experiment are: (i) Gray Matter, (ii) White Matter, and (iii) Cerebrospinal Fluid (CSF). We use the FMRIB's Automated Segmentation Tool [192] to compute the area of the regions which are expressed as percentages of the Total Brain to account for individual differences. The error is calculated as the mean absolute error between the area on the region from the predicted scan and ground truth MRIs.

Comparison Results. In Tab. A.1, we present quantitative results obtained by our method compared to other state-of-the-art approaches [183, 182, 5]. The results demonstrate that we outperform the state-of-the-art in SSIM and PSNR by +0.03 and +0.84, respectively. Results on the size of the region show that our method

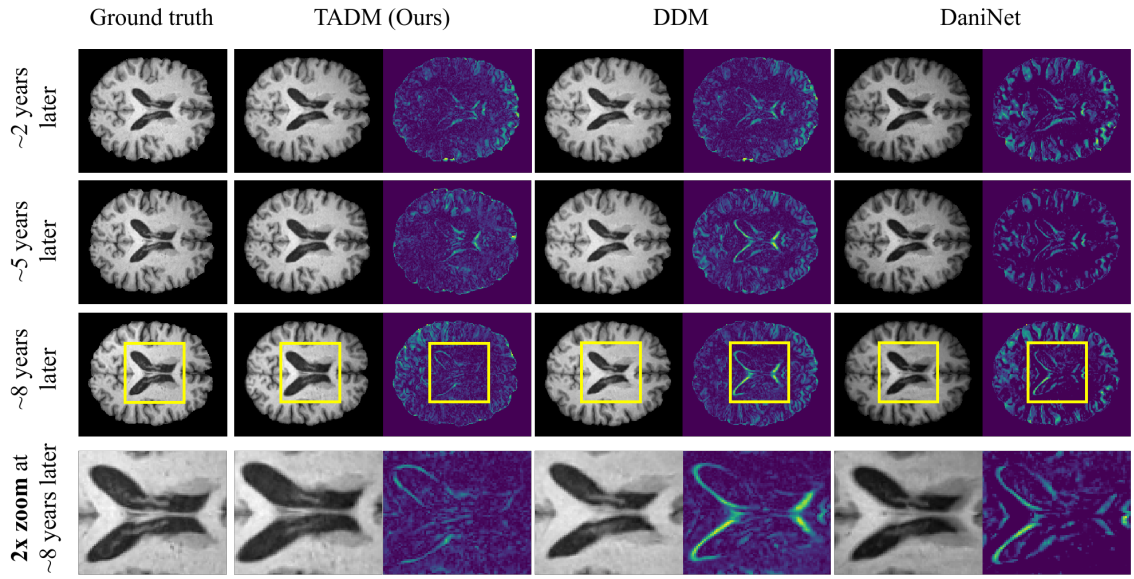

Figure A.2: Comparison of the temporal progression on a 66-year-old subject with AD, obtained by our approach against other state-of-the-art methods. We show predicted slice-MRIs on the left and the corresponding error with the subject's real brain MRI on the right.

achieves the lowest error in all the considered brain regions. In particular, we reduce the error on grey and white matter regions by approximately 30% and 8%, respectively. Regarding the CSF and Total Brain, the reduction of error is nearly 29%, demonstrating that our model generates high-quality follow-up scans compared to the current state-of-the-art. Additionally, in Figure A.2, we show one example of qualitative results in predicting MRI scans at different time points. The figure shows that TADM offers a better approximation of the brain's temporal evolution compared to other methods. Specifically, our predictions depict a notably accurate alignment of the ventricular expansion over time. Finally, TADM exhibits fewer minor disparities in the brain cortex compared to other methods.

Additional Analysis. In this section, we explore the impact of integrating BAE into our pipeline and conditioning the model on patient's patient-specific data. Additionally, we evaluate the model's performance when it is conditioned on either the age or the age gap as proposed in our method. In the first row in Tab. A.2, we show the results of our method without incorporating the patient's specific data.

Table A.2: Additional analysis: results showing the contribution of the different components of TADM, including BAE, the use of patient's specific data. Finally, we also show the impact of using the conditioning on age rather than age gap as we proposed.

			Region Size Error (%) ↓			
Method	SSIM ↑	PSNR ↑	Gray Matter	White Matter	Cerebrospinal Fluid	Total Brain
TADM w/o patient's data	0.71	20.32	1.78 ± 1.44	1.97 ± 2.14	2.72 ± 1.98	7.85 ± 5.17
TADM w/o BAE	0.69	20.08	2.44 ± 2.12	2.02 ± 2.13	3.85 ± 3.67	9.77 ± 8.23
TADM w/ age cond.	0.68	19.71	4.12 ± 3.48	4.98 ± 2.45	5.65 ± 3.32	11.95 ± 7.34
TADM	**0.72**	**20.51**	**1.69 ± 1.54**	**1.85 ± 2.20**	**2.70 ± 2.29**	**6.84 ± 5.00**

This outcome highlights that the absence of patient-specific data results in a minimal reduction in performance, indicating that the patient-specific data contributes minimally to the pipeline. When BAE is not used in our pipeline (second row in Tab. A.2), we notice an evident decrease in performance. This indicates that leveraging BAE is essential to support the generation process. Lastly, in the third row, we observe that conditioning the model on age rather than age gap drastically reduces performance by an average of approximately 60%, demonstrating the effectiveness of our idea of conditioning on age gaps.

A.5 Summary

In this work, we propose TADM, a novel approach designed to accurately mimic brain neurodegenerative progression in MRIs. We evaluated TADM on the OASIS-3 dataset, demonstrating superior performance compared to existing approaches. TADM focuses on 2D scans, as we strive to develop a new data-driven pipeline capable of improving the accuracy of current methods. Nonetheless, our method can be easily extended to 3D scans. Furthermore, our pipeline presents exciting prospects for data augmentation of underrepresented samples in medical imaging datasets. This feature holds significant promise, especially in the context of less common and more expensive modalities, *e.g.* PET and CT scans, where the generation of synthetic samples remains a critical challenge.

Appendix B

Temporally-Aware Diffusion Model for Brain Progression Modelling with Bidirectional Temporal Regularisation

This chapter presents our journal extension to the work illustrated in the previous chapter. The pipeline has been improved to work on 3D MRI scans and it has been empowered with a training regulariser that improves the temporal awareness of the diffusion model.

The findings and contributions of this chapter have been submitted to an international journal. [1]

B.1 Introduction

Predicting how brain structures evolve over time in MRI scans is a key challenge in medical imaging. The modelling of temporal brain trajectories has proven useful in multiple applications, including recovering missing scans in longitudinal data [176], acting as a virtual placebo [5], aiding in patient stratification [175, 176, 5, 177], and supporting both diagnosis and prognosis of Alzheimer's disease (AD) [178]. Although AD diagnosis is often based on neuropsychological and behavioural

[1] Mattia Litrico, Francesco Guarnera, Mario Valerio Giuffrida, Daniele Rav'ı, Sebastiano Battiato, "Temporally-Aware Diffusion Model for Brain Progression Modelling with Bidirectional Temporal Regularisation" [20]. Under Review, Computerized Medical Imaging and Graphics, 2025

evaluations, imaging data plays a crucial role in revealing structural brain changes associated with the disease, even at early stages [179].

Recent advancements in AI have led to the emergence of novel spatio-temporal technologies designed to model disease progression [180], enabling more accurate predictions of structural brain changes. Among these, generative approaches have gained attention for their ability to accurately simulate future MRI scans based on earlier scans. Early methods used Generative Adversarial Networks (GANs) to generate future MRIs [175, 5]. Recently, Denoising Diffusion Probabilistic Models (DDPMs) have become the state-of-the-art for generative models, demonstrating superior performance also in this domain. Sequence-Aware Diffusion Model (SADM) [181] combines a diffusion model with a transformer to predict follow-up MRIs. Similarly, Diffusion Deformable Model (DDM) [182] and DiffuseMorph [183] use diffusion models to estimate the deformation field between scans. Differently, BrLP [194] proposes the training of a latent diffusion model in conjunction with an autoencoder and an auxiliary model to generate individual MRI scans based on longitudinal data.

However, these methods often show poor performance in accurately modelling the temporal dynamics of morphological changes under neurodegenerative diseases. Some approaches [175, 5, 185, 195] do not explicitly capture the relationship between structural MRI changes and time intervals, while others are limited in their ability to generate future scans [182, 183] or require longitudinal data at inference time [181].

To address these challenges, we introduce TADM-3D, a novel 3D diffusion-based framework specifically designed to predict future brain MRI scans. Our method learns the statistical distribution of brain changes over specified time intervals, capturing the complex, non-linear patterns of neuroanatomical trajectories directly on MRI data. Rather than directly generating follow-up MRI scans, which can be prone to artefacts, TADM-3D focuses on predicting the voxel-wise intensity differences between baseline and follow-up images, reducing the complexity of the task and mitigating generation errors.

In contrast to other approaches, our model is conditioned on the age difference between the input and output scans, rather than on the patient's absolute age. This encourages the model to learn the relation between structural brain changes and the time interval between baseline and follow-up scans, rather than with specific age values. Since identical age gaps can occur at various absolute ages, this approach

alleviates the need to include samples from all age groups, which is particularly advantageous when certain age ranges are underrepresented in the training data. For instance, a 5-year gap could apply both to a patient scanned at ages 60 and 65, and to another scanned at 80 and 85. While the output ages differ (65 vs. 85), the age gap remains the same (5 years).

To further improve the temporal awareness, we use a Brain-Age Estimator (BAE) to estimate the age difference between the baseline and the generated scans. During training, these predicted age differences are incorporated into the loss function, encouraging the generation of scans that accurately reflect the expected temporal interval between input and prediction.

Lastly, we introduce a *Back-In-Time Regularisation* (BITR) strategy at training time, to improve the temporal awareness of the model. Based on the intuition that a temporal-aware model should be able to predict both forward and backwards in time, we randomly swap the roles of the baseline and follow-up scans, and we train the model bidirectionally to predict future or past scans, alternately. This simple regularization strategy encourages the model to learn how the brain anatomically changes going forward and backwards in time.

The method is trained and tested on the OASIS-3 dataset [187]. Moreover, we also use an external test set from the NACC dataset to evaluate the generalisation performance of TADM-3D on out-of-distribution data. As evaluation metrics, we use image-based similarity scores and brain region volumes between real and predicted follow-up MRIs. TADM-3D overcomes previous methods on both image-based and volumetric metrics, demonstrating its effectiveness on modelling the temporal brain evolution.

Additionally, our qualitative analysis demonstrates that our method more effectively reproduces the temporal progression of brains.

This work extends our previous MICCAI 2024 conference paper [19] in several ways:

- The method was extended to operate on 3D MRI scans, rather than single 2D slices, allowing the model to capture richer spatial context and more complex anatomical relationships across all three dimensions.

- We introduced BITR, a technique aimed to better model temporal awareness. This approach trains the model by considering not only forward progression but also backwards changes over time.
- The impact of the cognitive status as a conditioning variable was evaluated, allowing the model to account for clinically relevant differences. This conditioning helps capture variations in brain structure associated with different stages of cognitive decline [196].
- The comparisons were extended to include the most recent 3D methods, such as CounterSynth [175] and BrLP [194] and we evaluate TADM-3D on an external test set to assess its generalisability.

B.2 Related Works

The study of neurodegenerative diseases through MRIs has bloomed in the last years [197]. Most existing approaches propose simulators that model temporal changes in brain structure. Simulators receive high-dimensional data as input, such as 3D MRIs of a given subject, and predict the changes in the brain MRIs over a specified period and under specific subject's conditions, such as neurodegenerative diseases. Most of the existing simulators are based on recent deep generative methods, including Generative Adversarial Networks (GANs) [175, 5, 186], Variational Autoencoders (VAEs) [198], Normalising Flows (NFs) [199] and Diffusion Models (DMs) [194, 181].

Early methods [5, 186] used GANs to simulate subject-specific brain changes conditioned on the presence of a neurodegenerative disease. For instance, DANINet [5] employs adversarial training in conjunction with biological constraints to improve the generation process. Nevertheless, these approaches produce synthetic 2D slices, without making use of full 3D volumetric information. More recently, CounterSynth [175] introduced a 3D GAN-based framework that models 3D deformations, instead of directly manipulating image pixels. By leveraging morphology constraints, deformations are applied to the input MRI to reflect brain changes over time. However, CounterSynth only predicts structural changes without modelling temporal evolutions. Other related approaches proposed methodologies based on VAEs [198]

and NFs [199], but they produce low resolution scans and rely on morphologically constrained transformations, limiting their applicability.

More recently, the use of diffusion models has been introduced in medical image generation. DDM [182] proposed to combine a diffusion and a deformation module to learn how to interpolate between two MRIs. This combination enables the modelling of smooth anatomical transitions by simulating plausible intermediate brain states. Similarly, DiffuseMorph [183] leveraged a diffusion model to predict deformation fields between two MRIs. This field is then applied to warp one image into another, effectively enabling the synthesis of interpolated images along a continuous trajectory. Both DDM and DiffuseMorph are effective in learning realistic anatomical transformations, but they are limited to interpolation tasks. As a result, their applicability in clinical settings is constrained, since they are unable to generate predictive scans.

SADM [181] introduced a sequence-aware diffusion model, incorporating a transformer-based architecture to better handle temporal dependencies in longitudinal data. The transformer generates a sequence of prior MRIs to extract a latent representation that encodes the subject's disease trajectory. This representation is then used to condition the generation process of the diffusion model, allowing for the synthesis of future brain MRIs. It necessarily requires a sequence of input MRI scans, which are rarely available in a real-world context, especially at an early stage of the patient's treatment. Moreover, it does not use subject metadata, which is important to better model the disease progression. BrLP [194] introduced a methodology that leverages an autoencoder, a latent diffusion model, and an auxiliary model (ControlNet) to generate individual MRI scans. However, the quality of the produced scans is highly correlated to the effectiveness of the autoencoder features.

TADM-3D overcomes these limitations as follows: (i) unlike CounterSynth, our model learns to model the temporal progression of brain changes, making it capable of forecasting future scans; (ii) in contrast to DDM and DiffuseMorph, TADM-3D is conditioned with age differences, allowing it to predict scans at arbitrary future time points; (iii) unlike SADM, TADM-3D does not need a sequence of input MRI scans, improving its applicability in clinical context. Moreover, our model incorporates subject-specific metadata to improve the accuracy and personalisation of predictions; (iv) differently than BrLP, we train the diffusion model to directly

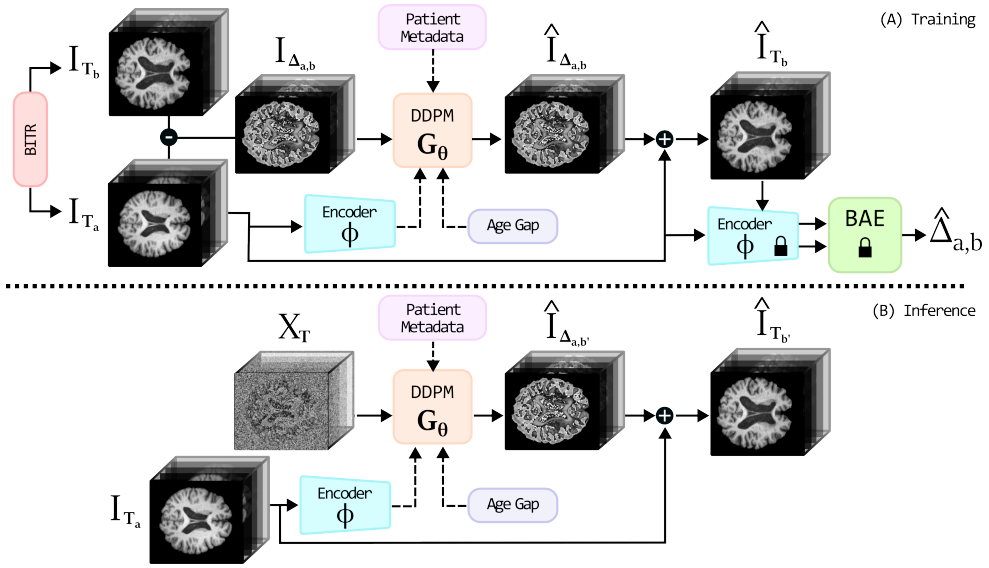

Figure B.1: **TADM-3D.** (A) Given a baseline scan I_{T_a} and a follow-up scan I_{T_b}, a residual image $I_{\Delta_{a,b}}$ is computed by subtracting the two scans. This residual is then corrupted with noise and denoised by a Denoising Diffusion Probabilistic Model (DDPM), producing the denoised residual $\widehat{I}_{\Delta a,b}$. The denoised residual image and the scan I_{T_a} are added together to estimate the scan $\widehat{I}_{T_b}$ at time T_b (see Appendix B.3.2). The baseline scan is encoded with Φ to produce a latent representation z_a, which, along with patient metadata, is used to condition the DDPM (see Appendix B.3.3). The estimated follow-up scan $\widehat{I}_{T_b}$ is also encoded to extract z_b. Representations z_a and z_b are then fed into a BAE to predict the time interval $\widehat{\Delta}_{a,b}$ between the scans (see Appendix B.3.4). To promote time awareness, the model is also trained to predict past scans with a probability $p = 0.5$ by swapping the roles of baseline and follow-up scans (see Appendix B.3.5). The PADLOCK indicates frozen parameters. Dashed lines indicate model conditioning. (B) At inference time, given the baseline scan I_{T_a} and random noise X_T, the DDPM predicts the residual $\widehat{\mathbf{I}}_{\Delta_{a,b'}}$ that summed to the baseline I_{T_a} produces the predicted follow-up $\widehat{\mathbf{I}}_{T_{b'}}$.

predict MRI scans without using any auxiliary model, reducing the computational requirements and improving the prediction quality.

B.3 Proposed Method

B.3.1 Background

DDPMs [188] are generative models that learn how to gradually convert a Gaussian data distribution into another distribution by applying a Markov chain process.

During the forward process (diffusion), Gaussian noise is progressively and incrementally added to the input data x_0 from the given data distribution $q(x_0)$. This occurs over a fixed number of steps, gradually transforming the sample into a latent variable distribution $q(x_t)$, as following:

$$q(x_1, x_2, \ldots, x_T|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}) \tag{B.1}$$

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}) \tag{B.2}$$

where $t \in 1, \ldots, T$ indicates the diffusion step, $\mathcal{N}$ is the Gaussian distribution, β_t is the variance of the noise, and $\mathbf{I}$ is the identity matrix.

During the reverse process (denoising), the model is trained to learn to invert the diffusion process, by progressively turning the noise latent variable distribution $p_\theta(x_t)$ into the data distribution $p_\theta(x_0)$, parameterised by θ. During this process, the model is trained to learn the Gaussian transformations, as follows:

$$p_\theta(x_0, x_1, \ldots, x_{T-1}|x_T) = \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t) \tag{B.3}$$

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_0(x_t, t), \sigma_0^2(x_t, t)\mathbf{I}) \tag{B.4}$$

$$p(x_t) = \mathcal{N}(x_T; \mathbf{0}, \mathbf{I}) \tag{B.5}$$

where $\sigma_0^2(x_t, t)$ is the variance at the step t and $\mu_0(x_t, t)$ is the mean of the Gaussian distribution. Once trained, the model is able to denoise random noise, generating high-quality images.

B.3.2 Temporally-Aware Diffusion Model

Our proposed training pipeline, depicted in Figure B.1, comprises three main components: (i) a Denoising Diffusion Probabilistic Model (DDPM), responsible for modelling the brain progression and generating high quality scans; (ii) an Encoder, which extracts latent representation from MRI scans; and (iii) a Brain-Age Estimator (BAE), which is leveraged to encourage temporal consistency by estimating the time interval between scans.

During training, we use MRI pairs, *i.e.* I_{T_a} and I_{T_b}, corresponding to scans acquired from the same subject at two distinct timepoints T_a and T_b, respectively. From these images, we compute a residual image $I_{\Delta_{a,b}} = I_{T_b} - I_{T_a}$, which captures the voxel-wise intensity changes occurring over the time interval $\Delta_{a,b} = T_b - T_a$. This residual is used to train the DDPM at predicting the residual $\widehat{I}_{\Delta_{a,b}}$. To generate the output follow-up scans $\widehat{I}_{T_b}$, we add the predicted residual to the baseline scan, such that $\widehat{I}_{T_b} = I_{T_a} + \widehat{I}_{\Delta a,b}$. Additionally, we leverage the BAE model to estimate the time interval $\widehat{\Delta}_{a,b}$ between I_{T_a} and the estimated $\widehat{I}_{T_b}$. The estimated time interval is then integrated in the loss function to regularise the DDPM generation process. At inference, we utilize the scan I_{T_a} acquired at time T_a along with the target time interval $\Delta_{a,b'}$ to produce a future scan $I_{T_{b'}}$ at time $T_{b'}$. To achieve patient individualization during the generation process, the DDPM is conditioned on a latent embedding obtained by applying an encoder Φ to the baseline scan I_{T_a}, along with auxiliary patient-specific metadata (*e.g.*, cognitive status and age).

B.3.3 Conditioning Strategies

To generate residual images, we condition the DDPM using: (i) baseline latent representation $\Phi(I_{T_a})$ extracted by the encoder Φ on the baseline scan I_{T_a}; (ii) the time interval $\Delta_{a,b}$ between T_a and T_b; (iii) patient's specific metadata.

Baseline Scan Encoding. Our goal is to predict patient-specific brain trajectories. To achieve this, the DDPM is conditioned with the encoded representation z_a derived from the baseline scan I_{T_a}, using an encoder Φ. This latent features capture the patient-specific anatomy, enabling the generation of personalised outputs.

Age Difference. Using age directly to model progression fails to explicitly capture the temporal evolution of structural changes in brain MRIs, and it demands age-balanced datasets, which are often unavailable in real-world settings. To address this issue, we propose conditioning the model on the age difference between scans, denoted as $\Delta_{a,b}$. Since the same age gaps can arise from scan pairs acquired at different ages, this strategy reduces the need of uniformly sampling all age groups during training. This is especially beneficial when certain age ranges are underrepresented in the dataset. In our approach, we encode $\Delta_{a,b}$ using positional encoding [190] before integrating it into the model.

Patient-Specific Metadata. We also use both the age (A) of the patient at time T_a, and cognitive status (D) to further condition our model. Baseline age provides essential context to the model, as neurodegenerative diseases progress differently depending on the age of the patient. Note that, unlike previous approaches that condition using the age of the patient w.r.t. the follow-up scan, we use the age at baseline to accurately ground the starting point of the disease trajectory.

B.3.4 Estimating Brain Age for Improving the Temporal Awareness

To guide the model in generating scans that accurately reflect the expected age difference between the input and predicted scans, we incorporate a Brain Age Estimator (BAE) [191] into our training pipeline. During DDPM training, the BAE evaluates whether the generated follow-up scan exhibits brain changes consistent with the expected time interval from the baseline. By including the predicted age difference as part of the loss function, this mechanism encourages the diffusion model to generate outputs that not only appear realistic but also accurately capture the desired temporal progression. This approach provides feedback to the model, thereby enhancing the temporal consistency of the predictions. The BAE model is pre-trained on the training set, and its parameters remain fixed during the DDPM training.

Given a baseline scan I_{T_a} and a generated follow-up scan $\widehat{I}_{T_b}$, we compute the estimated age difference as $\widehat{\Delta}_{a,b} = \Psi(\Phi(\widehat{I}_{T_b})) - \Psi(\Phi(I_{T_a}))$, where Φ is the encoder and Ψ is the BAE model. When the DDPM prediction $\widehat{I}_{T_b}$ accurately matches the ground-truth image I_{T_b}, the estimated age difference $\widehat{\Delta}_{a,b}$ aligns with the actual age gap $\Delta_{a,b}$, ensuring that the generated output reflects the correct temporal progression. Any differences between the predicted and true age gaps are backpropagated during training, serving as a regularization. This feedback loop iteratively refines the generation process, enhancing the model's ability to produce scans that accurately reflect the temporal progression.

B.3.5 Back-In-Time Regularisation

TADM-3D is trained using pairs of MRI scans I_{T_a} and I_{T_b} acquired from the same patient at two distinct time points. Here I_{T_a} indicates the baseline scan acquired

at age A, and I_{T_b} is the follow-up scan acquired after a time interval $\Delta_{a,b}$. In this way, the diffusion model learns to predict the brain structural changes that occur due to the combined effect of ageing and disease progression over the time interval. However, a temporal-aware model should also be able to predict past scans, *e.g.* generating a baseline scan I_{T_a} from a follow-up scan I_{T_b}.

To this aim, we propose to include a *Back-In-Time Regularisation* (BITR) strategy. Specifically, at each training step, we randomly swap the roles of the two scans I_{T_a} and I_{T_b} with a probability $p = 0.5$. Consequently, the diffusion model is trained to predict the follow-up from the baseline scan for approximately half of the training iterations and the baseline from the follow-up in the other half of the training steps. Note that this swapping is exclusively applied in the training phase. Although simple, this strategy encourages the model to learn a better relationship between anatomical changes in the brain and time interval, leading to an improved temporal awareness.

B.3.6 Overall Framework

Training. A complete overview of the training process is described in Algorithm 1. During the diffusion process, the DDPM learns to estimate the noise ϵ incorporated into $I_{\Delta_{a,b}}$. This is achieved by minimising the following loss function:

$$\mathcal{L}^{DML} = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,1),\bar{\mathbf{I}}_{\Delta_{a,b}},t}\left[||G_\theta(\bar{\mathbf{I}}_{\Delta_{a,b}}, t; z_a, \Delta_{a,b}, A, D) - \epsilon||_2^2\right], \tag{B.6}$$

where, G_θ represents the DDPM with parameters θ, and t denotes the diffusion timestep. Furthermore, as discussed in Appendix B.3.4, the prediction of BAE is incorporated as an extra component in the loss function. Specifically, we define the loss on the expected brain age difference as follows:

$$\mathcal{L}^{BAE} = (\widehat{\Delta}_{a,b} - \Delta_{a,b})^2. \tag{B.7}$$

Finally, the overall loss is obtained by combining Eqs. (B.6) and (B.7):

$$\mathcal{L}^{Tot} = \mathcal{L}^{DML} + \mathcal{L}^{BAE} \tag{B.8}$$

Algorithm 1 Pseudocode of TADM-3D's training process.

Input: Pairs of MRI scans (I_{T_a}, I_{T_b}), time interval $\Delta_{a,b}$, patient metadata (age at baseline A, cognitive status D).
Output: Trained DDPM model G_θ for generating future MRI scans
Training:
for each training step **do**
  Sample a pair (I_{T_a}, I_{T_b}) from the training set
  Compute residual image $I_{\Delta_{a,b}} = I_{T_b} - I_{T_a}$
  Extract latent representation
  $z_a = \Phi(I_{T_a})$ using the encoder Φ
  Sample $p \sim \mathcal{B}(0.5)$
  if $p = 1$ **then**
    Flip sign in $\Delta_{a,b}$, A, and D
  end if
  Sample noise $\epsilon \sim \mathcal{N}(0, 1)$
  Compute noisy residual $\bar{\mathbf{I}}_{\Delta_{a,b}}$ at diffusion step t
  Predict noise using DDPM:
  $\hat{\epsilon} = G_\theta(\bar{\mathbf{I}}_{\Delta_{a,b}}, t; z_a, \Delta_{a,b}, A, D)$
  Compute DDPM loss: $\mathcal{L}^{DML} = ||\hat{\epsilon} - \epsilon||_2^2$
  Generate predicted residual $\widehat{I}_{\Delta_{a,b}}$ from the DDPM
  Compute predicted follow-up scan $\widehat{I}_{T_b} = I_{T_a} + \widehat{I}_{\Delta_{a,b}}$
  Extract latent representation $z_b = \Phi(\widehat{I}_{T_b})$
  Predict age difference using BAE: $\widehat{\Delta}_{a,b} = \Psi(z_b) - \Psi(z_a)$
  Compute BAE loss: $\mathcal{L}^{BAE} = (\widehat{\Delta}_{a,b} - \Delta_{a,b})^2$
  Compute total loss: $\mathcal{L}^{Tot} = \mathcal{L}^{DML} + \mathcal{L}^{BAE}$
  Update DDPM parameters θ using gradients of $\mathcal{L}^{Tot}$
end for

Inference. Given a baseline MRI $\mathbf{I}_{T_a}$, the model generates a future MRI $\widehat{\mathbf{I}}_{T_{b'}}$ at any desired time interval $\Delta_{a,b'}$ after the baseline. The generation process starts with a random Gaussian noise input X_T, which is iteratively refined by the network $G_\theta(X_t, t; z_a, \Delta_{a,b'}, A, D)$. The generated residual image $\hat{\mathbf{I}}_{\Delta a,b'}$ is finally combined with the baseline scan to yield the follow-up prediction.

B.4 Experimental Results

B.4.1 Datasets

We train TADM-3D on 2,535 T1-weighted (T1w) brain MRI scans from 634 subjects from the OASIS-3 dataset [187]. Scans span a longitudinal interval of approximately 15 years, capturing a broad spectrum of aging-related changes. The dataset includes participants aged between 42 and 95 years classified as cognitively normal (CN), mild cognitive impairment (MCI), and Alzheimer's disease (AD). We evaluate TADM-3D on both internal and external test sets, to also asses generalisation performance with out-of-distribution data. For the external dataset, we used data from the NACC dataset [200], including 2,257 T1w MRIs from 962 subjects. Scan have a maximum interval between the initial and follow-up MRI of 13 years, with an average of 3.8. About 75% of the patients are classified as CN at the last visit, while the remaining 25% are MCI or AD. Data from the external dataset is used solely for evaluation.

B.4.2 Implementation Details

All MRI volumes were linearly registered using the MNI152 template to normalise spatial orientation and scale, and skull stripping using the FSL library [192]. The dataset is divided into a training set (70%), a validation set (10%) and a test set (20%). For both the training of diffusion and BAE models, we use the AdamW optimiser with a learning rate of 0.0001 and a weight decay of 0.001, a batch size of 16, and a cosine-based learning rate scheduler. Following [193], we employ the U-Net architecture as backbone of the diffusion model G_θ. For the encoder Φ, we use a 3D UNet-based architecture trained together with the diffusion model in an end-to-end manner. Results for all state-of-the-art methods were obtained using their publicly available implementations.

B.4.3 Evaluation Metrics

To evaluate the performance of our method, we employed both image-based similarity metrics and region-specific volumetric analyses in anatomically relevant brain areas. For the image-based evaluation, we use the *Structural Similarity Index Measure* (SSIM) and *Mean Squared Error* (MSE), measuring the similarity between the

Table B.1: Results on the internal test set of OASIS-3. Performance are evaluated in terms of image-based and region volumes errors wth respect to other methods. The MAE in region volumes is expressed as a percentage of total brain volume.

			Region Volumes Error (%) ↓				
Method	MSE ↓	SSIM ↑	Hippocampus	Amygdala	Lat. Ventricle	Thalamus	CSF
DaniNet [5]	0.016 ± 0.007	0.623 ± 0.162	0.030 ± 0.030	0.018 ± 0.017	0.257 ± 0.222	0.038 ± 0.030	1.081 ± 0.814
CounterSynth [175]	0.008 ± 0.004	0.861 ± 0.052	0.030 ± 0.018	0.016 ± 0.010	0.273 ± 0.311	0.041 ± 0.035	0.881 ±0.672
BrLP [194]	0.005 ± 0.002	0.887 ± 0.017	0.028 ± 0.018	0.017 ± 0.009	0.264 ± 0.271	0.039 ± 0.021	0.882 ± 0.645
TADM-3D	**0.004 ± 0.001**	**0.902 ± 0.014**	**0.017 ± 0.019**	**0.015 ± 0.014**	**0.228 ± 0.187**	**0.027 ± 0.015**	**0.642 ± 0.412**

Table B.2: Results on the external test set of NACC evaluating TADM-3D's generalisation performance in terms of image-based and region volumes errors in comparison to other methods. The MAE in region volumes is expressed as a percentage of total brain volume.

			Region Volumes Error (%) ↓				
Method	MSE ↓	SSIM ↑	Hippocampus	Amygdala	Lat. Ventricle	Thalamus	CSF
DaniNet [5]	0.017 ± 0.007	0.611 ± 0.181	0.032 ± 0.031	0.018 ± 0.016	0.232 ± 0.210	0.039 ± 0.032	1.154 ± 0.871
CounterSynth [175]	0.011 ± 0.003	0.813 ± 0.042	0.030 ± 0.020	0.014 ± 0.010	0.283 ± 0.314	0.111 ± 0.034	1.173 ± 0.731
BrLP [194]	0.005 ± 0.002	**0.909 ± 0.023**	0.024 ± 0.023	0.014 ± 0.013	**0.213 ± 0.350**	0.030 ± 0.024	1.044 ± 0.788
TADM-3D	**0.004 ± 0.002**	0.902 ± 0.017	**0.020 ± 0.020**	**0.013 ± 0.011**	0.235 ± 0.200	**0.029 ± 0.021**	**0.833 ± 0.543**

generated and ground-truth MRI scans in terms of structural fidelity and pixel-wise reconstruction accuracy, respectively. On the other side, volumetric metrics in AD-related regions (lateral ventricles, cerebrospinal fluid as CSF, hippocampus, amygdala, and thalamus) are computed to evaluate the modelling of disease progression in TADM-3D. We use *SynthSeg* 2.0 [201] to segment the brain and compute the region volumes, which are expressed as percentages of the entire brain to account for personal variations.

The accuracy of predicted regional volumes is then assessed by computing the Mean Absolute Error (MAE) with the ground-truth volumes.

B.4.4 Comparative Analysis

Tab. B.1 shows the quantitative results on the internal test set obtained by TADM-3D compared to other state-of-the-art 3D approaches [175, 194, 5, 181]. Overall, TADM-3D outperforms them across the board, reducing the MSE and increasing the SSIM by +0.001 and +0.15, respectively. Results on the region's volumes show that TADM-3D achieves the lowest error in different brain regions. In particular,

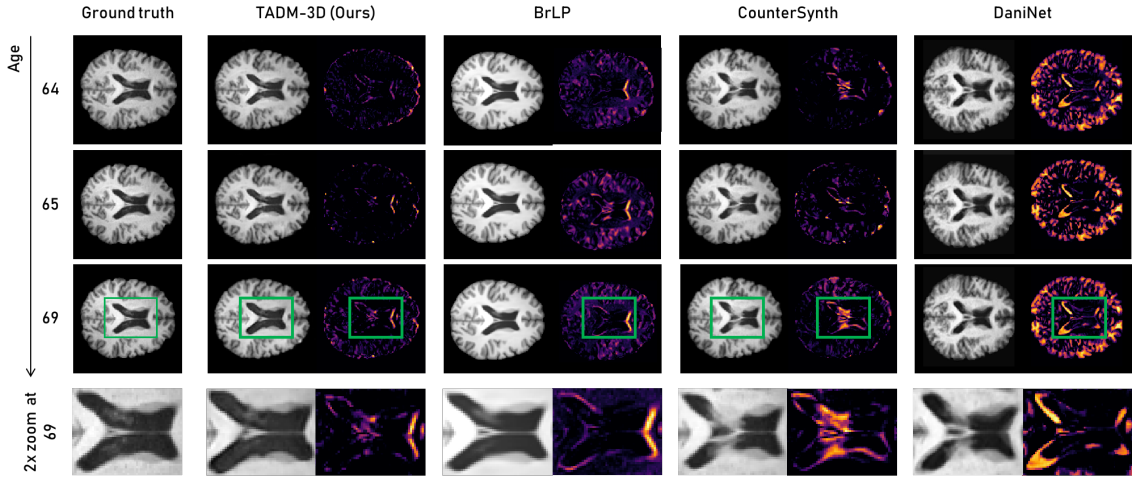

Figure B.2: Temporal progression on a 62-year-old subject with AD from the internal test set, generated by our approach against BrLP[194], CounterSynth[175] and DaniNet[5]. We show predicted MRIs on the centre slice on the left and the corresponding heatmap of prediction error on the right.

we reduce the error on Hippocampus, Amygdala and Ventricles by approximately 40%, 12%, and 13%, respectively. On Thalamus and CSF, TADM-3D reduces the error by 30% and 27%, respectively. These results show that TADM-3D generates accurate follow-up images with respect to existing approaches, achieving better performance in both image-based and volumetric metrics. On the external test set (see Tab. B.2), TADM-3D obtains similar performance w.r.t. the internal evaluation, showing robustness to generalisation challenges. TADM-3D achieves the best performance on the MSE and on volumetric errors of 4 out of 5 analysed brain regions, while BrLP unexpectedly improves its results w.r.t. the internal test set on SSIM and on volumetric error of the Lateral Ventricle, obtaining the lowest error.

Furthermore, Figure B.2 presents a sample of qualitative results illustrating the prediction of MRI scans at various ages for a 62-year-old patient diagnosed with AD.

Clearly, TADM-3D offers a better approximation of the temporal evolution of the brain compared to all other methods. Specifically, our predictions show a notable accuracy in modelling the ventricular expansion over time, maintaining a lower error at every age difference. In comparison to BrLP [194] and DaniNet [5], our model

Table B.3: Ablation studies: showing the results of the absence of the different components of TADM-3D. Finally, we also show the impact of using a 2D model to individually generate scans rather than a native 3D architecture.

			Region Volumes Error (%) ↓				
Method	MSE ↓	SSIM ↑	Hippocampus	Amygdala	Lat. Ventricle	Thalamus	CSF
TADM-3D w/o patient metadata	0.006 ± 0.001	0.899 ± 0.017	0.019 ± 0.011	**0.015 ± 0.018**	0.235 ± 0.201	0.030 ± 0.019	0.648 ± 0.557
TADM-3D w/o age gaps cond.	0.011 ± 0.006	0.812 ± 0.016	0.027 ± 0.013	0.017 ± 0.021	0.251 ± 0.231	0.037 ± 0.029	0.747 ± 0.555
TADM-3D w/o BAE	0.009 ± 0.005	0.872 ± 0.024	0.025 ± 0.014	0.016 ± 0.025	0.249 ± 0.212	0.036 ± 0.026	0.691 ± 0.512
TADM-3D w/o BITR	0.008 ± 0.003	874 ± 0.021	0.023 ± 0.016	0.016 ± 0.022	0.241 ± 0.200	0.031 ± 0.021	0.667 ± 0.491
TADM-2.5D	0.014 ± 0.006	0.779 ± 0.089	0.030 ± 0.027	0.016 ± 0.015	0.299 ± 0.277	0.060 ± 0.041	0.899 ± 0.643
TADM-3D	**0.004 ± 0.001**	**0.902 ± 0.014**	**0.017 ± 0.019**	**0.015 ± 0.014**	**0.228 ± 0.187**	**0.027 ± 0.015**	**0.642 ± 0.412**

reaches superior performance, especially on the ventricular region. CounterSynth [175] seems to produce similar results, but TADM-3D exhibits more balanced and consistently lower errors.

B.4.5 Ablation Studies

In this section, we present an ablation study to assess the contribution of the key components of TADM-3D. Specifically, we evaluate the effect of removing conditioning on age differences, patient-specific conditioning variables, and excluding the BAE module and BITR from training. Furthermore, we compare the performance of our 3D generative architecture against a 2D alternative, where slices are generated independently to reconstruct full 3D volumes (named TADM-2.5D).

Tab. B.3 (first row) presents the performance of TADM-3D without using patient metadata. Results show that, removing the conditioning with such metadata, we observe a minimal performance loss. In the second row, we observe that removing the conditioning of the model on age gaps highly drops performance, demonstrating the effectiveness of our strategy over the current state-of-the-art. Another case of performance drop is when BAE is not used (third row), indicating its usefulness in supporting the generation process. In the fourth row, we show the impact of removing the BITR with a significant drop of performance, demonstrating its contribution in improving the training. Finally, in the last row, we assess the performance using a 2D diffusion model (as our previous work [19]) rather than a 3D one. To reconstruct full 3D volumes, we train a 2D model to generate each slice independently, using the slice index as a conditioning variable to preserve spatial coherence across

Table B.4: Evaluating the impact of incorrect conditioning on cognitive status in TADM-3D predictions.

			Region Volumes Error (%) ↓				
Method	MSE ↓	SSIM ↑	Hippocampus	Amygdala	Lat. Ventricle	Thalamus	CSF
TADM-3D w/ wrong cond.	0.006 ± 0.005	0.895 ± 0.012	0.030 ± 0.013	0.017 ± 0.021	0.234 ± 0.187	0.030 ± 0.026	0.679 ± 0.417
TADM-3D	**0.004 ± 0.001**	**0.902 ± 0.014**	**0.017 ± 0.019**	**0.015 ± 0.014**	**0.228 ± 0.187**	**0.027 ± 0.015**	**0.642 ± 0.412**

the volume. Results show a drastically reduction in performance when using the 2D model, demonstrating that a 3D model captures better structural information during training.

B.4.6 Is TADM-3D Modelling the Disease Progression?

Following the protocol proposed in [202], we perform an experiment to investigate the potential prediction bias of TADM-3D toward modelling healthy ageing trajectories. Specifically, we aim to assess whether the model's outputs are appropriately generated with respect to the patient's cognitive status. The goal is to assess whether TADM-3D can effectively learn to generate normal and pathological trajectories. To do so, we use our trained model with scans obtained from patients with AD, but we deliberately condition the model, specifying that the patient is Cognitively Normal (CN). We then generate predicted future scans under both the incorrect (CN) and correct (AD) cognitive conditions, and compare the resulting volumetric predictions. This setup allows us to isolate the influence of cognitive status on the generative behaviour of the model. Results in Tab. B.4 show that prediction errors generally increase when incorrect cognitive conditioning is applied, particularly in the hippocampal region, a structure notably affected in AD. This supports the hypothesis that the model is not biased toward healthy ageing and is capable of capturing distinct anatomical progressions associated with neurodegeneration.

B.5 Discussions

B.5.1 Limitations

Despite the promising results achieved by TADM-3D, some limitations still warrant further investigation. Specifically, Figure B.3 illustrates an example of the

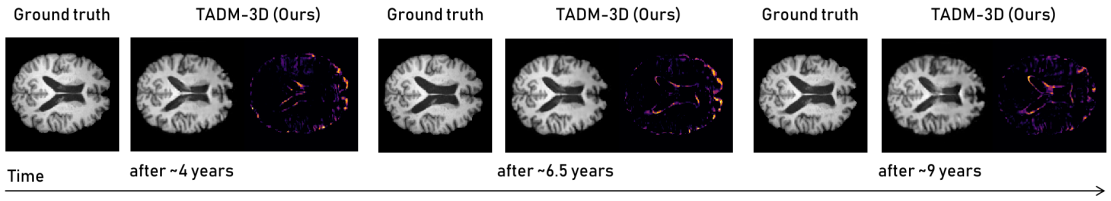

Figure B.3: Example of limitation of TADM-3D in predicting long-time trajectory for a 65-year-old patient. The brain's evolution in predictions after 6.5 and 9 years is inconsistent, showing limitations in capturing long-time dependencies.

challenges faced by TADM-3D in predicting long-term trajectories for a 65-year-old patient. The model tends to lose accuracy when predicting brain evolution over long time intervals, exhibiting increased errors, particularly for intervals spanning several years between scans. This highlights a difficulty in effectively capturing long-term temporal dependencies and predicting distant future anatomical changes. Another limitation is that, although TADM-3D is conditioned on age differences and some patient metadata, it does not yet incorporate other important clinical factors such as genetic information, comorbidities, or medication effects, that could significantly influence disease trajectories.

B.5.2 Applications

Disease progression modelling plays a crucial role in many aspects of clinical practice and holds significant promise for improving patient care. By simulating the anatomical changes that occur over time, TADM-3D enables earlier diagnosis and more accurate prognostic assessments, which are essential for timely and targeted interventions. Predicting personalised progression trajectories allows for individualized treatment planning, thereby optimising care strategies. In the context of clinical trials, our model can generate virtual follow-up scans or synthetic control arms, which can enhance trial design and reduce patient burden. TADM-3D also helps overcome limitations caused by missing or irregularly acquired longitudinal scans by effectively filling gaps and supporting robust longitudinal analyses. From a research perspective, modelling disease trajectories provides insights into the underlying pathophysiological mechanisms and temporal dynamics, contributing to novel biomarker discovery. Moreover, our model supports educational efforts by

delivering visual representations of expected progression, improving communication among clinicians, patients, and their families. Finally, TADM-3D may help address data scarcity and biases in medical research by synthesising data for underrepresented populations and costly imaging modalities, including uncommon and expensive scans such as PET and CT, where generating synthetic samples remains particularly challenging.

In summary, these applications highlight the importance of our model for precision medicine, enabling earlier, more personalised, and more effective care for progressive diseases such as Alzheimer's.

B.5.3 Future Works

Building upon our pipeline, multiple directions could be followed to further advance TADM-3D. One direction is to enhance the model's ability to capture longer-term temporal dependencies. Another one involves integrating multiple data modalities beyond structural MRI, such as other medical scans, genomic information and clinical records. This multi-modal integration could provide more comprehensive and personalised information on the neurodegenerative progression. Additionally, improving the model's robustness and adaptability to diverse clinical and imaging settings could be beneficial. Indeed, approaches based on federated learning frameworks could facilitate model deployment across institutions without requiring data sharing. Finally, extending the framework to other organs would further broaden its clinical impact.

B.6 Summary

In this work, we introduced TADM-3D, a diffusion-based approach for 3D brain progression modelling that directly predicts the intensity difference between baseline and follow-up MRI scans, effectively capturing structural changes over time. Our method addresses several key limitations of existing techniques by conditioning predictions on age differences rather than on target ages, allowing for more accurate temporal modelling without requiring age-balanced datasets. Furthermore, we proposed a Back-In-Time Regularization strategy, enhancing the model's temporal awareness by training the model to predict both forward and backwards in time.

We extensively evaluated TADM-3D on the OASIS-3 dataset, achieving superior results in both similarity metrics and region volume estimation. Moreover, to assess the generalisability of our method, we also tested TADM-3D on an external test set from the NACC dataset, achieving comparable results w.r.t. the internal evaluation. Qualitative analyses further highlighted the model's capacity to better capture the progression of brain structures over time, particularly in regions highly susceptible to neurodegenerative diseases.

Bibliography

[1] M. H. Mohd Noor and A. O. Ige. "A survey on state-of-the-art deep learning applications and challenges". In: *Engineering Applications of Artificial Intelligence* 159 (2025), p. 111225. ISSN: 0952-1976. DOI: https://doi.org/10.1016/j.engappai.2025.111225. URL: https://www.sciencedirect.com/science/article/pii/S0952197625012266.

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *ArXiv* abs/2010.11929 (2020). URL: https://api.semanticscholar.org/CorpusID:225039882.

[3] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows". In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), pp. 9992–10002. URL: https://api.semanticscholar.org/CorpusID:232352874.

[4] D. Nie, R. Trullo, J. Lian, L. Wang, C. Petitjean, S. Ruan, Q. Wang, and D. Shen. "Medical Image Synthesis with Deep Convolutional Adversarial Networks". In: *IEEE Transactions on Biomedical Engineering* 65.12 (2018), pp. 2720–2730. DOI: 10.1109/TBME.2018.2814538.

[5] D. Ravi, S. B. Blumberg, S. Ingala, F. Barkhof, D. C. Alexander, N. P. Oxtoby, A. D. N. Initiative, et al. "Degenerative adversarial neuroimage nets for brain scan simulations: Application in ageing and dementia". In: *Medical Image Analysis* 75 (2022), p. 102257.

[6] F. Rundo, F. Trenta, A. L. di Stallo, and S. Battiato. "Grid Trading System Robot (GTSbot): A Novel Mathematical Algorithm for Trading FX Market".

In: *Applied Sciences* 9.9 (2019). ISSN: 2076-3417. DOI: 10.3390/app9091796. URL: https://www.mdpi.com/2076-3417/9/9/1796.

[7] F. Ragusa, V. Tomaselli, A. Furnari, S. Battiato, and G. M. Farinella. "Food vs Non-Food Classification". In: *Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management*. MADiMa '16. Amsterdam, The Netherlands: Association for Computing Machinery, 2016, 77–81. ISBN: 9781450345200. DOI: 10.1145/2986035.2986041. URL: https://doi.org/10.1145/2986035.2986041.

[8] C. Jin, Z. Guo, Y. Lin, L. Luo, and H. Chen. "Label-Efficient Deep Learning in Medical Image Analysis: Challenges and Future Directions". In: *CoRR* abs/2303.12484 (2023). URL: https://doi.org/10.48550/arXiv.2303.12484.

[9] T. S. Kim, G. E. Jang, S. Lee, and T. Kooi. "Did you get what you paid for? Rethinking annotation cost of deep learning based computer aided detection in chest radiographs". In: *arXiv preprint arXiv:2209.15314* (2022). URL: https://arxiv.org/abs/2209.15314.

[10] H. W. Liao, C. Klugmann, D. Kondermann, and F. Mahmood. "Minority reports: balancing cost and quality in ground truth data annotation". In: *arXiv preprint arXiv:2504.09341* (2025). URL: https://arxiv.org/abs/2504.09341.

[11] C. Bouchard, R. Bernatchez, and F. Lavoie-Cardinal. "Addressing annotation and data scarcity when designing machine learning strategies for neurophotonics". In: *Frontiers in Neurophotonics* 7 (2023), p. 10447257. DOI: 10.3389/fnph.2023.10447257. URL: https://www.frontiersin.org/articles/10.3389/fnph.2023.10447257/full.

[12] A. Hosna, E. Merry, J. Gyalmo, Z. Alom, Z. Aung, and M. A. Azim. "Transfer learning: a friendly introduction". In: *Journal of Big Data* 9.1 (2022), p. 102. DOI: 10.1186/s40537-022-00652-w. URL: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-022-00652-w.

[13] Z.-H. Zhou. "A brief survey of weakly supervised learning". In: *ACM Computing Surveys* 51.2 (2018), pp. 1–35. DOI: 10.1145/3179993. URL: https://dl.acm.org/doi/10.1145/3179993.

[14] R. Mullapudi and et al. "Semi-supervised learning for deep neural networks". In: *Neurocomputing* 507 (2023), pp. 90–103. DOI: 10.1016/j.neucom.2022.12.051. URL: https://www.sciencedirect.com/science/article/pii/S0925231222013257.

[15] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al. "On the opportunities and risks of foundation models". In: *arXiv preprint arXiv:2108.07258* (2021). URL: https://arxiv.org/abs/2108.07258.

[16] M. Litrico, D. Talon, S. Battiato, A. D. Bue, M. V. Giuffrida, and P. Morerio. *Uncertainty-guided Open-Set Source-Free Unsupervised Domain Adaptation with Target-private Class Segregation*. 2024. arXiv: 2404.10574 [cs.CV]. URL: https://arxiv.org/abs/2404.10574.

[17] M. Litrico, M. V. Giuffrida, S. Battiato, and D. Tuia. *TRUST: Leveraging Text Robustness for Unsupervised Domain Adaptation*. 2025. arXiv: 2508.06452 [cs.CV]. URL: https://arxiv.org/abs/2508.06452.

[18] M. Litrico, F. Chen, M. Pound, S. A. Tsaftaris, S. Battiato, and M. V. Giuffrida. *Count2Density: Crowd Density Estimation without Location-level Annotations*. 2025. arXiv: 2509.03170 [cs.CV]. URL: https://arxiv.org/abs/2509.03170.

[19] M. Litrico, F. Guarnera, M. V. Giuffrida, D. Ravì, and S. Battiato. " TADM: Temporally-Aware Diffusion Model for Neurodegenerative Progression on Brain MRI ". In: *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*. Vol. LNCS 15002. Springer Nature Switzerland, 2024.

[20] M. Litrico, F. Guarnera, M. V. Giuffrida, D. Ravì, and S. Battiato. *Temporally-Aware Diffusion Model for Brain Progression Modelling with Bidirectional Temporal Regularisation*. 2025. arXiv: 2509.03141 [cs.CV]. URL: https://arxiv.org/abs/2509.03141.

[21] Y. Ganin and V. Lempitsky. "Unsupervised Domain Adaptation by Backpropagation". In: *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*. ICML'15. Lille, France: JMLR.org, 2015, 1180–1189.

[22] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang. "Moment matching for multi-source domain adaptation". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2019, pp. 1406–1415.

[23] S. Tan, X. Peng, and K. Saenko. "Class-Imbalanced Domain Adaptation: An Empirical Odyssey". In: *Computer Vision – ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part I*. Glasgow, United Kingdom: Springer-Verlag, 2020, 585–602. ISBN: 978-3-030-66414-5. DOI: 10.1007/978-3-030-66415-2_38. URL: https://doi.org/10.1007/978-3-030-66415-2_38.

[24] M. Long, H. Zhu, J. Wang, and M. I. Jordan. "Deep transfer learning with joint adaptation networks". In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. ICML'17. Sydney, NSW, Australia: JMLR.org, 2017, 2208–2217.

[25] G. Wei, C. Lan, W. Zeng, Z. Zhang, and Z. Chen. "ToAlign: task-oriented alignment for unsupervised domain adaptation". In: *Proceedings of the 35th International Conference on Neural Information Processing Systems*. NIPS '21. Red Hook, NY, USA: Curran Associates Inc., 2024. ISBN: 9781713845393.

[26] L. Chen, H. Chen, Z. Wei, X. Jin, X. Tan, Y. Jin, and E. Chen. "Reusing the Task-specific Classifier as a Discriminator: Discriminator-free Adversarial Domain Adaptation". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 7171–7180. URL: https://api.semanticscholar.org/CorpusID:248069256.

[27] T. Kalluri and M. Chandraker. "Cluster-to-adapt: Few Shot Domain Adaptation for Semantic Segmentation across Disjoint Labels". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Los Alamitos, CA, USA: IEEE Computer Society, 2022, pp. 4120–4130. DOI: 10.1109/CVPRW56347.2022.00457. URL: https://doi.ieeecomputersociety.org/10.1109/CVPRW56347.2022.00457.

[28] S. Li, M. Xie, F. Lv, C. H. Liu, J. Liang, C. Qin, and W. Li. "Semantic Concentration for Domain Adaptation". In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), pp. 9082–9091. URL: https://api.semanticscholar.org/CorpusID:236986856.

[29] D. Chen, D. Wang, T. Darrell, and S. Ebrahimi. "Contrastive Test-time Adaptation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022.

[30] M. Litrico, A. Del Bue, and P. Morerio. "Guiding Pseudo-labels with Uncertainty Estimation for Source-free Unsupervised Domain Adaptation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023.

[31] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell. "CyCADA: Cycle-Consistent Adversarial Domain Adaptation". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by J. Dy and A. Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, 2018, pp. 1989–1998. URL: https://proceedings.mlr.press/v80/hoffman18a.html.

[32] J. Zhu, H. Bai, and L. Wang. *Patch-Mix Transformer for Unsupervised Domain Adaptation: A Game Perspective*. 2023. arXiv: 2303.13434 [cs.CV]. URL: https://arxiv.org/abs/2303.13434.

[33] T. Xu, W. Chen, P. Wang, F. Wang, H. Li, and R. Jin. "CDTrans: Cross-domain Transformer for Unsupervised Domain Adaptation". In: *ArXiv* abs/2109.06165 (2021). URL: https://api.semanticscholar.org/CorpusID:237490346.

[34] Y. Sun, X. Wang, Z. Liu, J. Miller, A. Efros, and M. Hardt. "Test-Time Training with Self-Supervision for Generalization under Distribution Shifts". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by H. D. III and A. Singh. Vol. 119. Proceedings of Machine Learning Research. Virtual: PMLR, 2020, pp. 9229–9248. URL: https://proceedings.mlr.press/v119/sun20b.html.

[35] D. Wang, E. Shelhamer, S. Liu, B. Olshausen, and T. Darrell. "Tent: Fully Test-Time Adaptation by Entropy Minimization". In: *International Conference on Learning Representations*. 2021. URL: https://openreview.net/forum?id=uXl3bZLkr3c.

[36] J. Liang, D. Hu, and J. Feng. "Do We Really Need to Access the Source Data? Source Hypothesis Transfer for Unsupervised Domain Adaptation".

In: *Proceedings of the 37th International Conference on Machine Learning*. ICML'20. Virtual: JMLR.org, 2020.

[37] S. Yang, Y. Wang, J. van de Weijer, L. Herranz, and S. Jui. "Generalized Source-Free Domain Adaptation". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, pp. 8978–8987.

[38] R. Li, Q. Jiao, W. Cao, H.-S. Wong, and S. Wu. "Model Adaptation: Unsupervised Domain Adaptation Without Source Data". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 9638–9647. DOI: 10.1109/CVPR42600.2020.00966.

[39] J. Nath Kundu, N. Venkat, M. V. Rahul, and R. Venkatesh Babu. "Universal Source-Free Domain Adaptation". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 4543–4552. DOI: 10.1109/CVPR42600.2020.00460.

[40] D. Wang, S. Liu, S. Ebrahimi, E. Shelhamer, and T. Darrell. *On-Target Adaptation*. 2022. URL: https://openreview.net/forum?id=6ooiNCGZa5K.

[41] S. Bucci, M. R. Loghmani, and T. Tommasi. "On the Effectiveness of Image Rotation for Open Set Domain Adaptation". In: *Computer Vision – ECCV 2020*. Ed. by A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm. Cham: Springer International Publishing, 2020, pp. 422–438. ISBN: 978-3-030-58517-4.

[42] J. Jang, B. Na, D. Shin, M. Ji, K. Song, and I.-C. Moon. *Unknown-Aware Domain Adversarial Learning for Open-Set Domain Adaptation*. 2022. arXiv: 2206.07551 [cs.LG].

[43] H. Liu, Z. Cao, M. Long, J. Wang, and Q. Yang. "Separate to Adapt: Open Set Domain Adaptation via Progressive Separation". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 2922–2931. DOI: 10.1109/CVPR.2019.00304.

[44] K. Saito, S. Yamamoto, Y. Ushiku, and T. Harada. "Open Set Domain Adaptation by Backpropagation". In: *Computer Vision – ECCV 2018*. Ed. by V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss. Cham: Springer International Publishing, 2018, pp. 156–171. ISBN: 978-3-030-01228-1.

[45] Q. Wang, F. Meng, and T. Breckon. "Progressively Select and Reject Pseudo-labelled Samples for Open-Set Domain Adaptation". In: *ArXiv* abs/2110.12635 (2021). URL: https://api.semanticscholar.org/CorpusID:239768329.

[46] J. N. Kundu, S. Bhambri, A. R. Kulkarni, H. Sarkar, V. Jampani, and V. B. R. "Subsidiary Prototype Alignment for Universal Domain Adaptation". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. Curran Associates, Inc., 2022, pp. 29649–29662. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/bf121b033db3bac31c3193e8a0dcbf66-Paper-Conference.pdf.

[47] S. Yang, Y. Wang, K. Wang, S. Jui, and J. van de Weijer. "One Ring to Bring Them All: Towards Open-Set Recognition under Domain Shift". In: *arXiv preprint arXiv:2206.03600* (2022).

[48] J. N. Kundu, N. Venkat, A. Revanur, M. V. Rahul, and R. V. Babu. "Towards Inheritable Models for Open-Set Domain Adaptation". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, 2020, pp. 12373–12382. DOI: 10.1109/CVPR42600.2020.01239. URL: https://doi.ieeecomputersociety.org/10.1109/CVPR42600.2020.01239.

[49] Y. Shiqi, W. Yaxing, W. Kai, J. Shangling, and v. d. w. Joost. "Attracting and dispersing: A simple approach for source-free domain adaptation". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. New Orleans, LA (USA): Curran Associates, Inc., 2022.

[50] C. S. Jahan and A. Savakis. "Unknown Sample Discovery for Source Free Open Set Domain Adaptation". In: *ArXiv* abs/2312.03767 (2023). URL: https://api.semanticscholar.org/CorpusID:266052705.

[51] S. Qu, T. Zou, F. Röhrbein, C. Lu, G. Chen, D. Tao, and C. Jiang. "Upcycling Models under Domain and Category Shift". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023.

[52] S. Qu, T. Zou, L. He, F. Röhrbein, A. Knoll, G. Chen, and C. Jiang. "LEAD: Learning Decomposition for Source-free Universal Domain Adaptation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024.

[53] M. Caron, P. Bojanowski, A. Joulin, and M. Douze. "Deep Clustering for Unsupervised Learning of Visual Features". In: *Computer Vision – ECCV 2018*. Ed. by V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss. Cham: Springer International Publishing, 2018, pp. 139–156. ISBN: 978-3-030-01264-9.

[54] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. "A simple framework for contrastive learning of visual representations". In: *Proceedings of the 37th International Conference on Machine Learning*. ICML'20. JMLR.org, 2020.

[55] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. Hinton. "Big Self-Supervised Models Are Strong Semi-Supervised Learners". In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS'20. Vancouver, BC, Canada: Curran Associates Inc., 2020. ISBN: 9781713829546.

[56] X. Chen and K. He. "Exploring Simple Siamese Representation Learning". In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 15745–15753. DOI: 10.1109/CVPR46437.2021.01549.

[57] S. Gidaris, P. Singh, and N. Komodakis. "Unsupervised Representation Learning by Predicting Image Rotations". In: *ArXiv* abs/1803.07728 (2018).

[58] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko. "Bootstrap Your Own Latent a New Approach to Self-Supervised Learning". In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS'20. Vancouver, BC, Canada: Curran Associates Inc., 2020. ISBN: 9781713829546.

[59] K. He, H. Fan, Y. Wu, S. Xie, and R. B. Girshick. "Momentum Contrast for Unsupervised Visual Representation Learning". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 9726–9735. URL: https://api.semanticscholar.org/CorpusID:207930212.

[60] G. Larsson, M. Maire, and G. Shakhnarovich. "Colorization as a Proxy Task for Visual Understanding". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), pp. 840–849.

[61] M. Noroozi and P. Favaro. "Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles". In: *ECCV*. 2016.

[62] K. Saito, D. Kim, S. Sclaroff, and K. Saenko. "Universal Domain Adaptation through Self-Supervision". In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS'20. Vancouver, BC, Canada: Curran Associates Inc., 2020. ISBN: 9781713829546.

[63] K. Saito, D. Kim, S. Sclaroff, and K. Saenko. "Universal Domain Adaptation through Self Supervision". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 16282–16292. URL: https://proceedings.neurips.cc/paper/2020/file/bb7946e7d85c81a9e69fee1cea4a087c-Paper.pdf.

[64] A. Maracani, R. Camoriano, E. Maiettini, D. Talon, L. Rosasco, and L. Natale. "Key Design Choices in Source-Free Unsupervised Domain Adaptation: An In-depth Empirical Analysis". In: *arXiv preprint arXiv:2402.16090* (2024).

[65] D.-H. Lee. "Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks". In: 2013. URL: https://api.semanticscholar.org/CorpusID:18507866.

[66] K. Sohn, D. Berthelot, C.-L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, and C. Raffel. "FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence". In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS'20. Vancouver, BC, Canada: Curran Associates Inc., 2020. ISBN: 9781713829546.

[67] J. Liang, D. Hu, and J. Feng. "Domain Adaptation with Auxiliary Target Domain-Oriented Classifier". In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 16627–16637. DOI: 10.1109/CVPR46437.2021.01636.

[68] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. "Understanding Deep Learning (Still) Requires Rethinking Generalization". In: *Commun. ACM* 64.3 (2021), 107–115. ISSN: 0001-0782.

[69] Y. Wang, X. Ma, Z. Chen, Y. Luo, J. Yi, and J. Bailey. "Symmetric Cross Entropy for Robust Learning With Noisy Labels". In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), pp. 322–330.

[70] A. Ghosh, H. Kumar, and P. S. Sastry. "Robust Loss Functions under Label Noise for Deep Neural Networks". In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. AAAI'17. San Francisco, California (USA): AAAI Press, 2017, 1919–1925.

[71] X. Ma, H. Huang, Y. Wang, S. Romano, S. Erfani, and J. Bailey. "Normalized Loss Functions for Deep Learning with Noisy Labels". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by H. D. III and A. Singh. Vol. 119. Proceedings of Machine Learning Research. Virtual: PMLR, 2020, pp. 6543–6553.

[72] Y. Kim, J. Yim, J. Yun, and J. Kim. "NLNL: Negative Learning for Noisy Labels". In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), pp. 101–110.

[73] J. Goldberger and E. Ben-Reuven. "Training deep neural-networks using a noise adaptation layer". In: *ICLR*. 2017.

[74] T. Liu and D. Tao. "Classification with Noisy Labels by Importance Reweighting". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38.3 (2016), pp. 447–461. DOI: 10.1109/TPAMI.2015.2456899.

[75] L. Jiang, Z. Zhou, T. Leung, J. Li, and F.-F. Li. "MentorNet: Learning Data-Driven Curriculum for Very Deep Neural Networks on Corrupted Labels". In: *ICML*. 2018.

[76] K.-H. Lee, X. He, L. Zhang, and L. Yang. "CleanNet: Transfer Learning for Scalable Image Classifier Training with Label Noise". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), pp. 5447–5456.

[77] H. Wei, L. Feng, X. Chen, and B. An. "Combating Noisy Labels by Agreement: A Joint Training Method with Co-Regularization". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 13723–13732. DOI: 10.1109/CVPR42600.2020.01374.

[78] J. Li, R. Socher, and S. C. H. Hoi. "DivideMix: Learning with Noisy Labels as Semi-supervised Learning". In: *ArXiv* abs/2002.07394 (2020).

[79] W. Ahmed, P. Morerio, and V. Murino. "Cleaning Noisy Labels by Negative Ensemble Learning for Source-Free Unsupervised Domain Adaptation". In: *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2022, pp. 356–365. DOI: 10.1109/WACV51458.2022.00043.

[80] T. Kalluri, W. Xu, and M. Chandraker. "GeoNet: Benchmarking Unsupervised Adaptation across Geographies". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023).

[81] S. Chen, Y. Zhang, W. Jiang, J. Lu, and Y. Zhang. *VLLaVO: Mitigating Visual Gap through LLMs*. 2024. arXiv: 2401.03253 [cs.CV]. URL: https://arxiv.org/abs/2401.03253.

[82] L. Dunlap, C. Mohri, D. Guillory, H. Zhang, T. Darrell, J. E. Gonzalez, A. Raghunathan, and A. Rohrbach. *Using Language to Extend to Unseen Domains*. 2023. arXiv: 2210.09520 [cs.CV]. URL: https://arxiv.org/abs/2210.09520.

[83] Z. Wang, L. Zhang, L. Wang, and M. Zhu. "LanDA: Language-Guided Multi-Source Domain Adaptation". In: *ArXiv* abs/2401.14148 (2024). URL: https://api.semanticscholar.org/CorpusID:267212046.

[84] G. Liu and Y. Wang. *TDG: Text-guided Domain Generalization*. 2023. arXiv: 2308.09931 [cs.CV]. URL: https://arxiv.org/abs/2308.09931.

[85] Z. Huang, A. Zhou, Z. Lin, M. Cai, H. Wang, and Y. J. Lee. "A Sentence Speaks a Thousand Images: Domain Generalization through Distilling CLIP with Language Guidance". In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* (2023), pp. 11651–11661. URL: https://api.semanticscholar.org/CorpusID:262217079.

[86] S. Goyal, A. Kumar, S. Garg, Z. Kolter, and A. Raghunathan. "Finetune like you pretrain: Improved finetuning of zero-shot vision models". In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, 2023, pp. 19338–19347. DOI: 10.1109/CVPR52729.2023.01853. URL: https://doi.ieeecomputersociety.org/10.1109/CVPR52729.2023.01853.

[87] T. Kalluri, B. Majumder, and M. Chandraker. "Tell, Don't Show! Language Guidance Eases Transfer Across Domains in Images and Videos". In: *ICML* (2024). URL: https://arxiv.org/abs/2403.05535.

[88] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. "Learning Transferable Visual Models From Natural Language Supervision". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by M. Meila and T. Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 2021, pp. 8748–8763. URL: https://proceedings.mlr.press/v139/radford21a.html.

[89] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. V. Le, Y.-H. Sung, Z. Li, and T. Duerig. "Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision". In: *International Conference on Machine Learning*. 2021. URL: https://api.semanticscholar.org/CorpusID:231879586.

[90] F. Liang, B. Wu, X. Dai, K. Li, Y. Zhao, H. Zhang, P. Zhang, P. Vajda, and D. Marculescu. "Open-Vocabulary Semantic Segmentation with Mask-adapted CLIP". In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 7061–7070. URL: https://api.semanticscholar.org/CorpusID:252780581.

[91] J. Cho, G. Nam, S. Kim, H. Yang, and S. Kwak. "PromptStyler: Prompt-driven Style Generation for Source-free Domain Generalization". In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. Los Alamitos, CA, USA: IEEE Computer Society, 2023, pp. 15656–15666. DOI: 10.1109/ICCV51070.2023.01439. URL: https://doi.ieeecomputersociety.org/10.1109/ICCV51070.2023.01439.

[92] C. Ge, R. Huang, M. Xie, Z. Lai, S. Song, S. Li, and G. Huang. "Domain Adaptation via Prompt Learning". In: *IEEE transactions on neural networks and learning systems* PP (2022). URL: https://api.semanticscholar.org/CorpusID:246823759.

[93] S. Tang, W. Su, M. Ye, and X. Zhu. "Source-Free Domain Adaptation with Frozen Multimodal Foundation Model". In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 23711–23720. URL: https://api.semanticscholar.org/CorpusID:265466880.

[94] L.-C. Chen, G. Papandreou, I. Kokkinos, K. P. Murphy, and A. L. Yuille. "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40 (2016), pp. 834–848. URL: https://api.semanticscholar.org/CorpusID:3429309.

[95] S. Ren, K. He, R. Girshick, and J. Sun. "Faster R-CNN: towards real-time object detection with region proposal networks". In: *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1*. NIPS'15. Montreal, Canada: MIT Press, 2015, 91–99.

[96] H. Fan, H. Su, and L. Guibas. " A Point Set Generation Network for 3D Object Reconstruction from a Single Image ". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, July 2017, pp. 2463–2471. DOI: 10.1109/CVPR.2017.264. URL: https://doi.ieeecomputersociety.org/10.1109/CVPR.2017.264.

[97] Z. Yan, R. Zhang, H. Zhang, Q. Zhang, and W. Zuo. "Crowd counting via perspective-guided fractional-dilation convolution". In: *IEEE Transactions on Multimedia* 24 (2021), pp. 2633–2647.

[98] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma. "Single-Image Crowd Counting via Multi-Column Convolutional Neural Network". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 589–597.

[99] W. Liu, M. Salzmann, and P. Fua. "Context-aware crowd counting". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 5099–5108.

[100] S. Yang, W. Guo, and Y. Ren. "CrowdFormer: An Overlap Patching Vision Transformer for Top-Down Crowd Counting". In: *International Joint Conference on Artificial Intelligence*. 2022.

[101] Y. Ranasinghe, N. G. Nair, W. G. C. Bandara, and V. M. Patel. "CrowdDiff: Multi-Hypothesis Crowd Density Estimation Using Diffusion Models". In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 12809–12819. DOI: 10.1109/CVPR52733.2024.01217.

[102] D. Liang, J. Xie, Z. Zou, X. Ye, W. Xu, and X. Bai. "CrowdCLIP: Unsupervised Crowd Counting via Vision-Language Model". In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, 2023, pp. 2893–2903.

[103] X. Liu, J. van de Weijer, and A. D. Bagdanov. "Leveraging Unlabeled Data for Crowd Counting by Learning to Rank". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), pp. 7661–7669.

[104] Y. Lei, Y. Liu, P. Zhang, and L. Liu. "Towards using count-level weak supervision for crowd counting". In: *Pattern Recognition* 109 (2021), p. 107616. ISSN: 0031-3203.

[105] Y. Liu, L. Liu, P. Wang, P. Zhang, and Y. Lei. "Semi-supervised Crowd Counting via Self-training on Surrogate Tasks". In: *Computer Vision – ECCV 2020*. Ed. by A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm. Cham: Springer International Publishing, 2020, pp. 242–259. ISBN: 978-3-030-58555-6.

[106] V. A. Sindagi, R. Yasarla, D. S. Babu, R. V. Babu, and V. M. Patel. "Learning to Count in the Crowd from Limited Labeled Data". In: *Computer Vision – ECCV 2020*. Ed. by A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm. Cham: Springer International Publishing, 2020, pp. 212–229. ISBN: 978-3-030-58621-8.

[107] Z. Zhao, M. Shi, X. Zhao, and L. Li. "Active Crowd Counting with Limited Supervision". In: *Computer Vision – ECCV 2020*. Ed. by A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm. Cham: Springer International Publishing, 2020, pp. 565–581. ISBN: 978-3-030-58565-5.

[108] Y. Xu, Z. Zhong, D. Lian, J. Li, Z. Li, X. Xu, and S. Gao. "Crowd Counting With Partial Annotations in an Image". In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, pp. 15550–15559.

[109] S. Gong, S. Zhang, J. Yang, D. Dai, and B. Schiele. "Bi-Level Alignment for Cross-Domain Crowd Counting". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 7542–7550.

[110] Z. Xiong, L. Chai, W. Liu, Y. Liu, S. Ren, and S. He. "Glance to count: Learning to rank with anchors for weakly-supervised crowd counting". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2024, pp. 343–352.

[111] M. Litrico, S. Battiato, S. A. Tsaftaris, and M. V. Giuffrida. "Semi-Supervised Domain Adaptation for Holistic Counting under Label Gap". In: *Journal of Imaging* 7.10 (2021). ISSN: 2313-433X.

[112] Y. Yang, G. Li, Z. Wu, L. Su, Q. Huang, and N. Sebe. "Weakly-Supervised Crowd Counting Learns from Sorting Rather Than Locations". In: *Computer Vision – ECCV 2020*. Ed. by A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm. Cham: Springer International Publishing, 2020, pp. 1–17. ISBN: 978-3-030-58598-3.

[113] S. Kumagai, K. Hotta, and T. Kurita. "Mixture of counting CNNs: Adaptive integration of CNNs specialized to specific appearance for crowd counting". In: *arXiv preprint arXiv:1703.09393* (2017).

[114] D. Liang, X. Chen, W. Xu, Y. Zhou, and X. Bai. "Transcrowd: weakly-supervised crowd counting with transformers". In: *Science China Information Sciences* 65.6 (2022), p. 160104.

[115] J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. *Dataset Shift in Machine Learning*. Cambridge, MA (USA): The MIT Press, Dec. 2008. ISBN: 9780262255103. DOI: 10.7551/mitpress/9780262170055.001.0001.

[116] A. Gretton, A. Smola, J. Huang, M. Schmittfull, K. Borgwardt, and B. Schölkopf. "Covariate shift and local learning by distribution matching". In: *Dataset Shift in Machine Learning*. Cambridge, MA, USA: MIT Press, 2009, pp. 131–160.

[117] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada. "Maximum Classifier Discrepancy for Unsupervised Domain Adaptation". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2017), pp. 3723–3732. URL: https://api.semanticscholar.org/CorpusID:4619542.

[118] R. Xu, G. Li, J. Yang, and L. Lin. "Larger Norm More Transferable: An Adaptive Feature Norm Approach for Unsupervised Domain Adaptation". In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. Los Alamitos, CA, USA: IEEE Computer Society, 2019, pp. 1426–1435. DOI: 10.1109/ICCV.2019.00151. URL: https://doi.ieeecomputersociety.org/10.1109/ICCV.2019.00151.

[119] A. Sharma, T. Kalluri, and M. Chandraker. "Instance Level Affinity-Based Transfer for Unsupervised Domain Adaptation". In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 5357–5367. URL: https://api.semanticscholar.org/CorpusID:233025433.

[120] M. Mancini, L. Porzi, S. R. Buló, B. Caputo, and E. Ricci. "Inferring Latent Domains for Unsupervised Deep Domain Adaptation". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 43.2 (2021), 485–498. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2019.2933829. URL: https://doi.org/10.1109/TPAMI.2019.2933829.

[121] W. Li, Z. Xu, D. Xu, D. Dai, and L. Van Gool. "Domain Generalization and Adaptation Using Low Rank Exemplar SVMs". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.5 (2018), pp. 1114–1127. DOI: 10.1109/TPAMI.2017.2704624.

[122] Q. Zhou, Q. Gu, J. Pang, X. Lu, and L. Ma. "Self-Adversarial Disentangling for Specific Domain Adaptation". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 45.7 (2023), 8954–8968. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2023.3238727. URL: https://doi.org/10.1109/TPAMI.2023.3238727.

[123] P. P. Busto and J. Gall. "Open Set Domain Adaptation". In: *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 754–763. DOI: 10.1109/ICCV.2017.88.

[124] E. Arazo, D. Ortego, P. Albert, N. E. O'Connor, and K. McGuinness. "Pseudo-Labeling and Confirmation Bias in Deep Semi-Supervised Learning". In: *2020 International Joint Conference on Neural Networks (IJCNN)* (2019), pp. 1–8. URL: https://api.semanticscholar.org/CorpusID:199501839.

[125] P. Morerio, R. Volpi, R. Ragonesi, and V. Murino. "Generative Pseudo-label Refinement for Unsupervised Domain Adaptation". In: *Winter Conference on Applications of Computer Vision (WACV)*. 2020.

[126] Y. Kim, J. Yun, H. Shon, and J. Kim. "Joint Negative and Positive Learning for Noisy Labels". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 9442–9451.

[127] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. "Adapting Visual Category Models to New Domains". In: *Computer Vision – ECCV 2010*. Ed. by K. Daniilidis, P. Maragos, and N. Paragios. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 213–226. ISBN: 978-3-642-15561-1.

[128] H. Venkateswara, J. Eusebio, S. Chakraborty, and S. Panchanathan. "Deep Hashing Network for Unsupervised Domain Adaptation". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, 2017, pp. 5385–5394. DOI: 10.1109/CVPR.2017.572. URL: https://doi.ieeecomputersociety.org/10.1109/CVPR.2017.572.

[129] H. B. Mitchell and P. A. Schaefer. "A "soft" K-nearest neighbor voting scheme". In: *International Journal of Intelligent Systems* 16 (2001).

[130] X. Peng, B. Usman, N. Kaushik, J. Hoffman, D. Wang, and K. Saenko. "VisDA: The Visual Domain Adaptation Challenge". In: *CoRR* abs/1710.06924 (2017). arXiv: 1710.06924. URL: http://arxiv.org/abs/1710.06924.

[131] W. Li, J. Liu, B. Han, and Y. Yuan. "Adjustment and Alignment for Unbiased Open Set Domain Adaptation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 24110–24119.

[132] K. He, X. Zhang, S. Ren, and J. Sun. "Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[133] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. "ImageNet: A large-scale hierarchical image database". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

[134] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[135] Y. Luo, Z. Wang, Z. Huang, and M. Baktashmotlagh. "Progressive graph learning for open-set domain adaptation". In: *Proceedings of the 37th International Conference on Machine Learning*. ICML'20. Virtual: JMLR.org, 2020.

[136] Z. Fang, J. Lu, F. Liu, J. Xuan, and G. Zhang. "Open Set Domain Adaptation: Theoretical Bound and Algorithm". In: *IEEE Transactions on Neural Networks and Learning Systems* 32.10 (2021), pp. 4309–4322. DOI: 10.1109/TNNLS.2020.3017213.

[137] L. Chen, Y. Lou, J. He, T. Bai, and M. Deng. "Geometric Anchor Correspondence Mining with Uncertainty Modeling for Universal Domain Adaptation". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 16113–16122. DOI: 10.1109/CVPR52688.2022.01566.

[138] J. Liang, D. Hu, J. Feng, and R. He. "UMAD: Universal Model Adaptation under Domain and Category Shift". In: *ArXiv* abs/2112.08553 (2021). URL: https://api.semanticscholar.org/CorpusID:245218822.

[139] K. You, M. Long, Z. Cao, J. Wang, and M. I. Jordan. "Universal Domain Adaptation". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 2715–2724. DOI: 10.1109/CVPR.2019.00283.

[140] M. Yang, L. Wang, C. Deng, and H. Zhang. "Bootstrap Your Own Prior: Towards Distribution-Agnostic Novel Class Discovery". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 3459–3468.

[141] E. Fini, E. Sangineto, S. Lathuilière, Z. Zhong, M. Nabi, and E. Ricci. "A Unified Objective for Novel Class Discovery". In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), pp. 9264–9272. URL: https://api.semanticscholar.org/CorpusID:237213460.

[142] M. Long, Z. Cao, J. Wang, and M. I. Jordan. "Conditional adversarial domain adaptation". In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. NIPS'18. Montréal, Canada: Curran Associates Inc., 2018, 1647–1657.

[143] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. van der Maaten. "Exploring the Limits of Weakly Supervised Pretraining". In: *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part II*. Munich, Germany: Springer-Verlag, 2018, 185–201. ISBN: 978-3-030-01215-1. DOI: 10.1007/978-3-030-01216-8_12. URL: https://doi.org/10.1007/978-3-030-01216-8_12.

[144] J. Li, D. Li, S. Savarese, and S. Hoi. "BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models". In: *Proceedings of the 40th International Conference on Machine Learning*. ICML'23. Honolulu, Hawaii, USA: JMLR.org, 2023.

[145] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *North American Chapter of the Association for Computational Linguistics*. 2019. URL: https://api.semanticscholar.org/CorpusID:52967399.

[146] V. Sanh, L. Debut, J. Chaumond, and T. Wolf. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter". In: *ArXiv* abs/1910.01108 (2019). URL: https://api.semanticscholar.org/CorpusID:203626972.

[147] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan. "Supervised contrastive learning". In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS '20. Vancouver, BC, Canada: Curran Associates Inc., 2020. ISBN: 9781713829546.

[148] T. Kalluri, A. Sharma, and M. Chandraker. "MemSAC: Memory Augmented Sample Consistency for Large Scale Domain Adaptation". In: *ArXiv* abs/2207.12389 (2022). URL: https://api.semanticscholar.org/CorpusID:251040159.

[149] Y. Zhang, T. Liu, M. Long, and M. I. Jordan. "Bridging Theory and Algorithm for Domain Adaptation". In: *International Conference on Machine Learning*. 2019. URL: https://api.semanticscholar.org/CorpusID:118638514.

[150] Z. Du, J. Li, H. Su, L. Zhu, and K. Lu. "Cross-Domain Gradient Discrepancy Minimization for Unsupervised Domain Adaptation". In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 3936–3945. URL: https://api.semanticscholar.org/CorpusID:235367640.

[151] T. Sun, C. Lu, T. Zhang, and H. Ling. "Safe Self-Refinement for Transformer-based Domain Adaptation". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, 2022, pp. 7181–7190. DOI: 10.1109/CVPR52688.2022.00705. URL: https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.00705.

[152] V. Lempitsky and A. Zisserman. "Learning to count objects in images". In: *Advances in neural information processing systems* 23 (2010).

[153] S. Jaiswal, A. S. Gadgil, A. M. Kaslikar, and K. S. Kothari. "Comprehensive Study of Various Methods for Estimating Crowd Density". In: *Innovations and Advances in Cognitive Systems*. Ed. by S. D. P. Ragavendiran,

V. D. Pavaloaia, M. S. Mekala, and A. S. Cabezuelo. Cham: Springer Nature Switzerland, 2024, pp. 383–400. ISBN: 978-3-031-69201-7.

[154] Z. Ma, X. Wei, X. Hong, and Y. Gong. "Bayesian Loss for Crowd Count Estimation With Point Supervision". In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), pp. 6141–6150.

[155] H. Lin, Z. Ma, R. Ji, Y. Wang, and X. Hong. "Boosting Crowd Counting via Multifaceted Attention". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 19596–19605.

[156] W. Liu, N. Durasov, and P. Fua. "Leveraging Self-Supervision for Cross-Domain Crowd Counting". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 5331–5342.

[157] Wang et al. "Joint CNN and Transformer Network via weakly supervised Learning for efficient crowd counting". In: *arXiv:2203.06388* (2022).

[158] H. Idrees, M. Tayyab, K. Athrey, D. Zhang, S. A. Al-Maadeed, N. M. Rajpoot, and M. Shah. "Composition Loss for Counting, Density Map Estimation and Localization in Dense Crowds". In: *European Conference on Computer Vision*. 2018.

[159] V. A. Sindagi, R. Yasarla, and V. M. Patel. "JHU-CROWD++: Large-Scale Crowd Counting Dataset and A Benchmark Method". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44 (2020), pp. 2594–2609.

[160] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu. "Global Contrast Based Salient Region Detection". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.3 (2015), pp. 569–582.

[161] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan. "Amulet: Aggregating Multi-level Convolutional Features for Salient Object Detection". In: *2017 IEEE International Conference on Computer Vision (ICCV)* (2017), pp. 202–211.

[162] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand. "BASNet: Boundary-Aware Salient Object Detection". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 7471–7481.

[163] A. van den Oord, Y. Li, and O. Vinyals. "Representation Learning with Contrastive Predictive Coding". In: *ArXiv* abs/1807.03748 (2018).

[164] J. Wan and A. Chan. "Modeling Noisy Annotations for Crowd Counting". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 3386–3396.

[165] J. Wan, Z. Liu, and A. B. Chan. "A Generalized Loss Function for Crowd Counting and Localization". In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 1974–1983.

[166] Q. Wang, J. Gao, W. Lin, and X. Li. "NWPU-Crowd: A Large-Scale Benchmark for Crowd Counting and Localization". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020). DOI: 10.1109/TPAMI.2020.3013269.

[167] Q. Xie, Z. Dai, E. Hovy, M.-T. Luong, and Q. V. Le. "Unsupervised Data Augmentation for Consistency Training". In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS'20. Vancouver, BC, Canada: Curran Associates Inc., 2020. ISBN: 9781713829546.

[168] A. Tarvainen and H. Valpola. "Mean Teachers Are Better Role Models: Weight-Averaged Consistency Targets Improve Semi-Supervised Deep Learning Results". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, 1195–1204. ISBN: 9781510860964.

[169] V. Verma, A. Lamb, J. Kannala, Y. Bengio, and D. Lopez-Paz. "Interpolation Consistency Training for Semi-supervised Learning". In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, July 2019, pp. 3635–3641.

[170] C. LI, X. Hu, S. Abousamra, and C. Chen. "Calibrating Uncertainty for Semi-Supervised Crowd Counting". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023, pp. 16731–16741.

[171] T. Han, J. Gao, Y. Yuan, and Q. Wang. "Focus on Semantic Consistency for Cross-Domain Crowd Understanding". In: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2020), pp. 1848–1852.

[172] Han and et al. "STEERER: Resolving Scale Variations for Counting and Localization via Selective Inheritance Learning". In: *ICCV*. 2023.

[173] J. Wan and et al. "Robust Unsupervised Crowd Counting and Localization with Adaptive Resolution SAM". In: *ArXiv* abs/2402.17514 (2024).

[174] Zhou et al. "Texture-guided Saliency Distilling for Unsupervised Salient Object Detection". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023).

[175] G. Pombo, R. Gray, M. J. Cardoso, S. Ourselin, G. Rees, J. Ashburner, and P. Nachev. "Equitable modelling of brain imaging by counterfactual augmentation with morphologically constrained 3D deep generative models". In: *Medical Image Analysis* 84 (2023), p. 102723. ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2022.102723. URL: https://www.sciencedirect.com/science/article/pii/S1361841522003516.

[176] D. Ravi, D. C. Alexander, N. P. Oxtoby, and A. D. N. Initiative. "Degenerative adversarial neuroimage nets: generating images that mimic disease progression". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 164–172.

[177] A. L. Young, N. P. Oxtoby, S. Garbarino, N. C. Fox, F. Barkhof, J. M. Schott, and D. C. Alexander. "Data-driven modelling of neurodegenerative disease progression: thinking outside the black box". In: *Nature Reviews Neuroscience* 25.2 (Jan. 2024), 111–130. ISSN: 1471-0048. DOI: 10.1038/s41583-023-00779-6. URL: http://dx.doi.org/10.1038/s41583-023-00779-6.

[178] C. Bowles, R. Gunn, A. Hammers, and D. Rueckert. "Modelling the progression of Alzheimer's disease in MRI using generative adversarial networks". In: *Medical Imaging 2018: Image Processing* 10574 (2018), 105741K.

[179] C. R. Jack Jr, D. A. Bennett, K. Blennow, M. C. Carrillo, B. Dunn, S. B. Haeberlein, D. M. Holtzman, W. Jagust, F. Jessen, J. Karlawish, et al. "NIA-AA research framework: toward a biological definition of Alzheimer's disease". In: *Alzheimer's & Dementia* 14.4 (2018), pp. 535–562.

[180] D. Liu, M. Kelly, and P. Gong. "A spatial–temporal approach to monitoring forest disease spread using multi-temporal high spatial resolution imagery". In: *Remote sensing of environment* 101.2 (2006), pp. 167–180.

[181] J. S. Yoon, C. Zhang, H.-I. Suk, J. Guo, and X. Li. "SADM: Sequence-Aware Diffusion Model for Longitudinal Medical Image Generation". In: *Information Processing in Medical Imaging*. 2022. URL: https://api.semanticscholar.org/CorpusID:254823541.

[182] B. Kim and J. C. Ye. "Diffusion Deformable Model for 4D Temporal Medical Image Generation". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part I*. Singapore, Singapore: Springer-Verlag, 2022, 539–548. ISBN: 978-3-031-16430-9.

[183] B. Kim, I. Han, and J. C. Ye. "DiffuseMorph: Unsupervised Deformable Image Registration Using Diffusion Model". In: *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI*. Tel Aviv, Israel: Springer-Verlag, 2022, 347–364. ISBN: 978-3-031-19820-5.

[184] D. L. Dickstein, D. Kabaso, A. B. Rocher, J. I. Luebke, S. L. Wearne, and P. R. Hof. "Changes in the structural complexity of the aged brain". In: *Aging cell* 6.3 (2007), pp. 275–284.

[185] T. Xia, A. Chartsias, and S. A. Tsaftaris. "Consistent Brain Ageing Synthesis". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. Ed. by D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, and A. Khan. Cham: Springer International Publishing, 2019, pp. 750–758. ISBN: 978-3-030-32251-9.

[186] T. Xia, A. Chartsias, C. Wang, and S. A. Tsaftaris. "Learning to synthesise the ageing brain without longitudinal data". In: *Medical image analysis* 73

(2019), p. 102169. URL: https://api.semanticscholar.org/CorpusID:208637395.

[187] P. J. LaMontagne, T. L. Benzinger, J. C. Morris, S. Keefe, R. Hornbeck, C. Xiong, E. Grant, J. Hassenstab, K. Moulder, A. G. Vlassenko, et al. "OASIS-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimer disease". In: *MedRxiv* (2019), pp. 2019–12.

[188] J. Ho, A. Jain, and P. Abbeel. "Denoising diffusion probabilistic models". In: *Advances in neural information processing systems* 33 (2020), pp. 6840–6851.

[189] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy. "ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks". In: *Computer Vision – ECCV 2018 Workshops*. Ed. by L. Leal-Taixé and S. Roth. Cham: Springer International Publishing, 2019, pp. 63–79.

[190] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc., 2017.

[191] B. A. Jónsson, G. Bjornsdottir, T. Thorgeirsson, L. M. Ellingsen, G. B. Walters, D. Gudbjartsson, H. Stefansson, K. Stefansson, and M. Ulfarsson. "Brain age prediction using deep learning uncovers associated sequence variants". In: *Nature communications* 10.1 (2019), p. 5409.

[192] M. Jenkinson, C. F. Beckmann, T. E. Behrens, M. W. Woolrich, and S. M. Smith. "FSL". In: *Neuroimage* 62.2 (2012), pp. 782–790.

[193] H. Li, Y. Yang, M. Chang, S. Chen, H. Feng, Z. Xu, Q. Li, and Y. Chen. "Srdiff: Single image super-resolution with diffusion probabilistic models". In: *Neurocomputing* 479 (2022), pp. 47–59.

[194] L. Puglisi, D. C. Alexander, and D. Ravì. "Enhancing Spatiotemporal Disease Progression Models via Latent Diffusion and Prior Knowledge". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*. Ed. by M. G. Linguraru, Q. Dou, A. Feragen, S. Giannarou, B. Glocker, K. Lekadir, and J. A. Schnabel. Cham: Springer Nature Switzerland, 2024, pp. 173–183. ISBN: 978-3-031-72069-7.

[195] T. Xia, A. Chartsias, C. Wang, and S. A. Tsaftaris. "Learning to synthesise the ageing brain without longitudinal data". In: *Medical Image Analysis* 73 (2021), p. 102169. ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2021.102169. URL: https://www.sciencedirect.com/science/article/pii/S1361841521002152.

[196] J. H. Cole, S. J. Ritchie, M. E. Bastin, V. Hernández, S Muñoz Maniega, N. Royle, J. Corley, A. Pattie, S. E. Harris, Q. Zhang, et al. "Brain age predicts mortality". In: *Molecular psychiatry* 23.5 (2018), pp. 1385–1392.

[197] A. Rondinella, F. Guarnera, O. Giudice, A. Ortis, G. Russo, E. Crispino, F. Pappalardo, and S. Battiato. "Enhancing multiple sclerosis lesion segmentation in multimodal MRI scans with diffusion models". In: *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2023, pp. 3733–3740.

[198] R. He, G. Ang, and D. Tward. *Individualized multi-horizon MRI trajectory prediction for Alzheimer's Disease*. 2024. arXiv: 2408.02018 [cs.CV]. URL: https://arxiv.org/abs/2408.02018.

[199] M. Wilms, J. J. Bannister, P. Mouches, M. E. MacDonald, D. Rajashekar, S. Langner, and N. D. Forkert. "Invertible modeling of bidirectional relationships in neuroimaging with normalizing flows: Application to brain aging". en. In: *IEEE Trans. Med. Imaging* 41.9 (Sept. 2022), pp. 2331–2347.

[200] D. L. Beekly, E. M. Ramos, W. W. Lee, W. D. Deitrich, M. E. Jacka, J. Wu, J. L. Hubbard, T. D. Koepsell, J. C. Morris, W. A. Kukull, and NIA Alzheimer's Disease Centers. "The National Alzheimer's Coordinating Center (NACC) database: the Uniform Data Set". en. In: *Alzheimer Dis Assoc Disord* 21.3 (July 2007), pp. 249–258.

[201] B. Billot, D. N. Greve, O. Puonti, A. Thielscher, K. Van Leemput, B. Fischl, A. V. Dalca, and J. E. Iglesias. "SynthSeg: Segmentation of brain MRI scans of any contrast and resolution without retraining". In: *Medical Image Analysis* 86 (2023), p. 102789. ISSN: 1361-8415. DOI: 10.1016/j.media.2023.102789.

[202] L. Puglisi, D. C. Alexander, and D. Ravi. "Brain Latent Progression: Individual-based Spatiotemporal Disease Progression on 3D Brain MRIs via Latent Diffusion". In: *arXiv preprint arXiv:2502.08560* (2025).