國立臺灣大學電機資訊學院資訊網路與多媒體研究所

碩士論文

Graduate Institute of Networking and Multimedia

College of Electrical Engineering and Computer Science

National Taiwan University

Master's Thesis

循序漸進：用分階段訓練提升半監督學習方法在領域自適應姿態估計中的適用性

Phase2Phase: Enhancing the Applicability of Semi-Supervised Learning Methods in Domain Adaptive Pose Estimation through Phased Training

蕭昀豪

Yun-Hao Hsiao

指導教授: 許永真博士 & 鄭文皇博士

Advisor: Jane Yung-Jen Hsu, Ph.D. & Wen-Huang Cheng, Ph.D.

中華民國 113 年 10 月

October, 2024

doi:10.6342/NTU202404439

國立臺灣大學碩士學位論文
口試委員會審定書
MASTER'S THESIS ACCEPTANCE CERTIFICATE
NATIONAL TAIWAN UNIVERSITY

循序漸進: 用分階段訓練提升半監督學習方法在領域自適應姿態估計中的適用性

Phase2Phase: Enhancing the Applicability of Semi-Supervised Learning Methods in Domain Adaptive Pose Estimation through Phased Training

本論文係蕭昀豪（學號 R10944033）在國立臺灣大學資訊網路與多媒體研究所完成之碩士學位論文，於民國 113 年 3 月 21 日承下列考試委員審查通過及口試及格，特此證明。

The undersigned, appointed by the Graduate Institute of Networking and Multimedia on 21 March 2024 have examined a Master's Thesis entitled above presented by YUN-HAO HSIAO (student ID: R10944033) candidate and hereby certify that it is worthy of acceptance.

口試委員 Oral examination committee:

許永真		
（指導教授 Advisor）		
陳駿丞	鄭文皇	楊智淵

系（所）主管 Director: 鄭卜壬

doi:10.6342/NTU202404439

Acknowledgements

感謝所有在我學術生涯幫助過我的貴人。首先，我要感謝永真老師總是提供寶貴的建議協助我在研究之路上不斷前進。老師不僅在學術上引領我，也常在人生路上帶給我啟迪。當我想爭取實習機會或海外交流等，老師的分析與建議時常成為我做重大決定的依據。在學術內外，老師的指導都影響深刻，我銘謝在心。

碩士生活中，能有一群好夥伴同甘共苦至關重要。我很感謝那些在我實驗成功時分享喜悅的夥伴，也很珍惜當我面對低潮時願意接住我的手。因為你們，我得以輕鬆釋放壓力、記起歡笑快樂，我很感謝能與你們相遇相知。

除了校內的支持，實習也是我研究進步的一大助力。其中我特別感謝我在玩美移動公司實習時的主管 Detta，在一年的實習中給予了我十足的耐心和充份的資源，我對您的支持深表謝意。

在撰寫論文的過程中，我特別感謝楊智淵學長的寶貴建議。學長始終以嚴謹的態度審視我的論文，並在關鍵時刻提供精確的指導。透過與學長的反覆討論，我得以使論文去蕪存菁。學長治學嚴謹的態度讓我深感敬佩，也由衷感激。

在碩士生涯的最後一程，我來到了德國成為交換學生，在這裡我對於自己一路走來是基於多少幫助有了更深刻的理解。這篇論文的出現匯聚了太多人的好意與助力，這些恩澤我銘感五內。我的碩士學術生涯將要結束，但我會活用學習到的經驗與知識，前進下一階段，持續追求卓越。

doi:10.6342/NTU202404439

doi:10.6342/NTU202404439

摘要

近幾年，領域自適應姿態估計 (DAPE) 受到越來越多的關注。在解決該任務的方法中，半監督式學習 (SSL) 的方法因為與無監督領域自適應 (UDA) 有相似的目標而被廣泛使用。然而，雖然 SSL 和 UDA 同樣都旨在用未標記的數據增強已在標記數據上訓練的模型，兩個任務對於資料分布的不同期望依然讓 SSL 方法無法完美地契合於 UDA 中。有鑑於此，我們提出了 Phase2Phase，這一策略整合了三種方法：Adaptive Mean Teacher、T-VAT-based UniMatch，以及 Mixup Augmentation。

傳統的 Mean Teacher 方法通常採用較大的平滑係數，但在 UDA 任務中，較大的平滑細數容易阻礙教師模型迅速達到學生模型的性能。為了克服這一問題，我們提出 Adaptive Mean Teacher。藉由在領域自適應初期加入緩衝階段，並在該緩衝階段使用較低的平滑係數，我們使教師模型在保持 Mean Teacher 穩定性特點的同時，使其在初期就能快速追上學生模型的表現。

Mean Teacher 已經被證實在 UDA 的情境下有效，而除了如 Mean Teacher 這樣以模型為中心出發，用不同模型生成兩個輸出以進行一致性調節的方法，SSL 中同樣存在像 FixMatch 和 UniMatch 這樣在輸入和特徵層面引入干擾項的方法。儘管這些策略理論上能相輔相成，但實驗卻顯示生硬地結合兩個方法會導致效能下降。為此，我們提出的 Phase2Phase 在採用傳統 Mean Teacher 的訓練階段後又引

doi:10.6342/NTU202404439

入了一個額外階段，在該階段利用 T-VAT-based UniMatch 讓模型校能進一步提升。最後，我們還在訓練全程都引入 Mixup 技術，以提升模型的穩健性和整體性能。

實驗結果顯示，我們提出的方法確實有益於提升 DAPE 的表現。值得注意的是，Phase2Phase 超越了能使用標記資料的其他 DAPE 方法，同時他的表現也逼近當前最先進、在不使用標記資料的情況下依然表現優異 DAPE 方法，SFDAHPE。這凸顯了 Phase2Phase 在實際情境中的實用性。

關鍵字：領域自適應、姿態估計、關鍵點檢測、半監督式學習

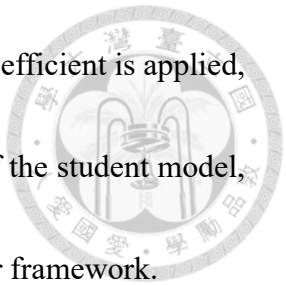
doi:10.6342/NTU202404439

Abstract

Domain Adaptive Pose Estimation (DAPE) has received increasing attention recently. Many methods commonly used in semi-supervised learning (SSL) tasks are now applied to DAPE tasks because SSL shares a similar goal with unsupervised domain adaptation (UDA). Although both SSL and UDA aim to enhance a model trained on labeled data with unlabeled data, the differences in data distribution expectations present a challenge, preventing these SSL methods from seamlessly adapting to DAPE. With an awareness of inconsistent expectations in the data, we introduce Phase2Phase, a novel approach that integrates three core strategies: Adaptive Mean Teacher, T-VAT-based UniMatch, and Mixup augmentation.

In traditional Mean Teacher methods, a significant smoothing coefficient is typically employed. However, in UDA tasks, a large smoothing coefficient can hinder the teacher model from quickly achieving the performance of the student model. To overcome this issue, we propose the Adaptive Mean Teacher. By introducing a ramp-up phase during

doi:10.6342/NTU202404439



the initial stages of domain adaptation, where a reduced smoothing coefficient is applied, we enable the teacher model to rapidly align with the performance of the student model, while preserving the stability inherent in the traditional Mean Teacher framework.

Mean Teacher has been proven effective in UDA, and in addition to model-centered approaches like Mean Teacher, which generate two outputs from different models for consistency regularization, SSL also includes methods like FixMatch and UniMatch that introduce disturbances at the input and feature levels. Although these strategies theoretically complement each other, their direct combination can lead to degraded effectiveness. Our thesis introduces an additional phase for T-VAT-based UniMatch to facilitate integration. Finally, we incorporate Mixup augmentation to boost the model's robustness, further elevating the overall performance.

The experimental results show that all the proposed methods enhance the performance of DAPE. Notably, Phase2Phase surpasses previous source-dependent DAPE approaches and achieves comparable results to the current state of the art in source-free DAPE, SFDAHPE. This underscores Phase2Phase' s practical effectiveness in the real-world scenarios.

Keywords: Domain Adaptation, Pose Estimation, Keypoint Detection, Semi-supervised Learning

doi:10.6342/NTU202404439

Contents

doi:10.6342/NTU202404439

doi:10.6342/NTU202404439

doi:10.6342/NTU202404439

doi:10.6342/NTU202404439

List of Figures

doi:10.6342/NTU202404439

doi:10.6342/NTU202404439

List of Tables

doi:10.6342/NTU202404439

doi:10.6342/NTU202404439

List of Algorithms

doi:10.6342/NTU202404439

doi:10.6342/NTU202404439

Symbols

Data Representation

S	Source domain dataset
T	Target domain dataset
x	Input image in S or T
n	Number of data in S or T
y	Label, only accessible in S in DAPE
H	Height
W	Width
K	Number of the keypoints
H^{gd}	The prediction heatmap that serve as the guide in consistency regularization
H^{flw}	The prediction heatmap that serve as the follower in consistency regularization

Model

η	Smoothing coefficient (EMA rate) in EMA

doi:10.6342/NTU202404439

Loss

L	Loss
L_{sup}	Supervised loss
L_{con}	Consistency regularization
L_{mix}	Mixup loss

Acronym & Abbreviation

Phase2Phase	Our proposed method
SSL	Semi-Supervised Learning
UDA	Unsupervised Domain Adaptation
DAPE	Domain Adaptive Pose Estimation
EMA	Exponential Moving Average
TGP	Teacher Guidance Phase
ABEP	Augmentation-Based Enhancement Phase
MSE	Mean Square Error
RHD	Rendered Hand Pose Dataset
H3D	Hand-3D-Studio Dataset
VAT	Virtual Adversarial Training
T-VAT	Virtual Adversarial Training from the Teacher Model
PCK	Percentage of Correct Keypoint
MCP	Metacarpophalangeal
PIP	Proximal Interphalangeal

doi:10.6342/NTU202404439

DIP	Distal Interphalangeal
Fin	Fingertip
SF	Source-free setting
MMD	Maximum Mean Discrepancy

doi:10.6342/NTU202404439

doi:10.6342/NTU202404439

Chapter 1 Introduction

Our research aims to explore the challenges of applying semi-supervised learning methods (SSL) in unsupervised domain adaptation (UDA) and propose solutions. We focus specifically on pose estimation, with an emphasis on the hand pose estimation.

In this chapter, Section 1.1 will explore the background of domain adaptive pose estimation (DAPE) for hands. Subsequently, Section 1.2 will discuss the motivations behind this research. Finally, Section 1.3 will outline the structure of the entire thesis.

1.1 Background

A substantial amount of annotated hand data is crucial in developing a robust pose estimation (keypoint detection) model. However, the resources required are often considerable. Specifically, annotating a single-hand image involves a person pinpointing the locations of 21 keypoints, which typically takes about 1 minute per image. Consequently, assuming each image takes approximately one minute for an individual to annotate, compiling a dataset comparable to H3D [14], which contains 22,000 images, would necessitate roughly 2.3 months of cumulative labor time.

In such a situation, synthetic data is becoming more and more popular. Synthetic data

doi:10.6342/NTU202404439

is more accessible and more flexible than real-world data. The development of computer graphics and game engines allows us to easily acquire images from different perspectives and skin tones, facilitating the collection of varied data to develop models capable of handling various scenarios. Specifically, if the dataset includes more images with exaggerated angles, like fingers facing the camera, the model is less likely to make erroneous predictions due to unusual angles. Such the capability to tackle abnormal situations can be advantageous in contexts like video recognition involving dynamic human activities.

Nonetheless, the disparity in distribution between synthetic and real-world data casts a shadow on overall performance. An effective domain adaptation strategy is imperative in response to this challenge. Domain adaptation enables us to train the model in one domain and broaden its expertise to encompass another, corresponding to the issues we encounter.

Various UDA techniques are rooted in SSL methodologies, such as consistency regularization and pseudo-labeling. SSL and UDA are two typical strategies for utilizing unlabeled data. While SSL leverages unlabeled data that is within the same distribution of labeled data, UDA focuses on transferring knowledge from one data distribution to another. Given similar goals and methods, Zhang et al. highlight SSL and UDA are intricately linked [37]. Their findings suggest that SSL is a particular case of domain adaptation when the target support covers source support. Given the similarities in their objectives that enhance the performance with unlabeled data, numerous studies in domain adaptation have successfully incorporated semi-supervised techniques. Although SSL has been widely applied in UDA tasks and has achieved outstanding performance, the methods developed for SSL might only partially align with UDA application scenarios. This misalignment stems from the inherent differences in the initial contexts for which developers

doi:10.6342/NTU202404439

originally intended SSL and UDA.

1.2 Motivation

Despite the advancement in the DAPE, little research focuses on the necessary modifications for seamlessly integrating SSL methods into UDA scenarios. Although existing methods has significantly enhanced UDA performance using approaches derived from SSL, we observe that applying these SSL techniques to UDA can yield results that vary from their typical use in SSL tasks. For instance, when we apply Mean Teacher in UDA scenarios, the teacher model can underperform and be less stable than the student model. Motivated by these observations, I delve into this issue and explore potential solutions to bridge this gap.

Furthermore, I have observed that in UDA scenarios, consistency regularization that features weak-to-strong augmentation fails to achieve the desired outcome. Theoretically, both the Mean Teacher paradigm and weak-to-strong consistency regularization aim to steer weaker predictions towards stronger ones. Although it seems possible to merge these two methods, the outcome differs from the expectation. These unexpected results have inspired me to investigate whether there is a training approach that can effectively leverage the strengths of both methods.

1.3 Thesis Organization

The thesis comprises five chapters and an appendix. Chapter 1 introduces the background information. Chapter 2 reviews related work. Chapter 3 outlines the problem

doi:10.6342/NTU202404439

definition and elaborates on the proposed methodology. Chapter 4 reports our experiments and analysis. Chapter 5 concludes the study and discusses future work. Finally, Appendix A contains an analogous explanation of the sudden learning acceleration at the start of domain adaptation using the human learning curve.

doi:10.6342/NTU202404439

Chapter 2 Literature Review

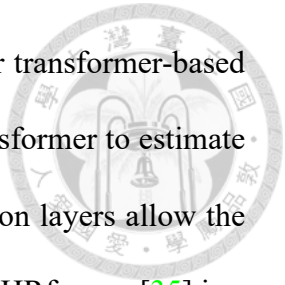
This chapter introduces the 2D pose estimation task in Section 2.1. Following this, we elaborate on the relationship between SSL and UDA in Section 2.2, where we discuss the similarities and differences between UDA and SSL methods in Subsections 2.2.1 and 2.2.2, respectively. Subsequently, Section 2.3 delves into consistency regularization, the focal point of our discussion in this thesis. Finally, we detail the evolution and current state of domain adaptive pose estimation (DAPE), our primary research focus, in Section 2.4.

2.1 2D Pose Estimation

In recent years, the research of 2D pose estimation, also called 2D keypoint detection, has become increasingly active [7, 9, 29]. Pose estimation has a wide range of applications, serving various roles and functions throughout different stages of the human lifespan [24]. The research on this topic can be categorized into two main methods: heatmap-based and regression-based.

Heatmap-based methods use heatmaps to estimate the likelihood of each keypoint on a per-pixel basis. Due to their outstanding performance, these methods currently dominate the research field. HRNet [27] maintains high-resolution representations by constantly

doi:10.6342/NTU202404439



fusing convolutions of different levels of resolution in parallel. After transformer-based architecture prospers in deep learning, some research utilizes the transformer to estimate keypoints and gain success. TransPose [34] indicates that the attention layers allow the model to efficiently and explicitly grasp global spatial dependencies. HRformer [35] improves performance by redesigning HRNet using the vision transformer (ViT) architecture. ViTPose [30] employs the ViT to encode features and utilizes a lightweight decoder to decode keypoints. ViTPose+ [31] further factorizes task-agnostic and task-specific knowledge to deal with different body pose estimation tasks.

Regression-based methods significantly differ from heatmap-based approaches, as they do not use heatmaps to represent the likelihood of keypoint positions. Regression-based methods directly map the input to the output keypoints. RLE [12] proposes an innovative and effective regression paradigm framed from a maximum likelihood perspective, marking the first time a regression-based method performs comparably to heatmap-based methods. Following this, Poseur [15] emerges as a notable work in the field, addressing drawbacks of earlier regression-based methods, such as information loss due to average pooling. Poseur also pioneers the formulation of the problem as a sequence prediction task. Simultaneously, Poseur introduces a novel perspective by treating pose estimation as a sequence prediction task and employs the Transformer architecture to solve it.

2.2 Unsupervised Domain Adaptation (UDA) and Semi-Supervised learning (SSL)

Unsupervised domain adaptation (UDA) and semi-supervised learning (SSL) are two typical strategies for using unlabeled data. These strategies enable the utilization of data

doi:10.6342/NTU202404439

whose annotations are challenging or impossible to obtain. Consequently, a considerable amount of time and resources can be saved. While UDA focuses on transferring knowledge between different data distributions, SSL enhances performance by utilizing unlabeled data that aligns closely with the distribution of labeled data. Given the similarities in their objectives that strengthen a model trained on labeled data by utilizing unlabeled data, numerous studies in UDA have successfully incorporated methods originating from SSL.

Furthermore, ongoing research investigates the interconnection between these two concepts. Zhang et al. highlight that despite differences, SSL and UDA are intricately linked given similar goals and methods. Their findings suggest that SSL is a particular case of domain adaptation when the target support covers source support [37].

In this review, I systematically categorize the methodologies that leverage unlabeled data to enhance the performance of models trained with labeled data. Specifically, I have identified two techniques with greater specificity to UDA. Additionally, five methodologies applicable to both UDA and SSL have been detailed.

2.2.1 Methods with Greater Specificity to UDA

The most significant characteristic of methods with greater specificity to UDA is their recognition of the differing data distributions between labeled and unlabeled data, accompanied by actions to mitigate this discrepancy. In my research, such methods can be categorized into two types: metric-based methods and style-transfer-based methods.

Metric-based approaches utilize distance metrics to minimize domain discrepancies. Long et al. minimizes the Maximum Mean Discrepancy (MMD) [14] for aligning dis-

doi:10.6342/NTU202404439

tinct domains. Following this advancement, the adaptation of weighted MMD [32] was proposed, targeting the previously underaddressed issue of class weight bias. Additionally, Sun et al. focus on aligning the second-order statistics between the source and target domains [25, 26].

Style-transfer-based methods adapt the aesthetic or stylistic aspects of input data from two domains during training, enhancing the model's robustness across various data distributions between the source and target domains. For instance, Ashish et al. introduced Simulated + Unsupervised (S + U) learning [22], which investigates the impact of synthetic images on gaze estimation and hand pose estimation tasks. Furthermore, in keypoint detection tasks, UniFrame [9] utilizes AdaIN [6] for domain adaptation. It adjusts the source domain data to match the style of the target domain. Simultaneously, it also transforms the target domain to reflect the style characteristics of the source.

2.2.2 Methods shared between UDA and SSL

According to Peláez Vegas et al., SSL approaches can be concluded into five categories [18]: adversarial methods, pseudo-labeling, contrastive learning, consistency regularization, and hybrid methods. These SSL methods are also applicable in UDA.

Adversarial-based methods involve constructing an adversarial mechanism to harmonize the feature representations between the source and target domains. DANN [4] confuses the feature representatives of two domains by adding a domain classifier and gradient reversal layer to the main architecture. On the other hand, RegDA [7] also designs a unique adversarial framework to achieve domain adaptation on keypoints.

Pseudo-labeling, often referred to as bootstrapping, wrapper, or self-labeled tech-

doi:10.6342/NTU202404439

niques, involve generating pseudo-labels for unlabeled images based on predictions from a model previously trained. These pseudo-labeled images are then added to the original dataset, creating an expanded set of image-label pairs. Subsequently, the model is enhanced through additional training utilizing the augmented dataset. The concept of pseudo-labeling is extensively applied in both SSL and UDA. For example, the Mixup process in MixMatch incorporates pseudo-labeling [2]. Besides, MAPS [3], a well-performed DAPE work, also utilizes pseudo-labeling as a critical component.

Contrastive learning aims to cluster similar samples while distancing them from dissimilar ones. In unlabeled scenarios, typical contrastive learning methods involve grouping augmented versions of the same samples and treating the rest as different. In DAPE, observations from RegDA indicate that errors in keypoint predictions often occur at other keypoints [7]. Based on this, SFDAHPE [19], the current state of the art, proposes the pose-specific contrastive loss. It groups predictions of the same keypoints from differently perturbed versions of the same image while distancing others, thereby achieving success.

Consistency regularization highlights the stability of the model under perturbations or domain shifts. These methods enhance the performance of the model by promoting consistent output across minor variations. They enable the model to disregard random noise and trivial discrepancies, thereby enhancing generalization in environments scarce in labeled data. Given the centrality of this topic to our study, a more comprehensive introduction is provided in Section 2.3.

doi:10.6342/NTU202404439

2.3 Consistency Regularization

Tarvainen indicates that there exist at least two approaches to achieve consistency regularization [28]. One strategy entails carefully choosing the teacher model, and the other approach involves meticulously selecting perturbations for the representations.

Regarding the approaches of carefully selecting the teacher model, Π model [10], temporal ensembling [10], and Mean Teacher [28] are representative methods. Π model utilizes the randomness of dropout and augmentation. Temporal ensembling alleviates the noise of prediction by aggregating the predictions of multiple previous network evaluations. Mean Teacher improves performance by solving a problem with temporal ensembling: slow updates caused by large datasets. In the Mean Teacher architecture, there exist two components: the student model, which learns via gradient descent, and the teacher model, which updates the student model's parameters through an exponential moving average (EMA) approach. As the weighted average of the student model over time, the teacher model is expected to be more stable and more robust to the variances than the student model.

Concerning methods that emphasize perturbation selection, Virtual Adversarial Training (VAT) [16] and weak-to-strong consistency regularization are two notable strategies.

Virtual Adversarial Training (VAT) is a technique designed to generate perturbations that guide the model towards diverging from its current state. Unlike typical adversarial training, which aims to train a robust model directly, VAT focuses on obtaining perturbations that facilitate effective training. VAT reflects the direction of deviation from the status quo, thus creating an adversarial training pattern with consistency regularization.

doi:10.6342/NTU202404439

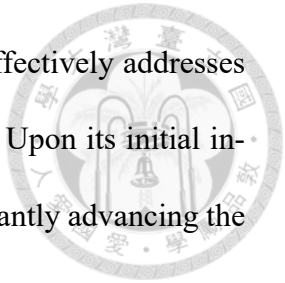
Building on VAT, PSMT [13] proposes VAT from the teacher model (T-VAT) for UDA. PSMT enhances VAT by incorporating it within the Mean Teacher framework. By leveraging the robustness of the teacher model, T-VAT is anticipated to be more effective than VAT alone.

Weak-to-strong consistency regularization evolved from Mixup-based approaches, with MixMatch [2] achieving unsupervised learning by calculating Mixup loss for both labeled and unlabeled data. Building on MixMatch, ReMixMatch [1] incorporates the concept of weak and strong augmentations. Subsequently, FixMatch [23] eliminates the Mixup concept and conducts consistency regularization between predictions obtained from weakly and strongly augmented inputs. Most recently, UniMatch [33] has enhanced the FixMatch framework by perturbing intermediate features—outputs derived prior to complete processing in deep neural networks—, thereby achieving more robust results.

2.4 Domain Adaptive Pose Estimation

Domain adaptive pose estimation (DAPE) has received increasing attention recently, particularly since the emergence of RegDA [7]. Initially, research in this area tended to concentrate on specific applications. For example, RegDA and MarsDA [8] are dedicated to human and hand pose estimation. Subsequently, CSSL [17] and UDA-Animal [11] concentrate on animal pose estimation. Based on the prior work, UniFrame [9] adopts Mean Teacher [28] and creates a versatile framework applicable across various pose estimation tasks. UniFrame indicates that the distinct focus of earlier research might stem from the unique properties of different datasets used in those benchmarks. Specifically, animal datasets often exhibit considerable variance at the input level, whereas human and hand

doi:10.6342/NTU202404439



datasets display substantial variance at the output level. UniFrame effectively addresses the domain gap from both input-level and output-level perspectives. Upon its initial introduction, UniFrame demonstrated remarkable performance, significantly advancing the state of the art.

From 2023, with growing concerns over privacy, source-free DAPE has emerged as a crucial research field. Source-free DAPE refers to the training scenario in which only a pretrained source domain model and target domain data are available. Its necessity is reflected in the services offered by cloud platforms, including Google Cloud Platform (GCP) and Azure. Many of these services provide models that have been pretrained but restrict access to the source data. In certain cases, an entity may refuse to disclose source data because they possess only usage rights, not distribution rights. In other cases, the reluctance to reveal source data may stem from concerns about protecting trade secrets or proprietary information.

Under the trend of source-free domain adaptation, MAPS [3] and SFDAHPE [19] display excellent performance in pose estimation tasks. Both MAPS and SFDAHPE employ EMA techniques. However, MAPS utilizes Mixup augmentation and a self-training mechanism to achieve performance comparable to source-dependent methods, while SFDAHPE designs a novel training framework consisting of one primary and two auxiliary models that work in tandem. The three-model architecture of SFDAHPE balances the dual tasks of preserving the source knowledge and exploring the target knowledge. The architecture breakthrough leads SFDAHPE to the current state of the art in the DAPE.

Instead of source-free DAPE, Phase2Phase focuses on the source-dependent scenario. Although the increasing societal emphasis on privacy rights has led to a rise in

doi:10.6342/NTU202404439

source-free research, this does not signify a decline in traditional source-dependent domain adaptation studies. In fact, with advances in 3D modeling and enhanced image quality rendered by game engines, obtaining large amounts of data to train a model has become easier. This development has enhanced accessibility to classic domain adaptation applications, particularly in cases where models are trained on synthetic data and deployed in real-world scenarios. Therefore, Phase2Phase continues exploring ways to improve source-dependent domain adaptation.

doi:10.6342/NTU202404439

doi:10.6342/NTU202404439

Chapter 3 Methodology

In this chapter, we begin by defining the problem of DAPE in Section 3.1. This is followed by the preliminaries in Section 3.2. Section 3.3 presents the pipeline of our proposed Phase2Phase, highlighting Adaptive Mean Teacher, Two-Phase Training Framework, T-VAT-based UniMatch, and Mixup augmentation within each respective subsection. Finally, Section 3.4 provides the pseudocode for Phase2Phase, showcasing the implementation details of both the Teacher Guidance Phase (TGP) and the Augmentation-Based Enhancement Phase (ABEP)—the two training phases of my proposed Two-Phase Training Framework.

3.1 Problem Statement

The objective of domain adaptive pose estimation is to develop a method that first trains a model in the labeled source domain (e.g., synthetic or laboratory environments) and then retains and expands this knowledge to perform effectively in the desired target domain (e.g., real-world scenarios).

Due to the shared need to handle both labeled and unlabeled data, many UDA approaches are closely aligned with SSL. Although prior successes show the practicality of SSL methods in UDA tasks, applying SSL techniques to UDA still poses particular

doi:10.6342/NTU202404439

challenges. The challenge results from the different expectations for the distributions of labeled and unlabeled data. This thesis addresses the challenges of applying SSL techniques to UDA tasks. Specifically, i) we propose Adaptive Mean Teacher in Section 3.3.1 to address the slow update issue of the regular Mean Teacher during the early stage of domain adaptation. ii) Section 3.3.2 demonstrates that the Two-Phase Training Framework enhances domain adaptation by leveraging three types of perturbations: model-level perturbations, input-level perturbations, and feature-level perturbations. Lastly, given the outstanding result of MAPS [3], Mixup augmentation [36] is introduced to the whole training process to increase the robustness of our model.

3.2 Preliminaries

In domain adaptive pose estimation, data of two distributions is involved [19]. The two distributions are the source domain and the target domain. Let S be a source pose dataset with n^s labeled image-pose pairs $(x_i^s, y_i^s)_{i=1}^{n^s}$. Correspondingly, let T be a target pose dataset containing n^t unlabeled images $(x^t)_{i=1}^{n^t}$. The source and target images are represented respectively as $x_i^s, x_i^t \in \mathbb{R}^{H \times W \times 3}$, where H, W denote the height and width of those images. It is important to note that the pose annotations are only available in the source domain and are represented as $y_i^s \in \mathbb{R}^{K \times 2}$, where K is the number of keypoints.

doi:10.6342/NTU202404439

3.3 Pipeline of Our Methods

3.3.1 Adaptive Mean Teacher

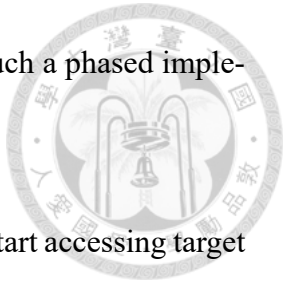
Mean Teacher is a classic consistency regularization method. As mentioned in Section 2.3, Mean Teacher introduces exponential moving average (EMA) to address the slow update issue of temporal ensembling and thus gains robust performance. While Mean Teacher performs adequately in the context of SSL, it experiences slow updates when applied to UDA, similar to the issue observed in temporal ensembling. The role of a teacher model is to guide the student model towards improvement and produce more reliable predictions than the student model. However, it is often observed that the opposite happens. Based on this observation, we propose the Adaptive Mean Teacher approach to mitigate this unexpected phenomenon. Adaptive Mean Teacher adopts the same EMA mechanism as Mean Teacher but emphasizes the importance of adjusting the smoothing coefficient (EMA decay rate) η.

$$\theta'_t = \eta\theta'_{t-1} + (1-\eta)\theta_t \tag{3.1}$$

The formula for EMA is presented in Formula 3.1. In this context, θ' denotes the parameter of the teacher model, θ represents that of the student model, and t signifies time. Typically, existing methods on DAPE often set the smoothing coefficient η to 0.999 [3, 9]. Nonetheless, Tarvainen and Valpola's experiments [28] indicate the efficacy of the ramp-up phase in SSL. During the ramp-up phase, the smoothing coefficient is initially set at a lower value of 0.99 and gradually transitions to the more commonly utilized value of 0.999. This systematic coefficient adjustment significantly enhances the performance of

doi:10.6342/NTU202404439

SSL compared to other experimental setups that do not incorporate such a phased implementation.



The reason is that the student model tends to learn rapidly when start accessing target data. This phenomenon can be analogously explained using the human learning curve, with the related explanation provided in Appendix A. Therefore, to maintain its guiding role, the teacher model must update quickly during this initial period to remain effective. Implementing a ramp-up strategy is critical in the realm of SSL. From our perspective, its importance is further amplified in UDA scenarios, where the distributions of labeled and unlabeled data often vary significantly.

Consequently, we place particular emphasis on the ramp-up phase in our study. We split the training involving Mean Teacher into two phases: an initial ramp-up phase and a subsequent regular training phase. The ramp-up phase, applied at the beginning of domain adaptation for five epochs, utilizes a smoothing coefficient of 0.99. Afterward, the smoothing coefficient is adjusted to 0.999 for the regular training phase. Our experiments indicate that Adaptive Mean Teacher accelerates model convergence and scores higher than traditional Mean Teacher.

3.3.2 Two-Phase Training Framework

In consistency regularization, various perturbations are applied to produce differing outputs, which are then used to calculate consistency. Some methods introduce perturbations within the model [10, 28], while others apply perturbation through the input or feature augmentation [23, 33]. Previous research in SSL and UDA typically focuses exclusively on one part. However, these two methods represent distinct systems within the

doi:10.6342/NTU202404439

consistency-based SSL framework and theoretically can be combined. It is reasonable to anticipate that combining these strategies could enhance outcomes.

Building upon the Adaptive Mean Teacher framework, we hypothesized that the integration of a weak-to-strong augmentation scheme and feature augmentation could further enhance its effectiveness. In the implementations of prior research, the augmentation pool is typically shared between the student and teacher models. Surprisingly, after modifying the shared augmentation pool to a weak-to-strong augmentation scheme, we observed a deterioration in performance.

This phenomenon may be attributed to the fact that a model trained exclusively on the source domain is not expected to demonstrate superior performance on weakly augmented data (e.g., using only flip-and-shift data augmentation) relative to strongly augmented data. Although in SSL, a model is typically expected to perform better with weakly augmented data due to the uniformity in distribution between labeled and unlabeled data, the situation differs in UDA. In UDA, the data distributions between the source (training) and target (application) domains often vary significantly. This difference in data distributions necessitates a distinct approach because the principles applicable in SSL might not hold for UDA scenarios. Therefore, we divide the whole training process into two separate phases: the Teacher Guidance Phase (TGP) and the Augmentation-Based Enhancement Phase (ABEP). In the TGP, Adaptive Mean Teacher is applied to train the model to its peak performance and ensure it performs well in the target domain. In ABEP, we develop an architecture that simultaneously perturbs both inputs and features by integrating the concepts from UniMatch [33], aiming to enhance the performance further.

Throughout the entire training process, whether in TGP or in ABEP, the loss can be

doi:10.6342/NTU202404439

divided into the following three parts as the following:

$$L = L_{sup} + L_{con} + L_{mix} \tag{3.2}$$

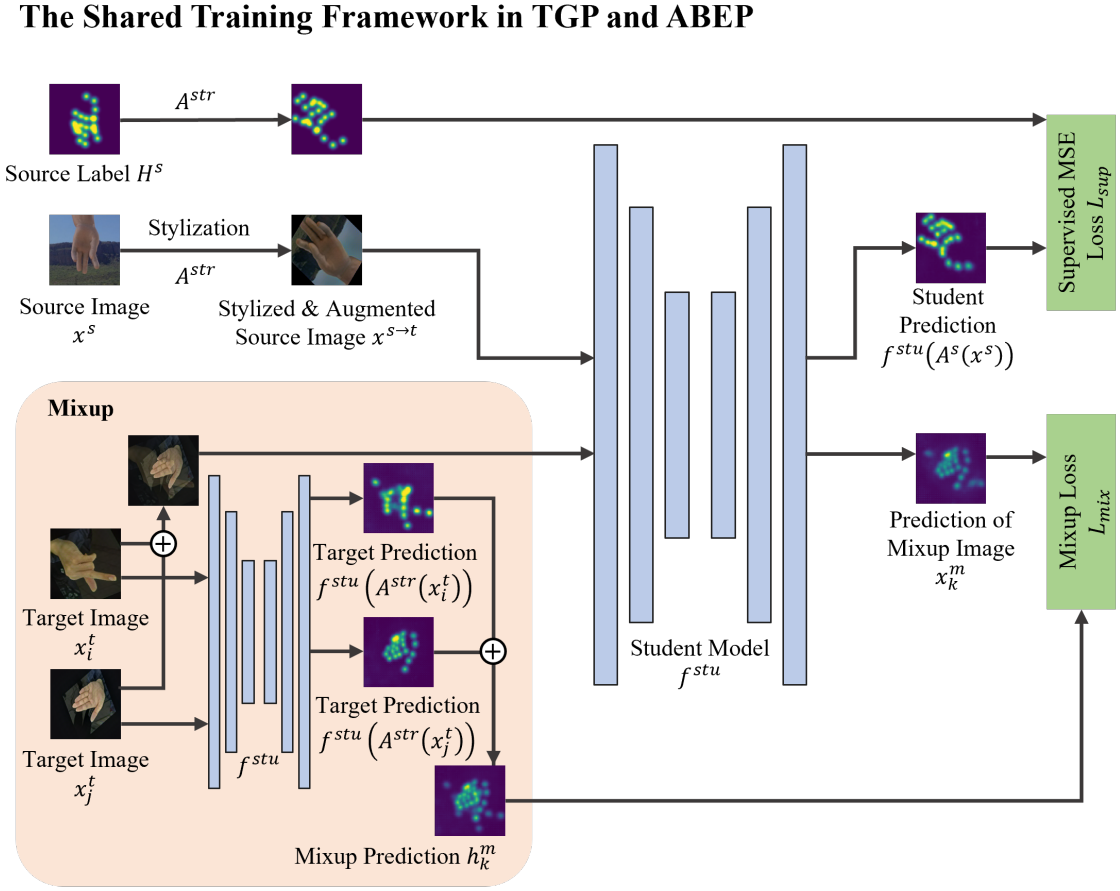


Figure 3.1: The Shared Training Framework in TGP and ABEP. The figure illustrates the supervised loss, L_{sup}, and the mixup loss, L_{mix}.

The supervised loss, denoted as L_{sup}, is implemented to preserve source domain knowledge throughout the domain adaptation process. The consistency loss, L_{con}, serves as a driving force in the training process with unlabeled target data. Furthermore, given the success of MAPS [3], the Mixup loss, represented as L_{mix}, is introduced to enhance the robustness of the model. In our work, L_{sup} and L_{con} are two shared components throughout TGP and ABEP as shown in Figure 3.1 while the implementation details of consistency

doi:10.6342/NTU202404439

loss differ in TGP and ABEP.

The supervised loss L_{sup} in the source domain is straightforward. L_{sup} is calculated by the output of the student model and the ground truth. The source images are subjected to augmentation A^s, and the ground truth also goes through the transformation part of these augmentations to reflect the changes in keypoints. Given the student model f^{stu} and the ground truth heatmap H^s, we calculate the supervised loss using Formula 3.3. We employ the Mean Squared Error (MSE) criterion to compare the two outputs.

$$L_{sup} = \frac{1}{n^s}\Sigma_{x^s \in S}\| f^{stu}(A^s(x^s)) - A^s(H^s)\|_2 \tag{3.3}$$

The core concept of consistency regularization L_{con} is to refine the accuracy of *follower* predictions by utilizing the more reliable *guide* predictions as a reference point for improvement. In Formula 3.4, H^{gd} symbolizes the *guide* heatmap predictions, which is then normalized to the pseudo label $\hat{H}^{gd}$. Similarly, H^{flw} signifies the *follower* predictions. The target images are processed with different augmentations, A_1^t and A_2^t, before being processed by the student and teacher models. To compare their consistency, we apply reversing augmentation $\tilde{A}_1^t$ to $\hat{H}^{gd}$ and $\tilde{A}_2^t$ to H^{flw}. We introduce τ_{conf} to ensure that stronger predictions guide weaker ones only when the predictions are confident. This condition is met when the maximum activation point on H^{gd}, denoted as $H_k^{gd,\hat{y}}$, is greater than or equal to the threshold τ_{conf}.

$$L_{con} = \frac{1}{n^t}\Sigma_{x^t \in T}\Sigma_{k=0}^{K}\mathbb{1}(H_k^{gd,\hat{y}} >= \tau_{conf})\|\tilde{A}_1^t(\hat{H}_k^{gd}) - \tilde{A}_2^t(H_k^{flw})\|_2 \tag{3.4}$$

The implementation of consistency regularization in the Teacher Guidance Phase

doi:10.6342/NTU202404439

(TGP) follows the method outlined in [9]. We denote this implementation loss as $L_{con,TGP}$, with its formulation presented in Formula 3.5. Augmentations A_1^t and A_2^t draw from the same pool, using the same strong augmentation setting in ABEP. Consistent with the traditional Mean Teacher model, we designate the teacher model f^{tea} as the *guide* and the student model f^{stu} as the *follower*.

$$L_{con,TGP} = \frac{1}{n^t}\Sigma_{x^t \in T}\Sigma_{k=0}^{K}\mathbb{1}(f^{tea}(A_1^t(x_k^t))^{\hat{y}} >= \tau_{conf}) \\ \|\tilde{A}_1^t(\hat{f}^{tea}(A_1^t(x_k^t))) - \tilde{A}_2^t(f^{stu}(A_2^t(x_k^t)))\|_2 \tag{3.5}$$

3.3.3 T-VAT-based UniMatch

The essence of weak-to-strong consistency regularization lies in using the prediction from a weakly perturbed image to supervise its strongly perturbed counterpart. In our method, the weakly perturbed category encompasses subtle transformations like rotation and translation, which do not affect the intrinsic texture of the images, while the strongly perturbed category includes more pronounced alterations like color jittering, significantly altering the appearance of images.

After the TGP, the model is expected to successfully perform the task and demonstrate at least basic capabilities within the target domain. In this situation, the model can be expected to perform better with weakly augmented images than with strongly augmented images, thus fulfilling the requirements of the weak-to-strong consistency regularization. Therefore, we integrate the UniMatch framework and introduce VAT from the teacher model (T-VAT) for feature perturbation, resulting in our T-VAT-based UniMatch approach. T-VAT-based UniMatch is the main feature in ABEP. Within that phase, the

doi:10.6342/NTU202404439

The Training Framework of Consistency Regulation Loss

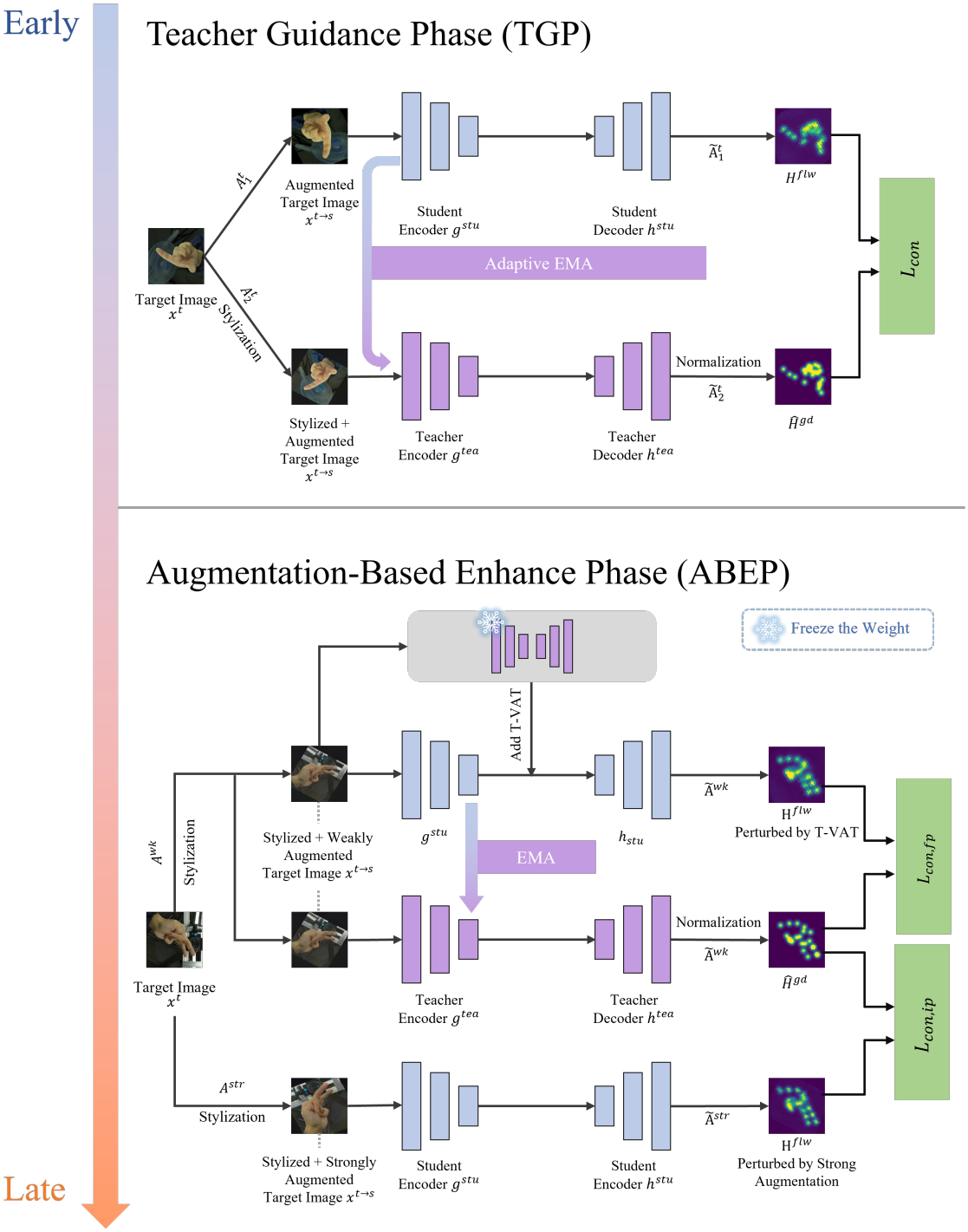


Figure 3.2: The Consistency Loss throughout TGP and ABEP

doi:10.6342/NTU202404439

consistency regularization is as follows:

$$L_{con,ABEP} = \frac{1}{2}(L_{con,ip} + L_{con,fp}) \tag{3.6}$$

Under the Mean Teacher framework with model-level perturbation, the total consistency loss in the ABEP $L_{con,ABEP}$ is divided into two parts, focusing separately on input-level perturbations $L_{con,ip}$ and feature-level perturbations $L_{con,fp}$. The input-level perturbations are obtained by introducing the weak-to-strong strategy in FixMatch [23]. The weak augmentation A^{wk} is employed in the input image to the Mean Teacher to yield stable predictions. Conversely, the strong augmentation A^{str} is utilized to introduce the perturbations to the input image of the student model. Regarding the model, a heatmap-based pose estimation model f can be decomposed into an encoder h and a decoder g. We used stu and tea as subscripts to mark the student model and the teacher model. For example, f^{stu} represents the student model while f^{tea} means the teacher model.

Based on Formula 3.4, we developed the consistency regularization loss for image-level perturbation in ABEP in Formula 3.7. In this specialized formula, A^{wk}, representing weak augmentation with minor image changes, corresponds to A_1^t in Formula 3.4. Similarly, A^{str}, which denotes strong augmentation causing more significant image alterations, aligns with A_2^t in Formula 3.4. Furthermore, we use the output heatmap of the teacher model as H^{gd}, and H^{flw} corresponds to that of the student model. To summarize, the overall formula is as follows.

$$\begin{aligned} L_{con,ip} =& \frac{1}{n^t}\Sigma_{x^t \in T}\Sigma_{k=0}^{K}\mathbb{1}(f^{tea}(A^{wk}(x_k^t))^{\hat{y}} >= \tau_{conf}) \\ & \|\tilde{A}^{wk}(\hat{f}^{tea}(A^{wk}(x_k^t))) - \tilde{A}^{str}(f^{stu}(A^{str}(x_k^t)))\|_2 \end{aligned} \tag{3.7}$$

doi:10.6342/NTU202404439

On the other hand, when addressing the consistency regularization for feature-level perturbation, A_1^t and A_2^t share the same weak augmentations, making it unnecessary to apply reversing augmentations for consistency regularization. In this context, H^{gd} corresponds to the direct output of the teacher model, while H^{flw} represents the output of the teacher model subjected to the feature perturbation, VAT, denoted as r_{adv} [13, 33]. The complete formula for $L_{con,fp}$ is presented below:

$$L_{con,fp} = \frac{1}{n^t}\Sigma_{x^t \in T}\Sigma_{k=0}^{K}\mathbb{1}(f^{tea}(A^{wk}(x_k^t))^{\hat{y}} >= \tau_{conf}) \\ \|f^{tea}(A^{wk}(x_k^t)) - g^{tea}(h^{tea}(A^{wk}(x_k^t)) + r_{adv})\|_2 \tag{3.8}$$

The accurate positions of the keypoints are required in pose estimation task, and Dropout tends to hide too much information when the model makes the predictions. Therefore, regarding the feature perturbation, we use T-VAT proposed by the PSMT [13]. PSMT adopts T-VAT to enhance the performance of the model in domain adaptive semantic segmentation. Given its success in semantic segmentation, we believe T-VAT can serve as an effective feature-level perturbation in DAPE, as the pose estimation task similarly values detailed image information.

During the feature-level consistency regularization, T-VAT is added to the feature $h^{tea}(A^{wk}(x^t))$ in Formula 3.8. The derivation of r_{adv} is achieved by optimizing the goal specified in the Formula 3.9. MSE is applied to compare the predictions with and without r_{adv}. Following backpropagation, the gradient of r_{adv} is extracted. We employ the gradient in T-VAT because it identifies perturbations that bring the model closer to its decision boundary, ensuring the model remains robust even in challenging scenarios. The process of obtaining the T-VAT is illustrated in Figure 3.3.

doi:10.6342/NTU202404439

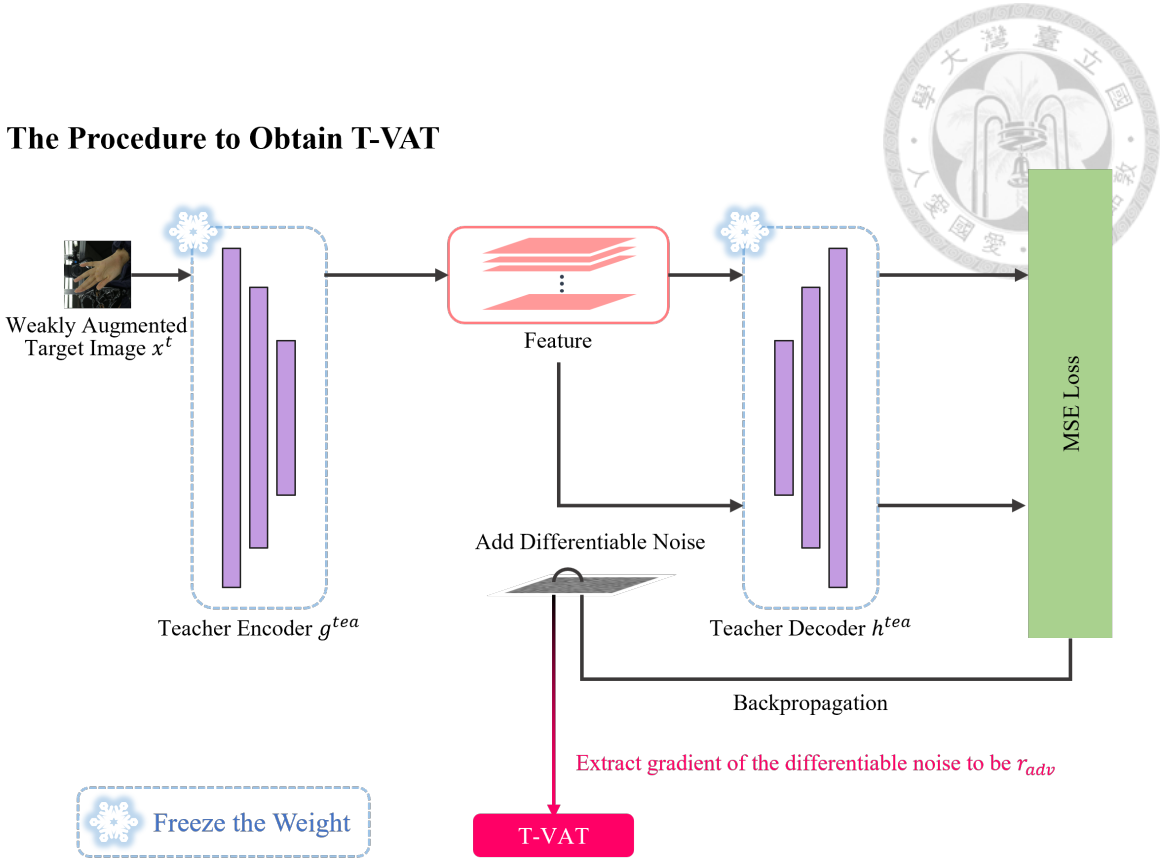


Figure 3.3: The Process of Obtaining T-VAT

$$\begin{aligned} &\text{maximize } d(\tilde{A}^{wk}(f^{tea}(A^{wk}(x^t))), \tilde{A}^{wk}(h^{tea}(g^{tea}(A^{wk}(x^t)) + r_{adv}))) \\ &\text{subject to } \|r_{adv}\|_2 \leq \epsilon. \\ &\text{The gradient of } r_{adv} \text{ is employed as T-VAT.} \end{aligned} \tag{3.9}$$

Compared to PSMT, our work also employs a teacher model due to its relative stability. However, while PSMT ensembles outputs from two teacher models for calculating VAT, we utilize only one teacher model to obtain VAT, denoted as T-VAT.

3.3.4 Mixup Augmentation

The efficacy of Mixup [36] is validated by MAPS [3] in the source-free DAPE. In Phase2Phase, Mixup augmentation is also added to boost the performance further in the

doi:10.6342/NTU202404439

source-dependent scenario. MAPS indicates that Mixup improves the handling of intermediate target data, thereby increasing the robustness of the model.

Benefiting from the advantage of source-dependent methods that enable the use of source data, we experiment with two configurations: a Mixup of source data combined with target data and a Mixup within the target domain itself. Ultimately, guided by the performance in the experiments in TGP, the Mixup of only target domain data is chosen as our final strategy.

In each mini-batch consisting of n samples, the strongly augmented target data $A^{str}(x^t)$ is shuffled. Two samples $A^{str}(x_i^t)$ and $A^{str}(x_j^t)$ are then mix based on mix ratio $\rho \sim Beta(\alpha, \beta)$. In Phase2Phase, the hyper-parameter α and β in the Beta distribution of the Mixup ratio are set to 0.75. The Mixup loss L_{mix} in the target domain is specified as follows:

$$
\begin{aligned}
L_{mix} &= \Sigma_{k=1}^{n} \| f^{stu}(x_k^m) - h_k^m \| \\
x_k^m &= \rho A^{str}(x_i^t) + (1-\rho) A^{str}(x_j^t) \\
h_k^m &= \rho f^{stu}(A^{str}(x_i^t)) + (1-\rho) f^{stu}(A^{str}(x_j^t)) \\
&\text{where } i, j \text{ are the image indices.}
\end{aligned} \tag{3.10}
$$

3.4 Algorithm

Phase2Phase design a Two-Phase Training Framework as shown in Section 3.3.2. The algorithms for the two training phases are detailed as follows:

doi:10.6342/NTU202404439

Algorithm 1: Algorithm of the Teacher Guidance Phase (TGP) of the Proposed Phase2Phase

Input: Source dataset S; Target dataset T; a well-trained source model f^s; Augmentation A^s for source data; Augmentation A_1^t and A_2^t for target data.

Output: New target student model model f^{stu} and target teacher model f^{tea}.

initialize the student f^{stu} and teacher f^{tea} with the weights from f^s;
while during epochs for TGP **do**
    // process data
    Stylized $x^s \in S$ and $x^t \in T$ with AdaIn;
    Augment source data with A^s;
    Augment target data with A_1^t and A_1^t;
    // the shared loss between TGP and ABEP
    Calculate L_{sup} by Formula 3.3;
    Calculate L_{mix} by Formula 3.10;
    // consistency regularization
    Get $f^{tea}(A_1^t(x^t))$ as H^{gd};
    Normalize H^{gd} to get $\hat{H}^{gd}$ as the pseudo label;
    Get $f^{stu}(A_1^t(x^t))$ as H^{flw};
    Calculate L_{con} by Formula 3.4;
    // update the model
    Update f^{stu} by Formula 3.2;
    /* When updating f^{tea} in TGP, the smoothing coefficient η is applied according to Adaptive Mean teacher, which means 0.99 for the first five epochs and 0.999 for the others. */
    Update f^{tea} by Formula 3.1.
end

doi:10.6342/NTU202404439

Algorithm 2: Algorithm of the Augmentation-Based Enhancement Phase (ABEP) of the Proposed Phase2Phase

Input: Source dataset S; Target dataset T; target student model trained after TGP f^{stu}; target teacher model trained after TGP f^{tea}; Augmentation A^s for source data; Augmentation A_1^t and A_2^t for target data.

Output: well trained target model f^{stu}.

while during epochs for TGP **do**

```
// process data
```

Stylized $x^s \in S$ and $x^t \in T$ with AdaIn;

Augment source data with A^s;

Augment target data with A^{wk} as A_1^t and A^{str} as A_2^t;

```
// the shared loss between TGP and ABEP
```

Calculate L_{sup} by Formula 3.3;

Calculate L_{mix} with $A^{str}(x^t)$ by Formula 3.10;

```
// consistency regularization
```

Calculate consistency loss for input-level perturbation by Formula 3.7;

Calculate consistency loss for feature-level perturbation by Formula 3.8;

Calculate the total consistency loss by Formula 3.6;

```
// update the model
```

Update f^{stu} by Formula 3.2;

```
/* When updating f^tea in ABEP, the smoothing coefficient η is applied
   according to the typical Mean Teacher, which means 0.999
   throughout the process.                                          */
```

Update f^{tea} by Formula 3.1.

end

doi:10.6342/NTU202404439

doi:10.6342/NTU202404439

Chapter 4 Experiment

In Chapter 4, we start by presenting the datasets in Section 4.1, followed by a detailed discussion of the evaluation metric in Section 4.2. Subsequently, Section 4.3 outlines the experimental protocol employed in Phase2Phase. The evaluation and results are then thoroughly examined in Section 4.4, where we compare Phase2Phase with other DAPE work, covering both source-dependent and source-free approaches. In Section 4.5, An ablation study is conducted to verify the effectiveness of the components proposed in Phase2Phase. In Section 4.6, we performed a sensitivity analysis, examining the impact of different configurations on Adaptive Mean Teacher, Mixup Augmentation, and Augmentation in ABEP.

4.1 Dataset

Our study concentrates on unsupervised domain adaptation for hand pose estimation tasks. Consistent with prior research in DAPE, we choose the Rendered Hand Pose Dataset (RHD) and the Hand-3D-Studio Dataset (H3D). In our experiments, RHD serves as the source domain, and H3D serves the target domain.

doi:10.6342/NTU202404439

4.1.1 Rendered Hand Pose (RHD)

The Rendered Hand Pose (RHD) dataset [14] is a synthetic dataset comprising 41,258 training images and 2,728 testing images. This dataset is crafted using freely available 3D human models and animations from Mixamo, rendered using the open-source software Blender. In creating RHD datasets, 20 characters and 39 actions are involved. Additionally, the camera position is randomly determined in a spherical area around one of the hands. As a result, RHD offers a diverse collection of images encompassing various viewpoints and complex scenarios. However, despite the diversity of the images, there is still a substantial difference in appearance between the data set and real-world scenarios.

Figure 4.1: Samples from the RHD dataset

4.1.2 Hand-3D-Studio (H3D)

Hand-3D-Studio (H3D) dataset [38] is a real-world dataset of multi-view color hand images. Ten people of different genders and skin colors are involved. In addition, 50 one-handed gestures and 27 hand-object interaction gestures are included. There are 22,000 frames in the dataset. Following the setting of an existing method, 3,200 frames are randomly picked for testing, and the other frames are utilized as the training set. Because the

doi:10.6342/NTU202404439

images in H3D are extracted from videos, it is common to observe a strong similarity in appearance among them.

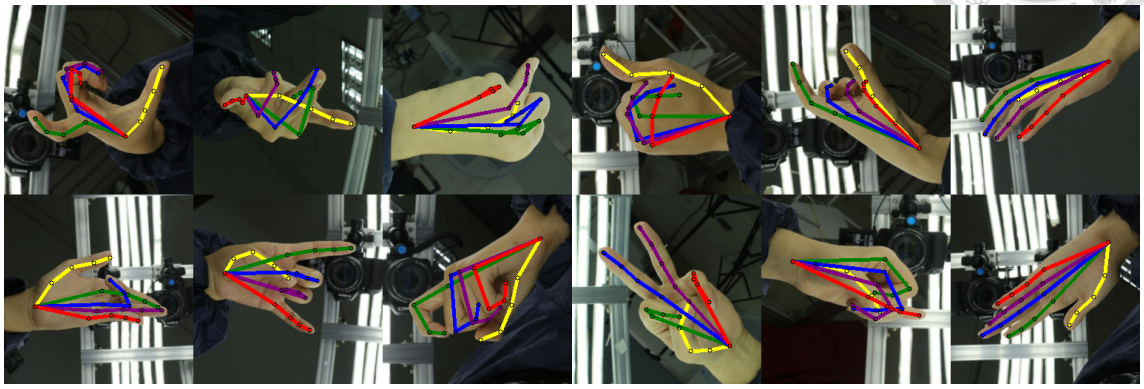


Figure 4.2: Samples from the H3D dataset

4.2 Evaluation Metrics

Percentage of Correct Keypoint (PCK) is applied as the evaluation metric in our research. This is particularly suitable for pose estimation tasks because keypoints often cover more than a single pixel. For instance, in the case of a hand image, the keypoints representing finger joints would occupy an area rather than a single point. Therefore, we utilize PCK@0.05 to accommodate this characteristic of keypoints in our evaluation. Specifically, the predictions within the range of 5% with respect to the image size are considered correct in the calculation of the PCK@0.05.

In the context of computing the PCK@0.05 metric as defined in Formula 4.1, the calculation involves the indicator function $\mathbb{1}$ to assess whether the Euclidean distance d between the predicted and ground truth positions of K keypoints in an input image falls within a specified threshold. When predictions fall within 5% of the original image size around the ground truth of keypoints, this function returns a value of 1; otherwise, it returns

doi:10.6342/NTU202404439

0.

$$PCK = \frac{\Sigma_{i=1}^{K} \mathbb{1}(d_i < \tau_{pck})}{K},$$
$$\text{where } \tau_{pck} = 0.05. \tag{4.1}$$

We report the average PCK@0.05 across all 21 keypoints. Additionally, we provide the PCK@0.05 for each hand part, including metacarpophalangeal (MCP), proximal interphalangeal (PIP), distal interphalangeal (DIP), and fingertip (Fin).

4.3 Experiments Setup

In our work, we employ the Simple Baseline [29] architecture with a ResNet101 [5] backbone. The training spans 90 epochs, each comprising 500 iterations with a batch size of 32. The training process is divided into two stages. The first 40 epochs are devoted to developing proficiency in the source domain. The subsequent stage, from the 41st to the 90th epoch, concentrates on domain adaptation to optimize performance on the target data, our primary objective. As detailed in Section 3.3.2, this domain adaptation stage consists of two phases. Initially, the Teacher Guidance Stage (TGP) spans 30 epochs, focusing on training the model to achieve peak performance in the source domain. This is followed by the Augmentation-Based Enhancement Stage (ABEP), lasting 20 epochs, which aims to further enhance performance using a weak-to-strong augmentation strategy.

As for the choice of the optimizer, we use Adam for the student model, following the prior work [7, 9]. The initial learning rate is 1e-4, and it is reduced to 1e-5 and 1e-6 after the 55th and the 85th epoch respectively. On the other hand, with the implementation of the

doi:10.6342/NTU202404439

Adaptive Mean Teacher, the smooth coefficient η, which updates the teacher model using EMA mechanism, experiences a ramp-up during the initial epochs of domain adaptation. Specifically, for the first five epochs following the 40^{th} epoch, η is set to 0.99. After this ramp-up phase, the smooth coefficient transitions to 0.999, aligning with the traditional practice in the typical Mean Teacher approach.

Our method enhances the UniFrame architecture [9]. We also employ AdaIN [6] to adjust source domain data to align with target domain characteristics, and vice versa, as illustrated in Figure 3.1 and 3.2.

4.4 Evaluation and Results

4.4.1 Quantitative Results

The Table 4.1 is divided into two sections: the upper section shows the results of source-dependent tasks, while the lower section focuses on the performance in source-free (SF) settings. In our experiments, Phase2Phase achieves state-of-the-art results in the source-dependent domain. Remarkably, even when compared to recent source-free research, Phase2Phase's performance remains competitive. Among all evaluated methods, Phase2Phase not only achieves the highest scores in source-dependent DAPE but also outperforms SFDAHPE by 0.5% in MCP and closely matches SFDAHPE's results in other parts of the hand.

During the domain adaptation stage, SFDAHPE employs three models for collaborative training to achieve state-of-the-art results, whereas Phase2Phase utilizes only two models for joint training. This configuration means that Phase2Phase requires only two-

doi:10.6342/NTU202404439

thirds of the GPU memory for model operation compared to SFDAHPE.

Table 4.1: PCK@0.05 on RHD → H3D Task. SF indicates whether these work are targeted for source-free tasks. The highest value is highlighted in orange and underlined, while the second highest value is distinguished by a purple and **bolded**.

Method	SF	MCP	PIP	DIP	Fin	All
RegDA [7]	×	79.6	74.4	71.2	62.9	72.5
UniFrame [9]	×	86.7	84.6	78.9	68.1	79.6
Phase2Phase (ours)	×	88.9	**86.7**	**79.9**	**71.1**	**81.7**
MAPS [3]	✓	86.9	84.8	79.1	69.3	80.0
SFDAHPE [19]	✓	**88.4**	89.2	80.9	71.4	82.2

4.4.2 Qualitative Results

Figure 4.3 presents the qualitative results for H3D. For a comprehensive comparison, we include results from *Source Only*, *UniFrame* [9], *MAPS* [3], ours *Phase2Phase*, and the *Ground Truth*. We have not implemented the current state of the art, SFDAHPE, as we are currently unable to reproduce the results of this work with the provided code.

4.5 Ablation Study on Framework

We conduct an ablation study on the framework to validate the robustness of our method, Phase2Phase. As indicated in Table 4.2, implementing the Adaptive Mean Teacher significantly enhances the training outcomes. Merely introducing a ramp-up phase during the initial epochs leads to a 1.1% increase in accuracy in the target domain. Additionally, the effectiveness of Mixup has also been validated. Upon incorporating Mixup augmentation, we observed an increase in the average PCK@0.05 by 0.5% with the Mean Teacher and 0.4% with the Adaptive Mean Teacher respectively. Last but not least, Table 4.2 illustrates that integrating the T-VAT-based UniMatch boosts performance, contributing

doi:10.6342/NTU202404439

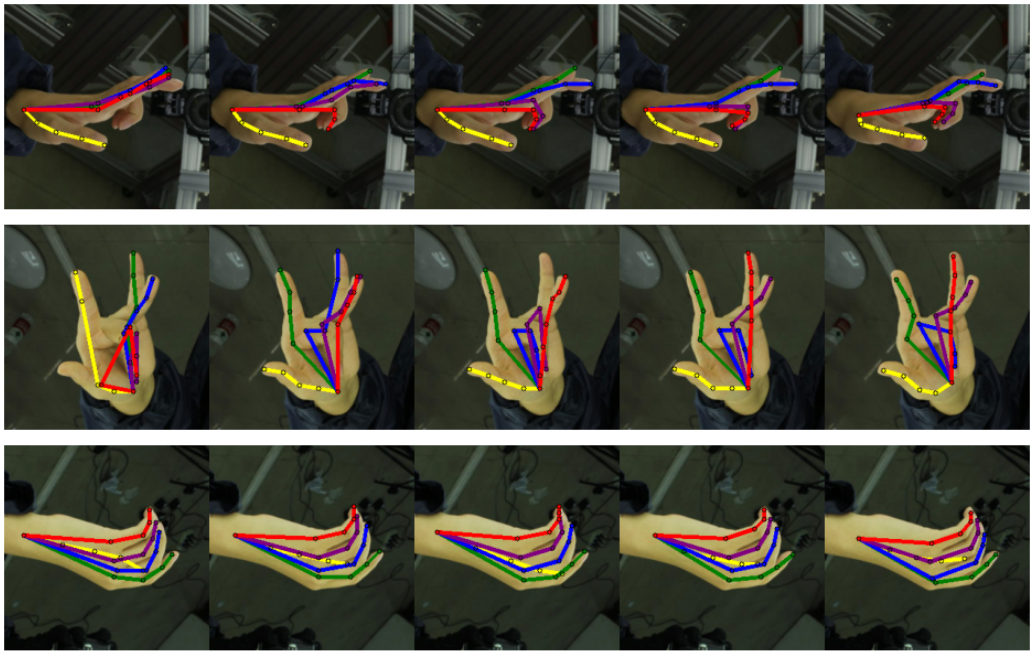


Figure 4.3: Qualitative Results on the H3D Dataset.

to a 0.6% increase in average PCK@0.05. These results underscore the effectiveness of our Two-Phase Training Framework and demonstrate that if properly employed, weak-to-strong consistency regularization can effectively complement the Mean Teacher approach.

Table 4.2: Ablation Study on Framework. The effectiveness of three main components and their interactions on task RHD → H3D. The highest value is highlighted in orange and underlined, while the second highest value is distinguished by a purple and **bolded**.

Method			RHD → H3D				
Adaptive MT	Mixup	UniMatch(T-VAT)	MCP	PIP	DIP	Fin	All
			86.7	84.6	78.9	68.1	79.6
✓			87.5	86.0	**80.1**	68.9	80.7
	✓		86.3	84.7	**80.1**	**69.6**	80.1
✓	✓		**87.8**	**86.4**	80.3	69.4	**81.1**
✓	✓	✓	88.9	86.7	79.9	71.1	81.7

doi:10.6342/NTU202404439

4.6 Sesensitiveitive Analysis

4.6.1 Sensitive Analysis on the Adaptive Mean Teacher

Figure 4.4: The PCK@0.05 Performance in the Target Domain. In this figure, The shared pretraining process, spanning from the 1st to the 40th epoch, is marked in orange. During TGP, from the 41st to the 70th epoch, four training curves are plotted. The blue and purple lines indicate the performance of the student and teacher models, respectively. Within these, the darker shades represent the performance of our Adaptive Mean Teacher (Adaptive MT), while the lighter shades depict that of the typical Mean Teacher (MT). To highlight the four training curves during the TGP phase, we magnify this training process in the yellow rectangle on the right side of the figure.

In the process of reproducing UniFrame [9], I found that the performance of the student model and the teacher model in the target domain contradicted the intuition. Mean Teacher is utilized for consistency regularization. Theoretically, the teacher model in this framework should perform more stably and reliably than the student model, serving as the guide during training. However, in DAPE, Figure 4.4 shows that the performance of the teacher model consistently trails behind that of the student model.

We hypothesize that the poorer predictions of the teacher model are due to an improp-

doi:10.6342/NTU202404439

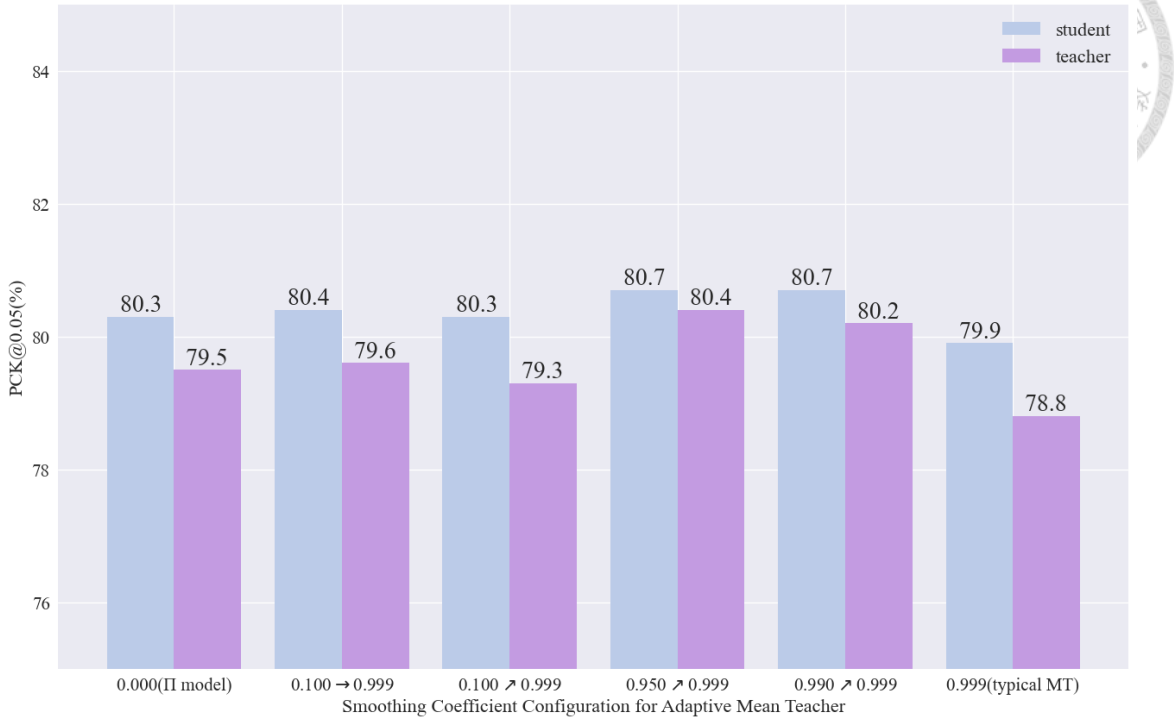


Figure 4.5: Sensitivity Analysis of the Smoothing Coefficient. All scores represent the results obtained after training for 30 epochs in the TGP.

erly set smoothing coefficient η. We can observe from Figure 4.4 that there is a sudden improvement in performance in the target domain after the 40th epoch, the start of the domain adaptation. If the smoothing coefficient is too high during the early stages of domain adaptation, it becomes challenging for the teacher model to capture the rapid pace of model improvement. Consequently, we conducted six experiments in TGP to evaluate the effect of varying the smoothing coefficient, as summarized in Figure 4.5. In the first setting, the smoothing coefficient is set to 0.000, indicating the adoption of the Π model [10]. In the second setting, the smoothing coefficient remains at 0.100 initially to allow the teacher model to adapt to the rapid improvement in the early epochs. Then, it is readjusted to 0.999 after 5 epochs. For the 3rd to 5th settings, the smoothing coefficient starts at 0.100, 0.950, and 0.999, respectively. Then, the smoothing coefficient progresses in equal increments to reach 0.999, starting from the 6th epoch. The last setting represented the typical

doi:10.6342/NTU202404439

Mean Teacher approach, where the smoothing coefficient remained at 0.999 throughout the domain adaptation.

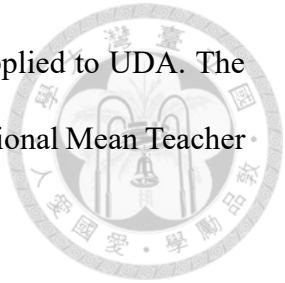
According to Figure 4.5, employing the typical Mean Teacher approach results in the poorest performance. Even the Π model yields better results. These outcomes suggest the necessity of a rapid update for the Mean Teacher during the early stage of the domain adaptation. The second and third settings, which begin with a low coefficient, show performance comparable to the Π model. Notably, settings starting with coefficients of 0.950 and 0.990 achieve the highest performance, surpassing their Mean Teacher counterparts by a significant approximately 1% in PCK@0.05.

Although the smoothing coefficient of 0.950 and 0.999 seems still high, in a scenario with 500 iterations per epoch, only 1% of the original model parameter can be preserved after an epoch. They achieve the fast update at the early stage and preserve the quality of the Mean Teacher to avoid fluctuation. Based on the outcomes of the sensitivity analysis and rand ramp-up phase outlined in the Mean Teacher paper [28], we adopt the strategy of applying 0.990 at the first five epochs of domain adaptation and 0.999 for the others.

Observing Figure 4.4, we can clearly see that the Adaptive Mean Teacher, having implemented a ramp-up phase, outperforms the typical Mean Teacher in both the student and teacher models' performance. During the ramp-up phase, we successfully enable the teacher model to catch up with the student model. This effectively addresses our initial hypothesis that the teacher model struggles to keep pace with the rapidly advancing student model in the early stages of UDA. However, in the later stage of TGP, we observed that the student model exhibits stronger and more stable predictive capabilities. The concrete reasons for this phenomenon remain unclear, but we believe that the Adaptive Mean

doi:10.6342/NTU202404439

Teacher partially resolves the problems Mean Teacher faces when applied to UDA. The Adaptive Mean Teacher indeed improves the applicability of the traditional Mean Teacher in UDA scenarios.



4.6.2 Sensitive Analysis on the Mixup Augmentation

Phase2Phase aims to address the task of source-dependent DAPE that allows access to source data. This suggests that for Mixup augmentation, we have the flexibility to either mix two sets of target data (t+t) like MAPS [3] or mix source and target data (s+t).

To determine which Mixup setting is more beneficial for DAPE, two types of Mixup are experimented in TGP. The first one implements both t+t and s+t, while the second one only adopts t+t. According to Table 4.3, we observed that the introduction of the Mixup of source data and target data does not improve the performance and even deteriorates the performance by 3% compared to not implementing Mixup. The reason for the deterioration cannot be determined at present. One possible explanation is that for tasks such as DAPE, which require more detailed predictions compared to classification tasks, the blend of inconsistent styles in composite images may be too complex for the model to handle. Based on the experimental findings, we have chosen the t+t setting as our definitive approach.

Table 4.3: Sensitivity Analysis of the Type of Mixup in TGP. The highest value is highlighted in orange and underlined.

Mixup Type	MCP	PIP	DIP	Fin	All
w/o Mixup	87.5	86.0	80.1	68.9	80.7
source + target(s+t), target + target(t+t)	86.9	85.4	79.3	69.5	80.4
target + target(t+t)	87.8	86.4	80.3	69.4	81.1

doi:10.6342/NTU202404439

4.6.3 Sensitive Analysis of the Augmentation in ABEP

During ABEP, the concept of weak-to-strong consistency regularization derived from FixMatch is utilized. The method generates theoretically stable and unstable predictions, and our goal is to guide the unstable outputs towards stability with the stable ones. In our experiments, the weak augmentation is quite straightforward. We employed only non-texture-altering techniques such as rotation and flipping, avoiding any methods that change textures, like color jittering. Regarding the strong augmentation, we experiment with two settings. The first, referred to as *soft FixMatch*, follows the UniFrame's approach but incorporates 360-degree rotation. The second setting, named *strict FixMatch*, adopts RandAugment in line with FixMatch. Compared to *soft FixMatch*, *strict FixMatch* is characterized by adding posterization and sharpening effects.

According to Table 4.4, we observe that stronger augmentation does not enhance performance; in fact, the implementation of *strict FixMatch* leads to a noticeable deterioration. This suggests that excessive augmentation may adversely affect the model's ability to generalize.

Table 4.4: Sensitivity Analysis of the Weak-to-Strong Augmentation in ABEP. The highest value is highlighted in orange and underlined.

Method	MCP	PIP	DIP	Fin	All
w/o ABEP	87.8	86.4	80.3	69.4	81.1
Strict FixMatch	88.1	85.9	79.7	69.9	80.9
Soft FixMatch	88.9	86.7	79.9	71.1	81.7

doi:10.6342/NTU202404439

Chapter 5 Conclusion

Domain adaptive pose estimation (DAPE) has increasingly captured the interest of researchers and practitioners in recent years. With similar goals between SSL and UDA, many SSL-derived methods are now being applied to tackle DAPE tasks. Although these approaches have achieved success, the differing distribution expectations between labeled and unlabeled data pose potential risks to the effectiveness of DAPE. In response, we propose Phase2Phase, a strategy meticulously designed to address the domain gap between labeled and unlabeled data. Phase2Phase resolves issues associated with the Mean Teacher paradigm in domain adaptation and effectively blends two distinct consistency-based SSL methods. Our experiments demonstrate that Phase2Phase outperforms previous source-dependent DAPE work and gains comparable results to the current state of the art in source-free DAPE, SFDAHPE.

doi:10.6342/NTU202404439

5.1 Contribution

5.1.1 Addressing Slow Update Challenges in Traditional Mean Teacher Framework with Adaptive Mean Teacher

We highlight an unusual phenomenon in DAPE where the Mean Teacher model underperforms compared to the student model. In response to this finding, We propose Adaptive Mean Teacher to mitigate this issue. The modification makes the traditional Mean Teacher more robust when applied in DAPE.

5.1.2 Integrating Weak-to-Strong Augmentation-Based Consistency Regularization into the Mean Teacher Framework

Weak-to-strong augmentation-based consistency regularization is a widespread SSL method that is challenging to apply in DAPE, and we successfully integrate it into DAPE. We note that a pretrained source model does not necessarily perform better on weakly augmented images than on strongly augmented ones. In response to this observation, we propose the Two-Phase Training Framework, which has led to significant success.

doi:10.6342/NTU202404439

5.2 Limitation and Future Work

5.2.1 Probing into the Underperformance of the Teacher Model Relative to the Student Model

By introducing the Adaptive Mean Teacher, Phase2Phase significantly outperforms the Mean Teacher counterparts in student and teacher scores. However, despite the success achieved by the Adaptive Mean Teacher, the issue of the teacher model lagging behind the student models persists. One potential hypothesis is that the ramp-up phase requires more epochs; however, the precise explanation remains uncertain. Consequently, addressing this challenge still presents a valuable opportunity for subsequent studies.

5.2.2 Expanding Experiments Across Various Tasks

In our thesis, we only conduct experiments on one domain adaptive pose estimation setup: RHD $\rightarrow$ H3D. In fact, we can take it a step further. For instance, we could test the generalization capability on the FreiHand dataset (FHD), similar to UniFrame [9]. On the other hand, given that the methods proposed in Phase2Phase are not limited to hand pose estimation, exploring their effectiveness in other subtasks of pose estimation, such as human or animal pose estimation, would be valuable. Moreover, the potential of Phase2Phase extends beyond pose estimation. We believe Phase2Phase could be effectively applied to other tasks, including classification and semantic segmentation.

doi:10.6342/NTU202404439

5.2.3 Exploring the Practicality in Source-Free DAPE

Our research focuses on the source-dependent DAPE. However, source-free DAPE has emerged as a popular research trend. Building upon the frameworks of UniFrame and MAPS, we achieved extraordinary performance in the RHD $\rightarrow$ H3D setup that is surpassed only by the source-free method, SFDAPE. We believe extending our research concepts to source-free DAPE holds great potential. Consequently, we look forward to future studies delving deeper into this area.

doi:10.6342/NTU202404439

References

[1] D. Berthelot, N. Carlini, E. D. Cubuk, A. Kurakin, K. Sohn, H. Zhang, and C. Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. arXiv preprint arXiv:1911.09785, 2019.

[2] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel. Mixmatch: A holistic approach to semi-supervised learning. Advances in neural information processing systems, 32, 2019.

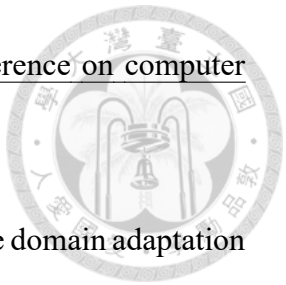
[3] Y. Ding, J. Liang, B. Jiang, A. Zheng, and R. He. Maps: A noise-robust progressive learning approach for source-free domain adaptive keypoint detection. arXiv preprint arXiv:2302.04589, 2023.

[4] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, and V. Lempitsky. Domain-adversarial training of neural networks. Journal of machine learning research, 17(59):1–35, 2016.

[5] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[6] X. Huang and S. Belongie. Arbitrary style transfer in real-time with adaptive instance

doi:10.6342/NTU202404439



normalization. In Proceedings of the IEEE international conference on computer vision, pages 1501–1510, 2017.

[7] J. Jiang, Y. Ji, X. Wang, Y. Liu, J. Wang, and M. Long. Regressive domain adaptation for unsupervised keypoint detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6780–6789, 2021.

[8] R. Jin, J. Zhang, J. Yang, and D. Tao. Multibranch adversarial regression for domain adaptative hand pose estimation. IEEE Transactions on Circuits and Systems for Video Technology, 32(9):6125–6136, 2022.

[9] D. Kim, K. Wang, K. Saenko, M. Betke, and S. Sclaroff. A unified framework for domain adaptive pose estimation. In Proceedings of European Conference on Computer Vision (ECCV), pages 603–620. Springer, 2022.

[10] S. Laine and T. Aila. Temporal ensembling for semi-supervised learning. arXiv preprint arXiv:1610.02242, 2016.

[11] C. Li and G. H. Lee. From synthetic to real: Unsupervised domain adaptation for animal pose estimation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 1482–1491, 2021.

[12] J. Li, S. Bian, A. Zeng, C. Wang, B. Pang, W. Liu, and C. Lu. Human pose regression with residual log-likelihood estimation. In Proceedings of the IEEE/CVF international conference on computer vision, pages 11025–11034, 2021.

[13] Y. Liu, Y. Tian, Y. Chen, F. Liu, V. Belagiannis, and G. Carneiro. Perturbed and strict mean teachers for semi-supervised semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4258–4267, 2022.

doi:10.6342/NTU202404439

[14] M. Long, Y. Cao, J. Wang, and M. Jordan. Learning transferable features with deep adaptation networks. In International conference on machine learning, pages 97–105. PMLR, 2015.

[15] W. Mao, Y. Ge, C. Shen, Z. Tian, X. Wang, Z. Wang, and A. v. den Hengel. Poseur: Direct human pose regression with transformers. In Proceedings of European Conference on Computer Vision (ECCV), pages 72–88. Springer, 2022.

[16] T. Miyato, S.-i. Maeda, M. Koyama, and S. Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. IEEE transactions on pattern analysis and machine intelligence, 41(8):1979–1993, 2018.

[17] J. Mu, W. Qiu, G. D. Hager, and A. L. Yuille. Learning from synthetic animals. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12386–12395, 2020.

[18] A. Peláez-Vegas, P. Mesejo, and J. Luengo. A survey on semi-supervised semantic segmentation. arXiv preprint arXiv:2302.09899, 2023.

[19] Q. Peng, C. Zheng, and C. Chen. Source-free domain adaptive human pose estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 4826–4836, 2023.

[20] M. V. Pusic, K. Boutis, R. Hatala, and D. A. Cook. Learning curves in health professions education. Academic Medicine, 90(8):1034–1042, 2015.

[21] M. V. Pusic, K. Boutis, S. A. Santen, and W. Cutrer. How does master adaptive learning ensure optimal pathways to clinical expertise? The Master Adaptive Learner, page 174, 2019.

doi:10.6342/NTU202404439

[22] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb. Learning from simulated and unsupervised images through adversarial training. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2107–2116, 2017.

[23] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Advances in neural information processing systems, 33:596–608, 2020.

[24] J. Stenum, K. M. Cherry-Allen, C. O. Pyles, R. D. Reetzke, M. F. Vignos, and R. T. Roemmich. Applications of pose estimation in human health and performance across the lifespan. Sensors, 21(21):7315, 2021.

[25] B. Sun, J. Feng, and K. Saenko. Return of frustratingly easy domain adaptation. CoRR, abs/1511.05547, 2015.

[26] B. Sun and K. Saenko. Deep coral: Correlation alignment for deep domain adaptation. In Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14, pages 443–450. Springer, 2016.

[27] K. Sun, Y. Zhao, B. Jiang, T. Cheng, B. Xiao, D. Liu, Y. Mu, X. Wang, W. Liu, and J. Wang. High-resolution representations for labeling pixels and regions. arXiv preprint arXiv:1904.04514, 2019.

[28] A. Tarvainen and H. Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems, 30, 2017.

doi:10.6342/NTU202404439

[29] B. Xiao, H. Wu, and Y. Wei. Simple baselines for human pose estimation and tracking. In Proceedings of the European conference on computer vision (ECCV), pages 466–481, 2018.

[30] Y. Xu, J. Zhang, Q. Zhang, and D. Tao. Vitpose: Simple vision transformer baselines for human pose estimation. Advances in Neural Information Processing Systems, 35:38571–38584, 2022.

[31] Y. Xu, J. Zhang, Q. Zhang, and D. Tao. Vitpose+: Vision transformer foundation model for generic body pose estimation. arXiv preprint arXiv:2212.04246, 2022.

[32] H. Yan, Y. Ding, P. Li, Q. Wang, Y. Xu, and W. Zuo. Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2272–2281, 2017.

[33] L. Yang, L. Qi, L. Feng, W. Zhang, and Y. Shi. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7236–7246, 2023.

[34] S. Yang, Z. Quan, M. Nie, and W. Yang. Transpose: Keypoint localization via transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 11802–11812, 2021.

[35] Y. Yuan, R. Fu, L. Huang, W. Lin, C. Zhang, X. Chen, and J. Wang. Hrformer: High-resolution vision transformer for dense predict. Advances in neural information processing systems, 34:7281–7293, 2021.

[36] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412, 2017.

doi:10.6342/NTU202404439

[37] Y. Zhang, H. Zhang, B. Deng, S. Li, K. Jia, and L. Zhang. Semi-supervised models are strong unsupervised domain adaptation learners, 2021.

[38] Z. Zhao, T. Wang, S. Xia, and Y. Wang. Hand-3d-studio: A new multi-view system for 3d hand reconstruction. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 2478–2482. IEEE, 2020.

doi:10.6342/NTU202404439

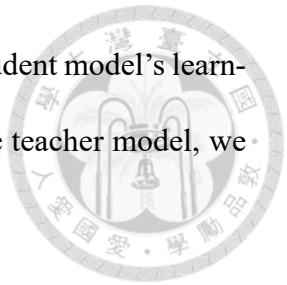
Appendix A — Analogous Explanation of the Sudden Learning Acceleration at the Start of Domain Adaptation Using the Human Learning Curve

The human learning curve can be divided into three phases [20, 21]. The first phase, known as the latent phase, involves acquiring basic knowledge. During this phase, the learning speed is slow. Once a foundation of knowledge is established, the learning enters the fast phase. In this phase, the model builds upon the foundation to learn more advanced concepts. The learning speed increases significantly. Finally, the learning enters the asymptotic phase. In this phase, the learning speed slows as simple knowledge in the field becomes scarce, leaving only the more difficult concepts to learn.

In UDA, the goal of the model is to acquire knowledge of the target domain. In the pretrain phase aimed at learning source knowledge, the model also acquires a basic understanding of the target domain. We can interpret the model's progression at the asymptotic phase for the source domain during pretraining as its progression through the latent phase for the target domain. Consequently, when domain adaptation begins and the model is exposed to a large amount of target data, we can consider the model to have entered the

doi:10.6342/NTU202404439

rapid phase of learning target domain knowledge. At this point, the student model's learning speed is remarkably high. To maintain the leading position of the teacher model, we propose the Adaptive Mean Teacher training strategy.



doi:10.6342/NTU202404439