Training-Free Model Merging for Multi-target Domain Adaptation

Wenyi Li[*1], Huan-ang Gao[*1], Mingju Gao[1], Beiwen Tian[1], Rong Zhi[2], and Hao Zhao[†1]

[1] Institute for AI Industry Research (AIR), Tsinghua University
[2] Mercedes-Benz Group China Ltd.
liwenyi19@mails.ucas.ac.cn, zhaohao@air.tsinghua.edu.cn

Abstract. In this paper, we study multi-target domain adaptation of scene understanding models. While previous methods achieved commendable results through inter-domain consistency losses, they often assumed unrealistic simultaneous access to images from all target domains, overlooking constraints such as data transfer bandwidth limitations and data privacy concerns. Given these challenges, we pose the question: How to merge models adapted independently on distinct domains while bypassing the need for direct access to training data? Our solution to this problem involves two components, merging model parameters and merging model buffers (*i.e.*, normalization layer statistics). For merging model parameters, empirical analyses of mode connectivity surprisingly reveal that linear merging suffices when employing the same pretrained backbone weights for adapting separate models. For merging model buffers, we model the real-world distribution with a Gaussian prior and estimate new statistics from the buffers of separately trained models. Our method is simple yet effective, achieving comparable performance with data combination training baselines, while eliminating the need for accessing training data. Project page: https://air-discover.github.io/ModelMerging.

Keywords: Multi-target Domain Adaptation · Mode Connectivity · Model Averaging

1 Introduction

Scene understanding models need to perform reliably across various domains (*i.e.*, diverse lighting [89], weather, and urban landscapes) to be truly useful for autonomous driving worldwide. The typical approach of supervised learning, however, relies heavily on costly pixel-level annotations by humans, which significantly hampers the scalability of these segmentation models. As such, the study of multi-target domain adaptation [20, 34, 55] (MTDA) is becoming increasingly significant. This area of research focuses on devising strategies to simultaneously utilize large-scale unlabeled real-world data from multiple domains along with

[*] [*] Indicates Equal Contribution. [†] Indicates Corresponding Author.

labeled synthetic data during training [78], providing a scalable approach to enhance the robustness of these models.

MTDA presents a greater challenge compared to traditional single-target domain adaptation (STDA, illustrated in Fig.1(a)), due to the inherent difficulty in developing a single model that performs effectively across multiple target domains. To tackle this, previous approaches [20, 34, 48, 55] have employed consistency learning among various expert models and online knowledge distillation to construct a domain-generic student model. Nonetheless, a significant limitation of these methods is their reliance on simultaneous access to all target data, as illustrated in Fig. 1(b), which is usually impractical. This impracticality stems partly from data transfer cost constraints, as a dataset comprising thousands of images can balloon to hundreds of gigabytes. Moreover, data privacy concerns are significant, with potential restrictions on the sharing or transferring of such data due to policies.

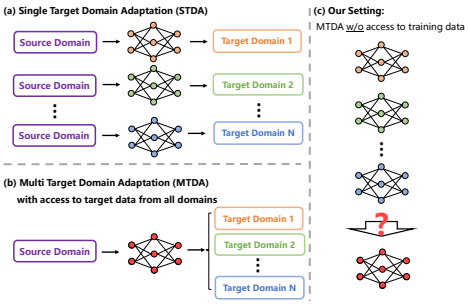


Fig. 1: Comparison of Domain Adaptation Settings. (a) Single Target Domain Adaptation (STDA) focuses on leveraging labeled synthetic data and unlabeled data from a single target domain together for optimal performance in that target domain. (b) Multi-target Domain Adaptation (MTDA) with data access involves utilizing data from target domains together to train a single model capable of excelling across all these domains. (c) MTDA without direct access to training data, employing model merging to enhance robustness.

Facing these challenges, in this paper, we consider a novel problem setting illustrated in Fig. 1(c). Our primary focus remains on multi-target domain adaptation, but instead of accessing data from multiple domains, we gain access to *models* that have been independently adapted to each of them. Our objective is to integrate these models into a single, effective model for various domains.

But how to merge multiple models into one while preserving their respective abilities on their respective domains? Our solution to this problem has two main components. **(I) Merging model parameters.** A straightforward method involves averaging the weights of models as a means of integration. Nevertheless, this approach raises questions about its underlying mechanisms and its reliability across different scenarios. In our research, we conduct careful investigation into the conditions under which such merging is effective, and when it is not, by examining the concept of *linear mode connectivity* of models. Through meticulous analysis, we find that pretraining significantly contributes to enhancing this linear connectivity among trained models. **(II) Merging model buffers.** We identify the importance of model buffers, *i.e.*, batch normalization (BN) layer statistics, which capture domain-specific characteristics in our multi-target domain adaptation setting. Leveraging the Gaussian prior assumption of BN

layers, we estimate new means and variances for the merged layers based on the statistics of separately trained models.

Our method, simple yet effective, demonstrates notable improvements in performance compared to a variety of baselines. For instance, when we apply our model merging technique to the state-of-the-art STDA method [29] using a ResNet101 [24] backbone, we observe a substantial increase of +5.6% in harmonic mean of mIoUs for individual domains. Remarkably, this level of performance rivals that achieved by baseline methods that involve training with multiple combined datasets, even when data availability is constrained. Furthermore, our technique outperforms previous top-performing multi-target domain adaptation methods that utilize explicit consistency training by a considerable margin, underscoring the critical role of exploring model connectivities. In summary, our contributions are as follows:

- We conduct a systematic exploration of mode connectivity in domain-adapted scene parsing models, revealing the underlying conditions of when model merging works.
- We introduce a model merging technique including parameter merging and buffer merging for multi-target domain adaptation tasks, applicable to *any* single-target domain adaptation model.
- Our approach achieves performance comparable to that of training with multiple combined datasets, even when data availability is constrained.

2 Related Works

2.1 Domain Adaptation for Semantic Segmentation

Unsupervised domain adaptation [14] helps to reduce annotation costs in pixel-wise semantic segmentation by utilizing unlabeled real-world images alongside labeled synthetic samples.

Single-target Domain Adaptation. The challenge of STDA mainly involves how to effectively reduce the distribution gap between labeled synthetic data and unlabeled single-domain real-world images. Methods to tackle this issue include domain-invariant representation learning [37, 46, 81], adversarial training [15,21,25,26,47,62,71,75], and teacher-student self-training [17,18,22,28–30]. Among these approaches, the last one has proven to be highly successful. It leverages synthetic data [7, 16, 45, 68] as source domains, generates pseudo labels [42] for target domains and enforces consistency regularization over data augmentation strategies [2, 8, 28, 31, 41, 51, 69, 86, 90, 91]. Despite STDA being a broadly studied topic, it has limitations for real-world applications like autonomous driving, where vehicles encounter diverse road scenarios beyond a single domain.

Multi-target Domain Adaptation. MTDA presents greater challenges than STDA, particularly in terms of integrating domain-specific knowledge from multiple sources into a single model. While initial efforts in this area have focused on classification tasks [20, 55, 83], MTDA has also been explored for semantic segmentation [34, 55], employing strategies like inter-domain consistency

regularization and online knowledge distillation. Nonetheless, the application of these explicit inter-domain regularization techniques limits the use of certain STDA strategies, like multi-resolution training, due to the excessive GPU memory demands associated with managing multiple student models. Consequently, these approaches [34, 55] significantly underperform when compared to STDA methodologies [28–30]. Moreover, they rely on the impractical assumption of having simultaneous access to data from all target domains. So we ask the question: is it possible to train several models under the best STDA setting and merge them without access to training data (*i.e.*, in a zero-shot way [23])?

2.2 Multi-target Learning with Constrained Data Assumption

Federated Learning (FL). While FL [50] focuses on the *joint training* of models, our research studies the direct merging of models *post training*. Representative FL methods confront challenges including communication cost [53], data heterogeneity [44], and support for device capabilities [27] in a distributed setting. In contrast, our approach diverges in methodology by leveraging the advancements in large-scale, centralized pre-training of vision models. We answer the question that *under which specific conditions* can we simply *merge* pre-trained models adapted to distinct data domains. **Federated MTDA.** In this setting [82], the distributed client data is unlabeled, and a centralized labeled dataset is available on the server. In our STDA stage, labeled dataset is accessible for each domain (comparable to clients). **Class-Incremental Domain Adaptation (CIDA).** CIDA [40] learns novel target-domain classes in a domain shift-aware manner, while they do not touch *multiple target* domains.

2.3 Mode Connectivity for Neural Networks

Mode connectivity refers to the fact where local minima in the loss functions of neural networks can be connected through a curve in the parameter space, along which the performance does not experience significant deterioration [19, 67]. Linear mode connectivity is a stronger constraint requiring that the convex combination of two minima stay in the same loss basin, which is closely related to the lottery ticket hypothesis [12] and has direct implications for continual learning [52]. Exploiting the property of linear mode connectivity, several works [32,35,54,58,74,79] integrate weights of neural networks for a more robust merged model. In our study, we also observe a similar phenomenon in models adapted to different domains, and we conduct an empirical analysis to trace the origin of it. We also identify the importance of batch normalization layer statistics in the multi-target domain adaptation context, as the diverse characteristics inherent in multiple domains are distinctly captured in these statistical layers. Different from the aforementioned work focused on language or image inputs from subsets of the same dataset, we propose a novel method to effectively merge these statistics.

3 Methodology

3.1 Overview

We introduce a novel pipeline designed to address the challenging problem of multi-target domain adaptation in semantic segmentation, which is illustrated in Fig. 2. What sets our approach apart from previous methods [3, 20, 34, 55, 76, 85] is the elimination of the impractical assumption that images of all target domains concurrently accessible during the adaptation phase. Instead, our pipeline comprises two distinct phases: a single-target domain adaptation phase and a model merging phase. In the first phase, we train models adapted for individual target domains separately, while the second phase focuses on merging these adapted *models* together to create a robust model, **without access to any training data.**

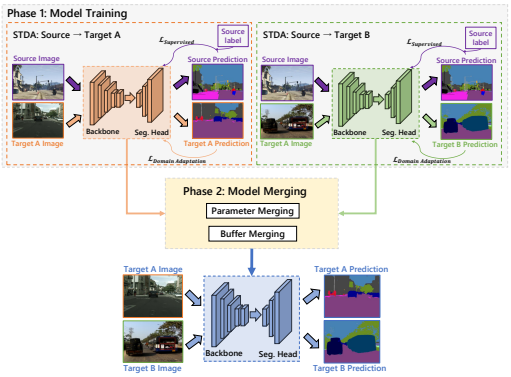


Fig. 2: Overview of Two-stage Pipeline of Our Proposed Multi-target Domain Adaptation Solution. After training STDA methods on separate domains, we integrate models together using our proposed merging techniques.

Our primary focus lies on the proposed model merging phase. In the initial phase, we simply adopt the state-of-the-art unsupervised domain adaptation approach, HRDA [29], leveraging various backbone architectures such as ResNet [24] and vision transformer [80]. Our approach encompasses two critical components of the models: parameters (*i.e.*, weights and biases for the learnable layers) and buffers (*i.e.*, running statistics for the normalization layers). We provide detailed explanations of the merging techniques and the underlying motivations for these techniques in Sec. 3.2 and Sec. 3.3 for merging parameters and buffers, respectively.

3.2 Merging Parameters

Permutation-based methods degenerate. In fact, the idea of merging the weights and biases of learnable layers between models has been a frontier research area. Among these efforts, a particularly promising line of research has emerged, known as permutation-based methods [1,11,36,57]. These methods operate under the assumption that, when accounting for all potential permutation symmetries of units in hidden layers of neural networks, the loss landscapes typically form a single basin. Therefore, when merging model parameters $\mathbf{\Theta}_A$ and $\mathbf{\Theta}_B$, the primary goal of these methods is to find a set of permutation transformations

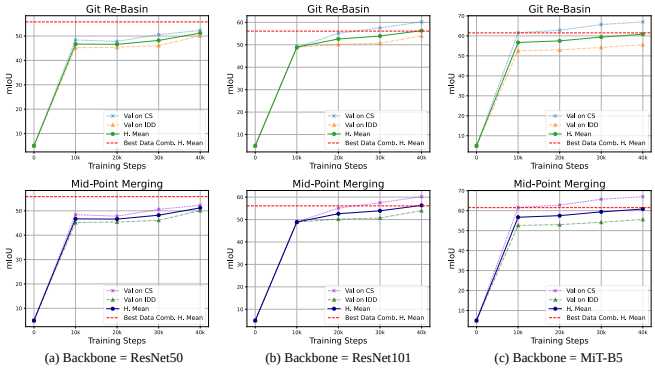


Fig. 3: Results of Git Re-Basin and Mid-Point Merging on Different Backbones. In our domain adaptation scenario, Git Re-Basin [1] reduced to a straightforward mid-point merging approach.

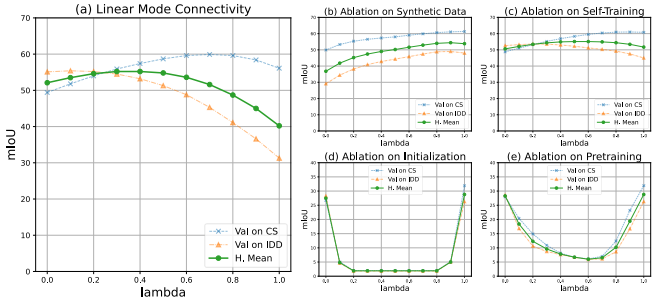


Fig. 4: Empirical Analysis for Linear Mode Connectivity. (a) Exploring the linear mode connectivity of two trained ResNet101 backbones targeted at two different domains. (b-e) Ablation studies on synthetic data, self-training architecture, initializaiton weights and pretrained weights to find the cause of the linear mode connectivity.

$\pi(\cdot)$ that ensures $\pi(\mathbf{\Theta}_B)$ is functionally equivalent to $\mathbf{\Theta}_B$, while also residing within an approximately convex basin near the reference model $\mathbf{\Theta}_A$. After that, with a simple mid-point merging ($\lambda = \frac{1}{2}$ in Eq. 1), we can acquire a merged model $\mathbf{\Theta}'$ that exhibits better generalization ability than single models,

$$\mathbf{\Theta}' = \lambda \mathbf{\Theta}_A + (1 - \lambda)\pi(\mathbf{\Theta}_B). \tag{1}$$

In our scenario, both $\mathbf{\Theta}_A$ and $\mathbf{\Theta}_B$ are trained during the first phase employing identical network architectures [29] and utilizing the same synthetic images and labels [60]. However, they are adapted to samples from distinct domains for semantic segmentation [9, 73]. Our initial attempt involved employing a representative permutation-based method known as Git Re-Basin [1]. This method transforms the task of finding permutation-symmetric transformations into a linear assignment problem (LAP), for which efficient and practical algorithms are

available. Surprisingly, in our experimental setup, Git Re-Basin's performance equaled that of a simple mid-point merging across all network architectures, including ResNet50 [24], ResNet101 [24], and MiT-B5 [80], and the results of mid-point merging (also results of Git Re-Basin [1]) are illustrated in Fig. 3. Further investigation revealed that the permutation transformation discovered by Git Re-Basin [1] remained identical permutation through iterations of solving LAP, suggesting that in our domain adaptation scenario, Git Re-Basin [1] reduced to a straightforward mid-point merging approach.

Empirical analysis of linear mode connectivity. We further investigate the above degeneration problem through the lens of *linear mode connectivity* [13, 19, 52]. Specifically, we use continuous curve $\phi(\lambda) : [0, 1] \rightarrow \mathbb{R}^{|\mathbf{\Theta}|}$ to connect $\mathbf{\Theta}_A$ and $\mathbf{\Theta}_B$ in the parameter space. In this specific case, we consider a linear path as follows,

$$\phi(\lambda) = \lambda\mathbf{\Theta}_A + (1 - \lambda)\mathbf{\Theta}_B. \tag{2}$$

After defining the curve connecting the models to be merged, we traverse along the curve and evaluate the performance of the interpolated models. To gauge the effectiveness of these models in adapting to the two specified target domains, denoted as $\mathcal{D}_1$ and $\mathcal{D}_2$, respectively, we employ the Harmonic Mean as our primary evaluation metric,

$$\text{Harmonic Mean} = \frac{2 \cdot \text{mIoU}_{\mathcal{D}_1} \cdot \text{mIoU}_{\mathcal{D}_2}}{\text{mIoU}_{\mathcal{D}_1} + \text{mIoU}_{\mathcal{D}_2}}. \tag{3}$$

We select harmonic mean as the metric as it gives more weight to smaller values, which corresponds to the worst-case performance among various cities over the world. It effectively penalizes scenarios where performance in one domain (e.g., in well-developed big cities) is disproportionately high while other domains (e.g., in a rural third-world town) have low performance. We believe this is aligned with the initial goal of the multi-target domain adaptation task. On the contrary, in an extreme case, scoring 100% in one city and zero on all other three cities still leads to 25% performance using arithmetic mean numbers, which we believe is of limited value in the multi-target domain adaptation problem.

The evaluation results of the interpolation are depicted in Fig. 4(a). 'CS' and 'IDD' denotes target datasets Cityscapes [9] and Indian Driving Dataset [73], respectively. Notably, it is evident that the two models from the first phase are already linearly mode connected without permutation, as the harmonic mean of the interpolated models outperforms the performance of individual models for both domains.

Understanding the cause of linear mode connectivity. Given the aforementioned revelation, we inquire: What is the underlying reason behind the linear mode connectivity property observed in previous domain adaptation methods? Subsequently, we conduct ablation experiments to investigate several constant factors during the training of Θ_A and Θ_B in the first phase.

Synthetic Data. The utilization of the same synthetic data may serve as a bridge between the two domains. To assess this, we partition the training

data from synthetic set [60] into two distinct non-overlapping subgroups, each comprising 30% of the original training samples. During the partitioning process, we group images with identical scene identifiers provided by the synthetic dataset into the same subgroup, while scenes with significant differences are placed in separate subgroups. We train two single-target domain adaptation models, using the source domain provided by these two distinct subgroups, while setting the target domain as the CityScapes [9] dataset. We subsequently investigate the linear mode connectivity of the two resulting models. The results, displayed in Fig. 4(b), reveal that there is no notable decline in performance along the linear curve connecting the two resulting models within the parameter space. This observation suggests that the usage of the same synthetic data is not the primary factor influencing the linear mode connectivity.

Self-training Architecture. The utilization of a teacher-student self-teaching architecture [66] may confine resulting models to the same basin within the loss landscape. To assess this possibility, we disable the exponential moving average (EMA) update for the teacher models. Instead, we copy the student weights to teacher models during each iteration. Subsequently, we proceed to train two single-target domain adaptation models, utilizing GTA [60] as the source domain and Cityscapes [9] and IDD [73] as the target domains, respectively. We then investigate the linear curve that connects the two resulting models within the parameter space, and the outcomes are presented in Fig. 4(c). We can see that the linear mode connectivity property remains intact.

Initialization and Pretraining. The practice of initializing the backbone with the same pretrained weights can potentially place it in a basin that is challenging to escape from during the training process. To examine this potential scenario, we initialize two separate backbones with distinct weights and then proceed with domain adaptation targeting the Cityscapes [9] and IDD [73] domains. During the evaluation of the linear interpolated models between the two converged models, we observe a notable deterioration in performance, as shown in Fig. 4(d). To gain a deeper understanding of the underlying factors, we explore whether it is the identical initial weight or the pretraining process that contributes to this effect. We initiate two backbones with the same weights but without pretraining and conduct the experiment once more. Interestingly, we still encounter a substantial performance barrier along the linear connecting curve in the parameter space, as reported in Fig. 4(e). This implies that it is the pretraining process that plays a pivotal role in facilitating linear mode connectivity in the fine-tuned models.

Summary. Our empirical analysis highlight that when commencing from identical pretrained weights, domain adaptation models can effectively transition to diverse target domains while still maintaining linear mode connectivity in the parameter space. Consequently, a straightforward mid-point merge between these trained models could generate models with robustness in both domains.

3.3 Merging Buffers

Buffers, namely running mean and variance for batch normalization (BN) layers, bear a close association with domains as they encapsulate domain-specific characteristics. The question of how to effectively merge buffers when merging models is usually overlooked, as existing methods [32, 35, 74, 79] predominantly address the merging of two models trained on separate subsets within the same domain. These approaches are reasonable, as buffers from any given model can be regarded as an unbiased estimation of the entire population, albeit solely derived from randomized data subsamples. Nevertheless, in our specific problem context, we are investigating the merging of two models trained in completely distinct target domains, rendering the buffer merging problem non-trivial anymore.

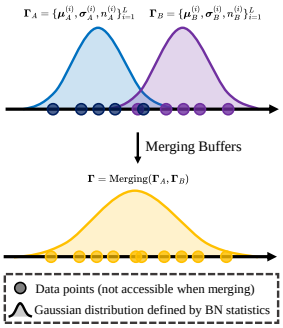


Fig. 5: Illustration on Merging Statistics in Batch Normalization (BN) Layers.

Since we assume no access to any form of training data during the merging phase of model A and model B, our available information is confined to the set of buffers $\mathbf{\Gamma} = \{\boldsymbol{\mu}^{(i)}, \boldsymbol{\sigma}^{(i)}, n^{(i)}\}_{i=1}^{L}$. Here, L represents the number of BN layers, while $\boldsymbol{\mu}^{(i)}$, $\boldsymbol{\sigma}^{(i)}$, and $n^{(i)}$ respectively signify the running statistics for mean, standard deviation, and the number of tracked batches for the i-th layer. The statistics of the resulting BN layer is given as,

$$
\begin{aligned}
n^{(i)} &= n_A^{(i)} + n_B^{(i)}, \\
\boldsymbol{\mu}^{(i)} &= \frac{1}{n^{(i)}}(n_A^{(i)}\boldsymbol{\mu}_A^{(i)} + n_B^{(i)}\boldsymbol{\mu}_B^{(i)}), \\
[\boldsymbol{\sigma}^{(i)}]^2 &= \frac{1}{n^{(i)}}\Big(n_A^{(i)}[\boldsymbol{\sigma}_A^{(i)}]^2 + n_B^{(i)}[\boldsymbol{\sigma}_B^{(i)}]^2 \\
&\quad + n_A^{(i)}[\boldsymbol{\mu}^{(i)} - \boldsymbol{\mu}_A^{(i)}]^2 + n_B^{(i)}[\boldsymbol{\mu}^{(i)} - \boldsymbol{\mu}_B^{(i)}]^2\Big).
\end{aligned}
\tag{4}
$$

The rationale behind Eq. 4 can be elucidated as follows. BN layers are introduced to alleviate the issue of *internal covariate shift* [4, 33, 65], where the means and variances of inputs undergo changes as they pass through internal learnable layers. In this context, our fundamental consideration is that subsequent learnable layers anticipate the output of the merged BN layer to follow a normal distribution. Since the resulting BN layer hold the the inductive bias of inputs conforming to a Gaussian prior, we estimate $\boldsymbol{\mu}^{(i)}$ and $[\boldsymbol{\sigma}^{(i)}]^2$ from what we get from $\mathbf{\Gamma}_A$ and $\mathbf{\Gamma}_B$. As depicted in Figure 5, we are provided with two sets of means and variances of data points sampled from this Gaussian prior, along with the sizes of these sets. We leverage these values collectively to estimate the parameters of this distribution.

Extending to more domains. When extending the merging method to $m (m \geq 2)$ Gaussian distributions, the number of tracked batches $n^{(i)}$, the

weighted average of the means $\boldsymbol{\mu}^{(i)}$ and the weighted average of the variances can be calculated as follows.

$$\begin{aligned} n^{(i)} =& n_1^{(i)} + n_2^{(i)} + \cdots + n_M^{(i)}, \\ \boldsymbol{\mu}^{(i)} =& \frac{1}{n^{(i)}}(n_1^{(i)}\boldsymbol{\mu}_1^{(i)} + n_2^{(i)}\boldsymbol{\mu}_2^{(i)} + \cdots + n_M^{(i)}\boldsymbol{\mu}_M^{(i)}), \\ \boldsymbol{\sigma}^2 =& \frac{\sum_{j=1}^{M} n^{(i)}(\boldsymbol{\sigma}_j^i)^2 + \sum_{j=1}^{M} n_j^i(\boldsymbol{\mu}_j^i - \boldsymbol{\mu}^i)^2}{\sum_{j=1}^{M} n_j^i}. \end{aligned} \tag{5}$$

4 Experiments

4.1 Datasets

In the context of multi-target domain adaptation experiments, we employ GTA [60] and SYNTHIA [61] as the synthetic dataset and the real-world datasets of Cityscapes [9], Indian Driving Dataset [73] (IDD), ACDC [64] and DarkZurich [63] as the target domains. For domain adaptation methods, training involves labeled source data and unlabeled target data from different domains. We employ our merging techniques to construct a model from the trained models, employing the discussed merging methods, all without the need for direct access to this data.

Synthetic Datasets. *GTA* [60] dataset comprises 24,966 synthetic images, each with a resolution of 1914×1052 pixels. These images are sourced from the video game GTA5 and come equipped with pixel-level annotations encompassing 19 categories, aligning with the annotation protocols of Cityscapes and IDD datasets. *SYNTHIA* [61] dataset comprises 9,400 rendered images, each with a resolution of 1280×760 pixels, generated from a virtual city.

Real-world Datasets. *Cityscapes* [9] is a real-world dataset featuring a collection of 5,000 street scenes captured in various European cities. These scenes have been meticulously labeled, classifying objects and elements into 19 distinct categories. In our experiments, we utilize a subset of this dataset, consisting of 2,975 images for training and an additional 500 images for validation purposes. *IDD* [73] offers greater diversity compared to Cityscapes, capturing unstructured traffic scenarios on Indian roads. It boasts a total of 10,003 images, with 6,993 designated for training, 981 for validation, and 2,029 for testing. *ACDC* [64] contains 1600 training, 406 validation, and 2000 test images, evenly distributed among adverse weather conditions, including fog, night, rain and snow. *Dark-Zurich* [63] contains 2416 training, 50 validation, and 151 test images specifically curated for nighttime scenarios.

4.2 Implementation Details

Our proposed model merging techniques are easy to implement. In the first phase of training single-target domain adaptation model, we leverage the state-of-the-art HRDA method [29]. We validate the effectiveness of our approach using a

range of image encoders as backbones, including ResNet50 [24], ResNet101 [24], and MiT-B5 [80], all pre-trained on ImageNet-1K [10]. Except for the backbone encoder and the factor under examination in Fig. 4, all other training hyper-parameters remain consistent with the original HRDA [29] implementation. In the second phase of model merging, we work directly with the state dictionaries of checkpoint files. For parameters, we perform a mid-point merging, and for buffers, we apply the formula presented in Eq. 4. When examining linear mode connectivity in Fig. 4, we evenly sample values of λ within the range of [0, 1] in Eq. 1, using a stride of 0.1 and including both endpoints.

Table 1: Performance Comparison of Our Method and Baselines. The mIoU (mean Intersection-over-Union) represents the average IoU across 19 categories. 'Enc.' denotes the encoder architecture, with 'R' representing ResNet101 and 'V' indicating MiT-B5. The 'Metric' column specifies whether evaluation was conducted on the Cityscapes ('C') or IDD ('I') dataset. The harmonic mean ('H'), representing adaptation ability across the two domains, is considered as the primary metric. Bold text highlights the best harmonic mean results, while underlined text indicates the second-best results. '†' signifies only merging backbones while keeping separate decode heads.

Method	Enc.	Metric	road	sidewalk	building	wall	fence	pole	light	sign	veg.	terrain	sky	person	rider	car	truck	bus	train	motor	bike	mIoU
Data Comb.	R	C	95.3	67.5	87.7	25.8	17.3	50.6	51.8	58.9	90.2	42.8	92.9	75.2	40.4	91.9	54.5	61.6	3.9	56.0	62.7	59.3
		I	95.8	22.3	73.0	29.2	12.8	34.2	26.0	54.6	82.2	36.2	95.3	67.0	64.4	80.3	70.2	60.9	0.0	74.8	34.1	53.3
		H	95.6	33.5	79.7	27.4	14.7	40.8	34.6	56.6	86.0	39.2	94.1	70.9	49.6	85.7	61.3	61.2	0.0	64.0	44.2	56.2
STDA (GTA → CS)	R	C	95.9	74.4	88.9	32.4	34.0	54.6	61.2	71.1	90.3	45.1	88.9	77.8	51.2	91.3	49.6	62.8	0.0	46.5	60.5	61.9
		I	90.7	30.6	64.8	14.6	17.5	28.9	17.9	18.8	80.1	17.5	92.3	62.3	51.6	74.6	40.2	36.1	0.0	67.7	35.2	44.3
		H	93.3	43.3	75.0	20.2	23.1	37.8	27.7	29.8	84.9	25.2	90.6	69.2	51.4	82.1	44.4	45.9	0.0	**55.1**	**44.5**	51.6
STDA (GTA → IDD)	R	C	81.4	29.1	79.0	22.1	30.4	41.7	45.0	42.4	88.9	39.0	92.0	72.6	43.1	80.3	31.1	55.3	7.9	22.4	34.4	49.4
		I	96.0	34.7	75.0	33.3	22.6	33.6	24.2	54.0	84.9	39.8	95.8	69.0	66.0	80.2	69.4	63.7	0.0	76.0	28.9	55.1
		H	88.1	31.6	76.9	26.6	25.9	37.2	31.5	47.5	86.8	39.4	**93.8**	70.8	52.1	80.2	43.0	**59.2**	0.0	34.6	31.4	52.1
Ours	R	C	91.8	60.0	86.3	39.3	31.0	47.6	53.9	53.6	89.8	46.4	91.8	75.5	48.6	89.9	54.1	60.5	25.2	31.6	39.9	58.8
		I	96.4	40.0	73.4	34.9	23.3	31.7	27.9	53.6	83.7	44.8	94.5	67.8	64.1	77.6	57.6	51.3	0.0	72.8	30.0	54.0
		H	94.0	48.0	79.3	**37.0**	26.6	38.1	36.8	53.6	86.6	45.6	93.2	71.4	55.2	83.3	55.8	55.5	0.0	44.0	34.3	56.3 (↑ 4.2)
Ours†	R	C	92.0	59.6	87.9	39.1	35.7	50.7	59.6	60.8	90.1	45.4	91.3	77.4	48.8	88.4	52.6	59.6	20.3	33.5	50.6	60.2
		I	96.9	41.1	74.0	34.6	25.1	34.5	29.0	50.6	85.8	48.1	95.8	69.6	65.3	79.3	61.8	55.2	0.0	74.4	33.3	55.5
		H	**94.4**	**48.7**	**80.4**	36.7	**29.5**	**41.1**	**39.0**	**55.2**	**87.9**	**46.7**	93.5	**73.3**	**55.9**	**83.6**	**56.8**	57.3	0.0	46.1	40.2	**57.7** (↑ 5.6)
Data Comb.	V	C	88.9	49.7	90.2	58.8	46.6	50.9	60.9	60.3	89.7	47.4	88.8	77.7	45.0	93.5	77.7	79.4	69.0	57.6	65.8	68.3
		I	94.0	32.2	77.9	44.6	30.3	42.6	33.2	53.0	80.6	24.8	91.2	71.0	67.2	81.2	73.4	66.1	0.0	75.6	24.6	56.0
		H	91.4	39.1	83.6	50.7	36.7	46.4	43.0	56.4	84.9	32.6	90.0	74.2	53.9	86.9	75.5	72.1	0.0	65.3	35.8	61.5
STDA (GTA → CS)	V	C	96.4	74.2	91.0	59.4	53.5	58.0	64.9	69.4	91.6	49.9	93.8	79.2	53.7	93.4	75.2	76.3	67.6	64.3	67.1	72.6
		I	85.8	10.0	72.1	30.7	27.4	34.0	32.9	53.8	79.9	36.4	95.2	65.2	54.2	80.2	47.2	48.8	0.0	72.4	34.9	50.6
		H	90.8	17.6	80.4	40.5	**36.2**	42.9	43.6	60.6	85.3	42.0	**94.5**	71.5	54.0	86.3	58.0	59.5	0.0	**68.1**	**45.9**	59.6
STDA (GTA → IDD)	V	C	87.2	30.0	88.9	53.1	35.3	52.0	56.6	49.1	89.7	47.3	88.4	74.9	40.0	91.1	76.7	64.2	29.5	15.2	30.5	57.9
		I	93.0	4.7	78.0	42.0	24.8	44.3	25.9	59.8	79.7	24.1	91.0	62.7	59.6	78.8	71.8	75.6	0.0	63.1	16.7	52.4
		H	90.0	8.2	83.1	**46.9**	29.1	**47.8**	35.5	53.9	84.4	31.9	89.7	68.3	47.9	84.5	**74.2**	69.4	0.0	24.5	21.6	55.0
Ours	V	C	93.6	57.8	89.9	58.5	41.8	55.5	58.8	56.5	90.8	52.2	92.0	77.5	46.3	93.3	76.6	75.7	53.1	44.9	57.4	67.0
		I	93.6	18.9	76.1	35.2	29.4	38.7	32.7	58.1	82.2	41.2	93.9	72.5	63.3	81.0	63.7	66.6	0.0	75.2	34.9	55.6
		H	**93.6**	**28.5**	82.4	44.0	34.5	45.6	42.0	57.3	**86.3**	**46.0**	92.9	**74.9**	53.5	86.7	69.6	70.9	0.0	56.2	43.4	60.8 (↑ 1.2)
Ours†	V	C	94.1	60.9	90.6	59.5	46.9	56.8	63.7	62.8	91.4	52.3	93.6	78.5	50.6	93.3	76.8	79.4	67.9	58.7	66.3	70.7
		I	93.0	15.7	77.0	36.6	29.2	41.1	35.8	62.1	80.2	36.7	92.3	70.6	62.7	81.9	65.3	69.5	0.0	71.6	29.2	55.3
		H	93.5	24.9	**83.2**	45.3	36.0	47.7	**45.8**	**62.4**	85.4	43.1	93.0	74.3	**56.0**	**87.2**	70.6	**74.1**	0.0	64.5	40.5	**62.1** (↑ 2.5)

4.3 Comparison with Baseline Methods

We present a comparison of our model merging-based multi-target domain adaptation approach with several baseline methods in Tab. 1. In this experiment, we evaluate our method using GTA [60] as the source domain and two target domains, Cityscapes [9] and IDD [73]. However, our approach can easily scale to handle a greater number of target domains, should the need arise. Additionally,

we assess the performance when SYNTHIA [61] serves as the source domain, and the results for this scenario are presented in the supplementary material.

Baseline methods include Data Combination ("Data Comb.") approaches, where a single domain adaptation model is trained on a mixture of data from two target domains. Note that these are only presented for reference as they contradict our considerations related to data transfer bandwidth and privacy issues. We also include Single-Target Domain Adaptation ("STDA") baselines, which involve training a single domain adaptation model for one domain and assessing its generalization to both domains. We evaluate our proposed methods (labeled as "Ours") which involve merging all models or merging only image backbones while maintaining separate decoding heads.

Results obtained using convolutional-based encoder architecture ResNet101 [24] and transformer-based architecture MiT-B5 [80] are presented in Tab. 1. Our method demonstrates a notable improvement of **+4.2%** and **+1.2%** in harmonic mean when applied to the ResNet101 [24] and MiT-B5 [80] backbones, respectively, compared to the strongest single-target domain adaptation model. Notably, this level of performance (56.3% harmonic mean with ResNet101 [24]) is already on par with data combination methods (56.2% harmonic mean), and we achieve it without requiring access to any training data. Furthermore, we explore a more relaxed setting where only the encoder backbone is merged while decoding heads are separated for various downstream domains. This is a feasible approach as the parameters of image backbones are typically orders of magnitude larger than those of decoding heads. Remarkably, this configuration results in a substantial performance improvement of **+5.6%** and **+2.5%** in harmonic mean for two backbones, respectively. We also find that our merging-based method consistently achieves the best harmonic means across most categories, indicating its ability to enhance adaptation globally instead of biasing to certain categories.

Table 2: Comparison of Our Method with State-of-the-Art Approaches. Prior MTDA methods used different training methods from ours, only for reference. '†' signifies results reproduced by us.

Setting	Method	Backbone	Metric		
			CS	IDD	H. Mean
STDA (GTA → X)	BDL [46]	R101	41.1	-	-
	AdaptSeg [71]	R101	42.4	-	-
	CLAN [49]	R101	43.2	-	-
	ADVENT [75]	R101	43.8	-	-
	MaxSquare [6]	R101	44.3	-	-
	AdaptPatch [72]	R101	44.9	-	-
	CBST [92]	R38	45.9	-	-
	IntraDA [56]	R101	46.3	-	-
	DACS [70]	R101	52.1	-	-
	DAFormer [28]	R101	56.0	-	-
	CorDA [77]	R101	56.6	-	-
	ProDA [87]	R101	57.5	-	-
	HRDA† [29]	R101	61.9	-	-
	DDB [5]	R101	62.7	-	-
	DAFormer [28]	MiT-B5	68.3	-	-
	HRDA† [29]	MiT-B5	72.6	-	-
DG	Yue et al. [84]	R101	42.1	42.8	42.4
	Kundu et al. [39]	R101	53.4	-	-
MTDA	MTDA-ITA [20]	R101	40.3	41.2	40.8
	MT-MTDA [55]	R101	43.2	44.0	43.6
	CCL [34]	R101	45.0	46.0	45.5
	Coast [88]	R101	47.1	49.3	48.2
MTDA (Merging)	Ours	R101	58.8	54.0	56.3 (↑8.1)
	Ours	MiT-B5	67.0	55.6	**60.8** (↑12.6)

4.4 Comparison with State-of-the-Arts

We begin by comparing our method with the single-target domain adaptation (STDA) on the GTA→Cityscapes task, as shown in Tab. 2. It is worthy to note that our method can be applied to any of these methods, provided they adapt to different domains using the same pretrained weights. This allows us to generalize to all target domains using a single model while keeping the relatively superior performance of STDA methods. We also compare our methods with domain

generalization approaches in Tab. 2, which aim to generalize a model trained on a source domain to multiple unseen target domains. Our approach stands out by achieving superior performance without requiring additional tricks, just by means of exploiting parameter space mode connectivity.

In the realm of multi-target domain adaptation, our method also stands out. We eliminate the need for explicit inter-domain consistency regularization or knowledge distillation of multiple student models, enabling techniques from STDA methods like multi-resolution training to transfer to MTDA tasks. Therefore, we witness a significant improvement over the best published results of MTDA, while eliminating the need of access to training data.

4.5 Extending to More Target Domains

In this section, we expand the application of our model merging technique to encompass four distinct target domains: Cityscapes [9], IDD [73], ACDC [64], and DarkZurich [63]. Each of these domains presents unique challenges and characteristics: Cityscapes [9] captures European urban settings, IDD [73] focuses on Indian road scenes, ACDC [64] is tailored to adverse weather conditions such as fog, rain, or snow, and DarkZurich [63] addresses night road scenes. We conduct a thorough evaluation of models that are trained separately for each domain as well as models created through the merging of these individually adapted models. The effectiveness of these approaches is quantified by reporting the harmonic mean of their performance across these diverse domains. All results are presented in Tab. 3.

Table 3: Application of Our Model Merging Techniques Across Four Target Domains. The datasets Cityscapes [9], IDD [73], ACDC [64], and DarkZurich [63] are represented by 'C', 'I', 'A', and 'D', respectively. The mIoU of each dataset and the harmonic mean (H) is reported.

Model #	Merging of				Metric					
	C	I	A	D	C	I	A	D	H	Δ
1	✓				61.3	47.0	42.1	16.1	32.4	-1.0%
2		✓			51.2	55.4	37.6	16.2	31.8	-1.7%
3			✓		44.9	38.5	42.4	18.6	31.9	-1.6%
4				✓	43.6	38.2	41.6	21.6	33.5	-
5	✓	✓			58.8	54.0	41.9	17.4	34.2	+0.8%
6	✓		✓		55.1	45.2	45.2	20.8	36.2	+2.8%
7	✓			✓	57.5	45.2	45.8	24.2	38.9	+5.5%
8		✓	✓		47.1	49.7	43.6	21.5	36.1	+2.7%
9		✓		✓	50.0	49.6	46.2	25.0	**39.3**	+5.8%
10			✓	✓	45.5	41.3	45.3	22.0	35.2	+1.7%
11	✓	✓	✓		53.0	49.3	44.5	20.0	35.9	+2.4%
12	✓	✓		✓	55.1	49.3	45.4	21.1	37.1	+3.6%
13	✓		✓	✓	51.9	43.8	46.3	22.1	36.7	+3.3%
14		✓	✓	✓	47.6	46.7	46.1	22.5	36.8	+3.4%
15	✓	✓	✓	✓	51.6	47.3	45.8	21.1	36.4	+3.0%

Our proposed model merging techniques demonstrate a significant improvement in performance, as illustrated in Tab. 3. While we use the method with the highest harmonic mean from separately trained models as our baseline for comparison, all the approaches based on model merging outperform it, with gains as substantial as +5.8%. Furthermore, despite the increasing complexity in merging models from multiple, diverse domains, we observe that the overall performance across all domains does not suffer any notable decline.

Through further analysis, we reveal that our approach is capable of simplifying domain consistency complexity. While existing methods like [38, 59] involve $O(n^2)$ considerations for inter-domain consistency regularization and online knowledge distillation, our approach reduces this to a more efficient $O(n)$, where n represents the number of target domains considered.

As shown in supplementary material, we have also included Mapillary as a target domain and compared it previous work [43].

4.6 Ablation Study

Weight Merging and Buffer Merging. We conduct ablation studies on our proposed parameter merging and buffer merging methods using ResNet101 [24] and MiT-B5 [80] as the image encoders in the segmentation network [29], with results reported in Tab. 4(a) and Tab. 4(b), respectively. We have observed variations in the generalization capabilities of single-target domain adaptation (STDA) models across different domains, which primarily arise from the varying diversity and quality of the target datasets used. Nonetheless, we select the higher harmonic means from STDA models as our baseline for comparison.

Table 4: Ablation Study on Different Vision Backbones.

(a) ResNet101 Backbone

Setting	Merging Params.	Merging Buffers	Metric C	I	H	Δ
STDA (G → C)	-	-	**61.9**	44.3	51.6	-0.5%
STDA (G → I)	-	-	49.4	**55.1**	52.1	-
Ours	✓	✗	58.7	51.3	**54.8**	+2.7%
	✓	✓	58.8	54.0	**56.3**	+4.2%

(b) MiT-B5 Backbone

Setting	Merging Params.	Merging Buffers	Metric C	I	H	Δ
STDA (G → C)	-	-	**72.6**	50.6	59.6	-
STDA (G → I)	-	-	57.9	52.4	55.0	-4.6%
Ours	✓	✗	66.6	54.9	60.2	+0.6%
	✓	✓	67.0	**55.6**	**60.8**	+1.2%

The data in Tab. 4(a) and Tab. 4(b) reveal that employing a straightforward mid-point merging approach for parameters leads to an increase in generalization ability by +2.7% and +0.6%. Furthermore, when buffer merging is incorporated, this enhancement in performance is further amplified to +4.2% and +1.2%. We also observe an intriguing phenomenon with the MiT-B5 backbone: the merged model outperforms the single-target adapted model when evaluating in the IDD domain. This finding implies that domain-invariant knowledge can be acquired from other domains. These results suggest that each part of our proposed model merging technique is effective.

Different STDA Methods & Tasks. To validate the versatility of our proposed model merging method, we conducted experiments on another STDA Method ADVENT [75]. We also apply our methods on image classification tasks. Details are in the supplementary material.

5 Conclusion

This paper introduces a novel model merging strategy aimed at addressing the multi-target domain adaptation (MTDA) challenge without relying on training data. Our findings reveal that when pretrained on extensive datasets, both deep convolutional neural networks and transformer-based vision models can confine the finetuned models with the same basin in the loss landscape. We also emphasize the significance of buffer merging in MTDA, as buffers are key to capturing the unique features of various domains. The methods we propose are straightforward yet highly effective, establishing new state-of-the-art results on the MTDA benchmark. We anticipate that the concepts and methodologies presented in this paper will inspire future explorations in this field.

Acknowledgements

This research is supported by Tsinghua University – Mercedes Benz Institute for Sustainable Mobility.

References

1. Ainsworth, S.K., Hayase, J., Srinivasa, S.: Git re-basin: Merging models modulo permutation symmetries. arXiv preprint arXiv:2209.04836 (2022)
2. Araslanov, N., Roth, S.: Self-supervised augmentation consistency for adapting semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15384–15394 (2021)
3. Ashmore, S., Gashler, M.: A method for finding similarity between multi-layer perceptrons by forward bipartite alignment. In: 2015 International Joint Conference on Neural Networks (IJCNN). pp. 1–7. IEEE (2015)
4. Bjorck, N., Gomes, C.P., Selman, B., Weinberger, K.Q.: Understanding batch normalization. Advances in neural information processing systems **31** (2018)
5. Chen, L., Wei, Z., Jin, X., Chen, H., Zheng, M., Chen, K., Jin, Y.: Deliberated domain bridging for domain adaptive semantic segmentation. Advances in Neural Information Processing Systems **35**, 15105–15118 (2022)
6. Chen, M., Xue, H., Cai, D.: Domain adaptation for semantic segmentation with maximum squares loss. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2090–2099 (2019)
7. Chen, M., Chen, J., Ye, X., Gao, H.a., Chen, X., Fan, Z., Zhao, H.: Ultraman: Single image 3d human reconstruction with ultra speed and detail. arXiv preprint arXiv:2403.12028 (2024)
8. Choi, J., Kim, T., Kim, C.: Self-ensembling with gan-based data augmentation for domain adaptation in semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6830–6840 (2019)
9. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)
10. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
11. Entezari, R., Sedghi, H., Saukh, O., Neyshabur, B.: The role of permutation invariance in linear mode connectivity of neural networks. arXiv preprint arXiv:2110.06296 (2021)
12. Frankle, J., Carbin, M.: The lottery ticket hypothesis: Finding sparse, trainable neural networks. arXiv preprint arXiv:1803.03635 (2018)
13. Frankle, J., Dziugaite, G.K., Roy, D., Carbin, M.: Linear mode connectivity and the lottery ticket hypothesis. In: International Conference on Machine Learning. pp. 3259–3269. PMLR (2020)
14. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: International conference on machine learning. pp. 1180–1189. PMLR (2015)
15. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. The journal of machine learning research **17**(1), 2096–2030 (2016)

16. Gao, H.a., Gao, M., Li, J., Li, W., Zhi, R., Tang, H., Zhao, H.: Scp-diff: Photo-realistic semantic image synthesis with spatial-categorical joint prior. arXiv preprint arXiv:2403.09638 (2024)
17. Gao, H.a., Tian, B., Li, P., Chen, X., Zhao, H., Zhou, G., Chen, Y., Zha, H.: From semi-supervised to omni-supervised room layout estimation using point clouds. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 2803–2810. IEEE (2023)
18. Gao, H.a., Tian, B., Li, P., Zhao, H., Zhou, G.: Dqs3d: Densely-matched quantization-aware semi-supervised 3d detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21905–21915 (2023)
19. Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D.P., Wilson, A.G.: Loss surfaces, mode connectivity, and fast ensembling of dnns. Advances in neural information processing systems **31** (2018)
20. Gholami, B., Sahu, P., Rudovic, O., Bousmalis, K., Pavlovic, V.: Unsupervised multi-target domain adaptation: An information theoretic approach. IEEE Transactions on Image Processing **29**, 3993–4002 (2020)
21. Gong, R., Li, W., Chen, Y., Gool, L.V.: Dlow: Domain flow for adaptation and generalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2477–2486 (2019)
22. Guan, T., Shen, W., Yang, X., Wang, X., Yang, X.: Bridging synthetic and real worlds for pre-training scene text detectors. arXiv preprint arXiv:2312.05286 (2023)
23. He, B., Ding, N., Qian, C., Deng, J., Cui, G., Yuan, L., Gao, H.a., Chen, H., Liu, Z., Sun, M.: Zero-shot generalization during instruction tuning: Insights from similarity and granularity. arXiv preprint arXiv:2406.11721 (2024)
24. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
25. Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A., Darrell, T.: Cycada: Cycle-consistent adversarial domain adaptation. In: International conference on machine learning. pp. 1989–1998. Pmlr (2018)
26. Hoffman, J., Wang, D., Yu, F., Darrell, T.: Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. arXiv preprint arXiv:1612.02649 (2016)
27. Horvath, S., Laskaridis, S., Almeida, M., Leontiadis, I., Venieris, S., Lane, N.: Fjord: Fair and accurate federated learning under heterogeneous targets with ordered dropout. Advances in Neural Information Processing Systems **34**, 12876–12889 (2021)
28. Hoyer, L., Dai, D., Van Gool, L.: Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9924–9935 (2022)
29. Hoyer, L., Dai, D., Van Gool, L.: Hrda: Context-aware high-resolution domain-adaptive semantic segmentation. In: European Conference on Computer Vision. pp. 372–391. Springer (2022)
30. Hoyer, L., Dai, D., Wang, H., Van Gool, L.: Mic: Masked image consistency for context-enhanced domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11721–11732 (2023)
31. Hoyer, L., Dai, D., Wang, Q., Chen, Y., Van Gool, L.: Improving semi-supervised and domain-adaptive semantic segmentation with self-supervised depth estimation. International Journal of Computer Vision pp. 1–27 (2023)
32. Huang, G., Li, Y., Pleiss, G., Liu, Z., Hopcroft, J.E., Weinberger, K.Q.: Snapshot ensembles: Train 1, get m for free. arXiv preprint arXiv:1704.00109 (2017)

33. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning. pp. 448–456. pmlr (2015)
34. Isobe, T., Jia, X., Chen, S., He, J., Shi, Y., Liu, J., Lu, H., Wang, S.: Multi-target domain adaptation with collaborative consistency learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8187–8196 (2021)
35. Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., Wilson, A.G.: Averaging weights leads to wider optima and better generalization. arXiv preprint arXiv:1803.05407 (2018)
36. Jordan, K., Sedghi, H., Saukh, O., Entezari, R., Neyshabur, B.: Repair: Renormalizing permuted activations for interpolation repair. arXiv preprint arXiv:2211.08403 (2022)
37. Kim, M., Byun, H.: Learning texture invariant representation for domain adaptation of semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12975–12984 (2020)
38. Koh, K.B., Fernando, B.: Consistency regularization for domain adaptation. In: European Conference on Computer Vision. pp. 347–359. Springer (2022)
39. Kundu, J.N., Kulkarni, A., Singh, A., Jampani, V., Babu, R.V.: Generalize then adapt: Source-free domain adaptive semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7046–7056 (2021)
40. Kundu, J.N., Venkatesh, R.M., Venkat, N., Revanur, A., Babu, R.V.: Class-incremental domain adaptation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16. pp. 53–69 (2020)
41. Lai, X., Tian, Z., Jiang, L., Liu, S., Zhao, H., Wang, L., Jia, J.: Semi-supervised semantic segmentation with directional context-aware consistency. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1205–1214 (2021)
42. Lee, D.H., et al.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on challenges in representation learning, ICML. vol. 3, p. 896. Atlanta (2013)
43. Lee, S., Choi, W., Kim, C., Choi, M., Im, S.: Adas: A direct adaptation strategy for multi-target domain adaptive semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19196–19206 (2022)
44. Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V.: Federated optimization in heterogeneous networks. Proceedings of Machine learning and systems **2**, 429–450 (2020)
45. Li, W., Xu, H., Zhang, G., Gao, H.a., Gao, M., Wang, M., Zhao, H.: Fairdiff: Fair segmentation with point-image diffusion. arXiv preprint arXiv:2407.06250 (2024)
46. Li, Y., Yuan, L., Vasconcelos, N.: Bidirectional learning for domain adaptation of semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6936–6945 (2019)
47. Long, M., Cao, Z., Wang, J., Jordan, M.I.: Conditional adversarial domain adaptation. Advances in neural information processing systems **31** (2018)
48. Long, X., Zheng, Y., Zheng, Y., Tian, B., Lin, C., Liu, L., Zhao, H., Zhou, G., Wang, W.: Adaptive surface normal constraint for geometric estimation from monocular images. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)

49. Luo, Y., Zheng, L., Guan, T., Yu, J., Yang, Y.: Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2507–2516 (2019)
50. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial intelligence and statistics. pp. 1273–1282. PMLR (2017)
51. Melas-Kyriazi, L., Manrai, A.K.: Pixmatch: Unsupervised domain adaptation via pixelwise consistency training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12435–12445 (2021)
52. Mirzadeh, S.I., Farajtabar, M., Gorur, D., Pascanu, R., Ghasemzadeh, H.: Linear mode connectivity in multitask and continual learning. arXiv preprint arXiv:2010.04495 (2020)
53. Mishchenko, K., Malinovsky, G., Stich, S., Richtárik, P.: Proxskip: Yes! local gradient steps provably lead to communication acceleration! finally! In: International Conference on Machine Learning. pp. 15750–15769. PMLR (2022)
54. Neyshabur, B., Sedghi, H., Zhang, C.: What is being transferred in transfer learning? Advances in neural information processing systems **33**, 512–523 (2020)
55. Nguyen-Meidine, L.T., Belal, A., Kiran, M., Dolz, J., Blais-Morin, L.A., Granger, E.: Unsupervised multi-target domain adaptation through knowledge distillation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1339–1347 (2021)
56. Pan, F., Shin, I., Rameau, F., Lee, S., Kweon, I.S.: Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3764–3773 (2020)
57. Peña, F.A.G., Medeiros, H.R., Dubail, T., Aminbeidokhti, M., Granger, E., Pedersoli, M.: Re-basin via implicit sinkhorn differentiation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20237–20246 (2023)
58. Qin, Y., Qian, C., Yi, J., Chen, W., Lin, Y., Han, X., Liu, Z., Sun, M., Zhou, J.: Exploring mode connectivity for pre-trained language models. arXiv preprint arXiv:2210.14102 (2022)
59. Reddy, N., Baktashmotlagh, M., Arora, C.: Towards domain-aware knowledge distillation for continual model generalization. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 696–707 (2024)
60. Richter, S.R., Vineet, V., Roth, S., Koltun, V.: Playing for data: Ground truth from computer games. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14. pp. 102–118. Springer (2016)
61. Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.M.: The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3234–3243 (2016)
62. Sajjadi, M., Javanmardi, M., Tasdizen, T.: Regularization with stochastic transformations and perturbations for deep semi-supervised learning. Advances in neural information processing systems **29** (2016)
63. Sakaridis, C., Dai, D., Van Gool, L.: Map-guided curriculum domain adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(6), 3139–3153 (2020)

64. Sakaridis, C., Dai, D., Van Gool, L.: Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10765–10775 (2021)
65. Santurkar, S., Tsipras, D., Ilyas, A., Madry, A.: How does batch normalization help optimization? Advances in neural information processing systems **31** (2018)
66. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems **30** (2017)
67. Tatro, N., Chen, P.Y., Das, P., Melnyk, I., Sattigeri, P., Lai, R.: Optimizing mode connectivity via neuron alignment. Advances in Neural Information Processing Systems **33**, 15300–15311 (2020)
68. Tian, B., Gao, H.a., Cui, L., Zheng, Y., Luo, L., Wang, B., Zhi, R., Zhou, G., Zhao, H.: Latency-aware road anomaly segmentation in videos: A photorealistic dataset and new metrics. arXiv preprint arXiv:2401.04942 (2024)
69. Tian, B., Liu, M., Gao, H.a., Li, P., Zhao, H., Zhou, G.: Unsupervised road anomaly detection with language anchors. In: 2023 IEEE international conference on robotics and automation (ICRA). pp. 7778–7785. IEEE (2023)
70. Tranheden, W., Olsson, V., Pinto, J., Svensson, L.: Dacs: Domain adaptation via cross-domain mixed sampling. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1379–1389 (2021)
71. Tsai, Y.H., Hung, W.C., Schulter, S., Sohn, K., Yang, M.H., Chandraker, M.: Learning to adapt structured output space for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7472–7481 (2018)
72. Tsai, Y.H., Sohn, K., Schulter, S., Chandraker, M.: Domain adaptation for structured output via discriminative patch representations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1456–1465 (2019)
73. Varma, G., Subramanian, A., Namboodiri, A., Chandraker, M., Jawahar, C.: Idd: A dataset for exploring problems of autonomous navigation in unconstrained environments. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1743–1751. IEEE (2019)
74. Von Oswald, J., Kobayashi, S., Meulemans, A., Henning, C., Grewe, B.F., Sacramento, J.: Neural networks with late-phase weights. arXiv preprint arXiv:2007.12927 (2020)
75. Vu, T.H., Jain, H., Bucher, M., Cord, M., Pérez, P.: Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2517–2526 (2019)
76. Wang, H., Yurochkin, M., Sun, Y., Papailiopoulos, D., Khazaeni, Y.: Federated learning with matched averaging. arXiv preprint arXiv:2002.06440 (2020)
77. Wang, Q., Dai, D., Hoyer, L., Van Gool, L., Fink, O.: Domain adaptive semantic segmentation with self-supervised depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8515–8525 (2021)
78. Wei, Y., Wang, Z., Lu, Y., Xu, C., Liu, C., Zhao, H., Chen, S., Wang, Y.: Editable scene simulation for autonomous driving via collaborative llm-agents. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15077–15087 (2024)
79. Wortsman, M., Ilharco, G., Kim, J.W., Li, M., Kornblith, S., Roelofs, R., Lopes, R.G., Hajishirzi, H., Farhadi, A., Namkoong, H., et al.: Robust fine-tuning of zero-shot models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7959–7971 (2022)

80. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in Neural Information Processing Systems **34**, 12077–12090 (2021)
81. Yang, J., An, W., Wang, S., Zhu, X., Yan, C., Huang, J.: Label-driven reconstruction for domain adaptation in semantic segmentation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16. pp. 480–498. Springer (2020)
82. Yao, C.H., Gong, B., Qi, H., Cui, Y., Zhu, Y., Yang, M.H.: Federated multi-target domain adaptation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1424–1433 (2022)
83. Yu, H., Hu, M., Chen, S.: Multi-target unsupervised domain adaptation without exactly shared categories. arXiv preprint arXiv:1809.00852 (2018)
84. Yue, X., Zhang, Y., Zhao, S., Sangiovanni-Vincentelli, A., Keutzer, K., Gong, B.: Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2100–2110 (2019)
85. Yurochkin, M., Agarwal, M., Ghosh, S., Greenewald, K., Hoang, N., Khazaeni, Y.: Bayesian nonparametric federated learning of neural networks. In: International conference on machine learning. pp. 7252–7261. PMLR (2019)
86. Zhang, K., Sun, Y., Wang, R., Li, H., Hu, X.: Multiple fusion adaptation: A strong framework for unsupervised semantic segmentation adaptation. arXiv preprint arXiv:2112.00295 (2021)
87. Zhang, P., Zhang, B., Zhang, T., Chen, D., Wang, Y., Wen, F.: Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12414–12424 (2021)
88. Zhang, Y., Roy, S., Lu, H., Ricci, E., Lathuilière, S.: Cooperative self-training for multi-target adaptive semantic segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5604–5613 (2023)
89. Zheng, Y., Zhong, C., Li, P., Gao, H.a., Zheng, Y., Jin, B., Wang, L., Zhao, H., Zhou, G., Zhang, Q., et al.: Steps: Joint self-supervised nighttime image enhancement and depth estimation. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 4916–4923. IEEE (2023)
90. Zheng, Z., Yang, Y.: Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. International Journal of Computer Vision **129**(4), 1106–1120 (2021)
91. Zhou, Q., Feng, Z., Gu, Q., Cheng, G., Lu, X., Shi, J., Ma, L.: Uncertainty-aware consistency regularization for cross-domain semantic segmentation. Computer Vision and Image Understanding **221**, 103448 (2022)
92. Zou, Y., Yu, Z., Kumar, B., Wang, J.: Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In: Proceedings of the European conference on computer vision (ECCV). pp. 289–305 (2018)