Transferring to Real-World Layouts: A Depth-aware Framework for Scene Adaptation

Mu Chen
ReLER Lab, AAII, University of Technology Sydney
Sydney, NSW, Australia
mu.chen@student.uts.edu.au

Zhedong Zheng
FST and ICI, University of Macau
Macau, China
zhedongzheng@um.edu.mo

Yi Yang[†]
ReLER Lab, AAII, University of Technology Sydney
Sydney, NSW, Australia
yi.yang@uts.edu.au

Abstract

Scene segmentation via unsupervised domain adaptation (UDA) enables the transfer of knowledge acquired from source synthetic data to real-world target data, which largely reduces the need for manual pixel-level annotations in the target domain. To facilitate domain-invariant feature learning, existing methods typically mix data from both the source domain and target domain by simply copying and pasting pixels. Such vanilla methods are usually sub-optimal since they do not take into account how well the mixed layouts correspond to real-world scenarios. Real-world scenarios are with an inherent layout. Real-world scenarios are with an inherent layout. We observe that semantic categories, such as sidewalks, buildings, and sky, display relatively consistent depth distributions, and could be clearly distinguished in a depth map. The model suffers from confusion in predicting the target domain due to the unrealistic mixing. For instance, it is not reasonable to directly paste the near "pedestrian" pixels into the remote "sky" area. Based on such observation, we propose a depth-aware framework to explicitly leverage depth estimation to mix categories and facilitate two complementary tasks, *i.e.*, segmentation and depth learning in an end-to-end manner. In particular, the framework contains a Depth-guided Contextual Filter (DCF) for data augmentation and a cross-task encoder for contextual learning. DCF simulates the real-world layouts, while the cross-task encoder further adaptively fuses the complementing features between two tasks. Besides, several public datasets do not provide depth annotation. Therefore, we leverage the off-the-shelf depth estimation network to obtain the pseudo depth. Extensive experiments show that our methods, even with pseudo depth, achieve competitive performance, *i.e.*, 77.7 mIoU on GTA→Cityscapes and 69.3 mIoU on Synthia→Cityscapes.

CCS Concepts

• **Computing methodologies → Scene understanding**; **Transfer Learning**.

Keywords

Unsupervised Scene Adaptation, Depth Fusion, Transfer Learning

MM '24, October 28–November 1, 2024, Melbourne, VIC, Australia.
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0686-8/24/10
https://doi.org/10.1145/3664647.3681041

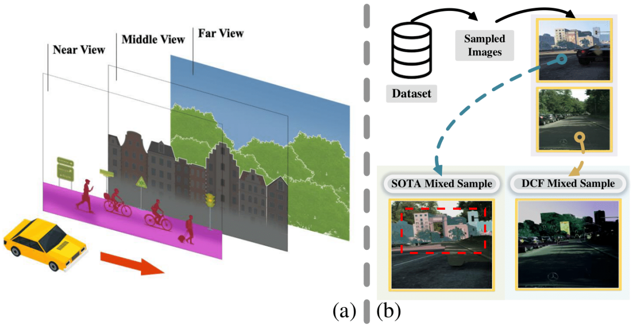

Figure 1: (a) Considering the driving scenario, we observe that the object location is relatively stable according to the distance from the camera. With such insight, we propose a Depth-guided Contextual Filter (DCF) which is aware of the semantic categories distribution in terms of Near, Middle, and Far view to facilitate cross-domain mixing. (b) Since we explicitly take the semantic layout into consideration, our method achieves more realistic mixed samples compared to existing state-of-the-art methods (Vanilla Mixed Sample) [10, 22]. As shown in the red box, "new" buildings are pasted before the parked cars.

ACM Reference Format:
Mu Chen, Zhedong Zheng, and Yi Yang[†]. 2024. Transferring to Real-World Layouts: A Depth-aware Framework for Scene Adaptation. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24), October 28–November 1, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3664647.3681041

1 Introduction

Semantic segmentation serves as a essential task in machine vision [15, 34, 35, 40, 41, 60, 73, 100], benefiting numerous vision applications [7, 12, 72, 83, 85, 91, 101]. It has achieved significant progress in the last few years [2, 6, 8, 12, 34–36, 45, 79]. It is worth noting that prevailing models usually require large-scale training datasets with high-quality annotations, such as ADE20K [98], to achieve good performance and but such pixel-level annotations in real-world are usually unaffordable and time-consuming [14]. One straightforward idea is to train networks with synthetic data so that the pixel-level annotations are easier to obtain [53, 54]. However, the network trained with synthetic data results in poor scalability when being deployed to a real-world environment due to multiple

[†] Corresponding author: yi.yang@uts.edu.au

factors, such as weather, illumination, and road design. Therefore, researchers resort to unsupervised domain adaptation (UDA) to further tackle the variance between domains [70, 71]. One branch of UDA methods attempts to mitigate the domain shift by aligning the domain distributions [21, 47, 52, 77]. Another potential paradigm to heal the domain shift is self-training [37, 84, 96, 102], which recursively refine the target pseudo-labels. Taking one step further, recent DACS [61] and follow-up works [10, 22–24, 27, 69] combine self-training and ClassMix [50] to mix images from both source and target domain. In this way, these works could craft highly perturbed samples to assist training by facilitating learning shared knowledge between two domains. Specifically, cross-domain mixing aims to copy the corresponding regions of certain categories from a source domain image and paste them onto an unlabelled target domain image. We note that such a vanilla strategy leads to pasting a large amount of objects to the unrealistic depth position. It is because that every category has its own position distribution. For instance, background classes such as "sky" and "vegetation" usually appear farther away, while the classes that occupy a small number of pixels such as "traffic signs" and "pole", usually appear closer as shown in Figure 1 (a). Such crafted training data compromise contextual learning, leading to sub-optimal location prediction performance, especially for small objects.

To address these limitations, we observe the real-world depth distribution and find that semantic categories are easily separated (disentangled) in the depth map since they follow a similar distribution under certain scenarios, *e.g.*, urban. Therefore, we propose a new depth-aware framework, which contains Depth Contextual Filter (DCF) and a cross-task encoder. In particular, DCF removes unrealistic classes mixed with the real-world target training samples based on the depth information. On the other hand, multi-modal data could improve the performance of deep representations and the effective use of the deep multi-task features to facilitate the final predictions is crucial. The proposed cross-task encoder contains two specific heads to generate intermediate features for each task and an Adaptive Feature Optimization module (AFO). AFO encourages the network to optimize the fused multi-task features in an end-to-end manner. Specifically, the proposed AFO adopts a series of transformer blocks to capture the information that is crucial to distinguish different categories and assigns high weights to discriminative features and vice versa.

The main contributions are as follows: **(1)** We propose a simple Depth-Guided Contextual Filter (DCF) to explicitly leverage the key semantic categories distribution hidden in the depth map, enhancing the realism of cross-domain information mixing and refining the cross-domain layout mixing. **(2)** We propose an Adaptive Feature Optimization module (AFO) that enables the cross-task encoder to exploit the discriminative depth information and embed it with the visual feature which jointly facilitates semantic segmentation and pseudo depth estimation. **(3)** Albeit simple, the effectiveness of our proposed methods has been verified by extensive ablation studies. Despite the pseudo depth, our method still achieves competitive accuracy on two commonly used scene adaptation benchmarks, namely 77.7 mIoU on GTA→Cityscapes and 69.3 mIoU on Synthia→Cityscapes.

2 Related Work

2.1 Unsupervised Domain Adaptation

Unsupervised domain adaptation (UDA) aims to train a model on a label-rich source domain and adapt the model to a label-scarce target domain. Some methods propose learning the domain-invariant knowledge by aligning the source and target distribution at different levels. For instance, AdaptSegNet [62], ADVENT [65], and CLAN [47] adversarially align the distributions in the feature space. CyCADA [21] diminishes the domain shift at both pixel-level and feature-level representation. DALN [4] proposes a discriminator-free adversarial learning network and leverages the predicted discriminative information for feature alignment. Both Wu *et al.*[77] and Yue *et al.* [86] learn domain-invariant features by transferring the input images into different styles, such as rainy and foggy, while Zhao *et al.* [94] and Zhang *et al.* [89] diversify the feature distribution via normalization and adding noise respectively. Another line of work refines pseudo-labels gradually under the iterative self-training framework, yielding competitive results. Following the motivation of generating highly reliable pseudo labels for further model optimization, CBST [102] adopts class-specific thresholds on top of self-training to improve the generated labels. Feng *et al.*[16] acquire pseudo labels with high precision by leveraging the group information. PyCDA [39] constructs pseudo-labels in various scales to further improve the training. Zheng *et al.*[95] introduce memory regularization to generate consistent pseudo labels. Other works propose either confidence regularization [96, 103] or category-aware rectification [87, 88] to improve the quality of pseudo labels. DACS [61] proposes a domain-mixed self-training pipeline to mix cross-domain images during training, avoiding training instabilities. Kim *et al.*[29], Li *et al.*[38] and Wang *et al.*[68] combine adversarial and self-training for further improvement. Chen *et al.*[5] establish a deliberated domain bridging (DDB) that aligns and interacts with the source and target domain in the intermediate space. SePiCo [78] and PiPa [10] adopt contrastive learning to align the domains. Liu *et al.*[44] addresses the label shift problem by adopting class-level feature alignment for conditional distribution alignment. Researchers also attempted to perform entropy minimization [9, 65], and image translation [19, 81]. consistency regularization[1, 13, 49, 99]. Recent multi-target domain adaptation methods enable a single model to adapt a labeled source domain to multiple unlabeled target domains [17, 32, 57]. However, the above methods usually ignore the rich multi-modality information, which can be easily obtained from the depth and other sensors.

2.2 Depth Estimation and Multi-task Learning in Semantic Segmentation

Semantic segmentation and geometric information are shown to be highly correlated [28, 42, 59, 64, 67, 74, 80, 90, 92, 93]. Recently depth estimation has been increasingly used to improve the learning of semantics within the context of multi-task learning, but the depth information should be exploited more precisely to help the domain adaptation. SPIGAN [31] pioneered the use of geometric information as an additional supervision by regularizing the generator with an auxiliary depth regression task. DADA [66] introduces an adversarial training framework based on the fusion of semantic and depth predictions to facilitate the adaptation. GIO-Ada [11] leverages the geometric information on both the input level and output level to reduce domain shift. CTRL [55] encodes task dependencies between the semantic and depth predictions to capture the cross-task relationships. CorDA [69] bridges the domain gap by utilizing self-supervised depth estimation on both domains. Wu *et al.* [76] propose to further support semantic segmentation by depth distribution density. Our work follows a similar spirit to leverage depth knowledge as auxiliary supervision. It is worth noting that our work is primarily different from existing works in the following two aspects: (1) from the data perspective, we explicitly delineate the depth distribution to refine data augmentation and construct realistic training samples to enhance contextual learning. (2) from the network perspective, our proposed multi-task learning network not only adopts auxiliary supervision for learning more robust deep representations but also facilitates the multi-task feature fusion by iterative deploying of transformer blocks to jointly learn the rich multi-task information for improving the final predictions.

Algorithm 1 Depth-guided Contextual Filter Algorithm with Cross-Image Mixing and Self Training

```
Input: Source domain: $(\mathbf{x}^S, \mathbf{y}^S, \mathbf{z}^S \sim \mathrm{X}^S, \mathrm{Y}^S, \mathrm{Z}^S)$, Target domain: $(\mathbf{x}^T, \mathbf{z}^T \sim \mathrm{X}^T, \mathrm{Z}^T)$. Semantic network $\mathcal{F}_\theta$.
Initialize network parameters $\theta$ randomly.
for iteration = 1 to $n$ do
    $\hat{\mathbf{y}}^T \leftarrow \mathcal{F}_\theta\left(\mathbf{x}^T\right)$, Generate pseudo label
    Pre-calculate the density value $\mathbf{p}$ for each class $i$ at each depth interval from the target depth map $\mathbf{z}^T$,
    $\hat{\mathbf{y}}^M \leftarrow \mathcal{M} \odot \mathbf{y}^S + (1 - \mathcal{M}) \odot \hat{\mathbf{y}}^T$, Randomly select 50% categories and copy the category ground truth label from the source image to target pseudo label
    $\mathbf{x}^M \leftarrow \mathcal{M} \odot \mathbf{x}^S + (1 - \mathcal{M}) \odot \mathbf{x}^T$, Copy the corresponding category region from the source image to the target image
    Re-calculate the density value $\hat{\mathbf{p}}$ after the mixing,
    Calculate the depth density distribution difference before and after mixing,
    Filter the category once the difference exceeds the threshold,
    Re-generate the depth-aware binary mask $\mathcal{M}^{DCF}$,
    $\hat{\mathbf{y}}^F \leftarrow \mathcal{M}^{\mathcal{DCF}} \odot \mathbf{y}^S + \left(1 - \mathcal{M}^{\mathcal{DCF}}\right) \odot \hat{\mathbf{y}}^T$, Generate the filtered training samples with new DCF mask
    $\mathbf{x}^F \leftarrow \mathcal{M}^{\mathcal{DCF}} \odot \mathbf{x}^S + \left(1 - \mathcal{M}^{\mathcal{DCF}}\right) \odot \mathbf{x}^T$,
    Compute predictions
    $\bar{\mathbf{y}}^S \leftarrow argmax\left(\mathcal{F}_\theta\left(\mathbf{x}^S\right)\right)$,
    $\bar{\mathbf{y}}^F \leftarrow argmax\left(\mathcal{F}_\theta\left(\mathbf{x}^F\right)\right)$,
    Compute loss for the batch:
    $\ell \leftarrow \mathcal{L}\left(\bar{\mathbf{y}}^S, \mathbf{y}^S, \bar{\mathbf{y}}^F, \hat{\mathbf{y}}^F\right)$.
    Compute $\nabla_\theta \ell$ by backpropagation.
    Perform stochastic gradient descent.
end for
return $\mathcal{F}_\theta$
```

3 Method

3.1 Problem Formulation

In a typical Unsupervised Domain Adaptation (UDA) scenario, we have a source domain, denoted S, which consists of abundant labeled synthetic data. On the other hand, the target domain, represented by T, contains unlabeled real-world data. For example, we have labeled training samples $\left(\mathbf{x}^S, \mathbf{y}^S, \mathbf{z}^S \sim \mathrm{X}^S, \mathrm{Y}^S, \mathrm{Z}^S\right)$ in the source domain, where $\mathbf{x}^S, \mathbf{y}^S$ are the training image and the corresponding ground truth for semantic segmentation. $\mathbf{z}^S$ is the label for the depth estimation task. Similarly, we have unlabeled target images sampled from target domain data $\left(\mathbf{x}^T, \mathbf{z}^T \sim \mathrm{X}^T, \mathrm{Z}^T\right)$, where $\mathbf{x}^T$ is the unlabeled sample in the target domain and $\mathbf{z}^T$ is the label for the depth estimation task. Since depth annotation is not supported by common public datasets, we adopt pseudo depth that can be easily generated by the off-the-shelf model [18].

3.2 Depth-guided Contextual Filter

In UDA, recent works Recent UDA works [10, 22–24, 50, 69] often employ pixel mixing to create cross-domain augmented samples. The basic idea is straightforward: take a portion of pixels from a source domain image and transplant them onto an equivalent area in a target domain image. However, this simple approach faces challenges due to the inherent differences in structure and layout between source and target domain data. To decrease noisy signals and simulate augmented training samples with real-world layouts, we propose Depth-guided Contextual Filter (DCF) to reduce the noisy pixels that are naively mixed across domains. The implementation of DCF is represented as pseudo-code in Algorithm 1, where the image $\mathbf{x}^S$ and the corresponding semantic labels $\mathbf{y}^S$ are sampled from source domain data. The image $\mathbf{x}^T$ and the depth label $\mathbf{z}^T$ are from target domain data. Pseudo label $\hat{\mathbf{y}}^T$ is then generated as $\hat{\mathbf{y}}^T = \mathcal{F}_\theta\left(\mathbf{x}^T\right)$, where $\mathcal{F}_\theta$ is a pre-trained semantic network. In practice, $\mathcal{F}_\theta$ usually has been trained on the source domain dataset via supervised learning. Based on the hypothesis that most semantic categories usually fall under a finite depth range, we introduce DCF, which divides the target depth map $\mathbf{z}^T$ into a few discrete depth intervals $(\Delta z_1, ..., \Delta z_n)$. For a given real-world target input image $\mathbf{x}^T$ combined with the pseudo label $\hat{\mathbf{y}}^T$ and target depth map $\mathbf{z}^T$, the density value at each depth interval $(\Delta z_1, ..., \Delta z_n)$ for each class $i \in (1, \ldots, C)$ can be counted and normalized as a probability. We denote the density value for class i at the depth interval Δz_1 as $p_i(\Delta z_1)$. All the density values make up the depth distribution in the target domain image. Then we randomly select half of the categories on the source images to paste on the target domain image. In practice, we apply a binary mask $\mathcal{M}$ to denote the corresponding pixels. Then naive cross-domain mixed image $\mathbf{x}^{Mix}$ and the mixed label $\hat{\mathbf{y}}^{Mix}$ can be formulated as:

$$\mathbf{x}^{Mix} = \mathcal{M} \odot \mathbf{x}^S + (1 - \mathcal{M}) \odot \mathbf{x}^T, \tag{1}$$

$$\hat{\mathbf{y}}^{Mix} = \mathcal{M} \odot \mathbf{y}^S + (1 - \mathcal{M}) \odot \hat{\mathbf{y}}^T, \tag{2}$$

where $\odot$ denotes the element-wise multiplication of between the mask and the image. The naively mixed images are visualized in Figure 2. It could be observed that due to the depth distribution difference between two domains, pixels of "Building" category are

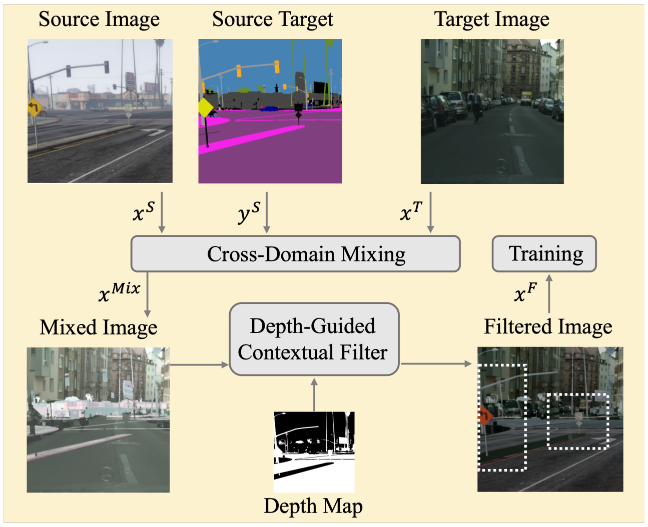

Figure 2: Source domain images x^S and x^T are mixed together, using the ground truth label y^S. The mixed images are denoised by our proposed Depth-guided Contextual Filter (DCF) and then trained by the network. We illustrate DCF with a set of practical sample. As illustrated, the unrealistic "Building" pixels from the source image are mixed pasted to the target image, leading to a noisy mixed sample. DCF removes these pixels and maintain mixed pixels of "Traffic Sign" and "Pole" shown in the white dotted boxes, enhancing the realism of cross-domain mixing. (Best viewed when zooming in.)

mixed from the source domain to the target domain, creating unrealistic images. Training with such training samples will compromise contextual learning. Therefore, we propose to filter the pixels that do not match the depth density distribution in the mixed image. After the naive mixing, we re-calculate the density value for each class at each depth interval. For example, the new density value for class i at the depth interval Δz_1 is denoted as $\hat{p}_i\,(\Delta z_1)$. Then we calculate the depth density distribution difference for each pasted category and denote the difference for class i at the depth interval Δz_1 as $\Delta p_i(\Delta z_1) = |p_i(\Delta z_1) - \hat{p}_i(\Delta z_1)|$. Once $\Delta p_i(\Delta z_1)$ exceeds the threshold of that category i, these pasted pixels are removed. After performing DCF, we confirm the final realistic pixels to be mixed and construct a depth-aware binary mask $\mathcal{M}^{\mathcal{DCF}}$, which is changed dynamically based on the depth layout of the current target image.

The filtered mixing samples are then generated. In practice, we directly apply the updated depth-aware mask to replace the original mask. Therefore, the new mixed sample and the label are as follows:

$$\mathbf{x}^F = \mathcal{M}^{\mathcal{DCF}} \odot \mathbf{x}^S + \left(1 - \mathcal{M}^{\mathcal{DCF}}\right) \odot \mathbf{x}^T, \tag{3}$$

$$\hat{\mathbf{y}}^F = \mathcal{M}^{\mathcal{DCF}} \odot \mathbf{y}^S + \left(1 - \mathcal{M}^{\mathcal{DCF}}\right) \odot \hat{\mathbf{y}}^T. \tag{4}$$

Because large objects such as "sky" and "terrain" usually aggregate and occupy a large amount of pixels and small objects only occupy a small amount of pixels in a certain depth range, we set different filtering thresholds for each category. DCF uses pseudo semantic labels for the target domain as there is no ground truth available. Since the label prediction is not stable in the early stage, we apply a warmup strategy to perform DCF after 10,000 iterations. Examples of the input images, naively mixed samples and filtered samples are presented in Figure 2. The sample after the process of the DCF module has the pixels from the source domain that match the depth distribution of the target domain, helping the network to better deal with the domain gap.

3.3 Multi-task Scene Adaptation Framework

To exploit the relation between segmentation and depth learning, we introduce a multi-task scene adaptation framework including a high resolution semantic encoder, and a cross-task shared encoder with a feature optimization module, which is depicted in Figure 3. The proposed framework incorporates and optimizes the fusion of depth information for improving the final semantic predictions.

High Resolution Semantic Prediction. Most supervised methods use high resolution images for training, but common scene adaptation methods usually use random crops of the image that is half of the full resolution. To reduce the domain gap between scene adaptation and supervised learning while maintaining the GPU memory consumption, we adopt a high-resolution encoder to encode HR image crops into deep HR features. Then a semantic decoder is used to generate the HR semantic predictions $\bar{\mathbf{y}}_{hr}$. We adopt the cross entropy loss for semantic segmentation:

$$\mathcal{L}^S_{hr} = \mathbb{E}\left[-\mathbf{y}^S \log \bar{\mathbf{y}}^S_{hr}\right], \quad \mathcal{L}^F_{hr} = \mathbb{E}\left[-\hat{\mathbf{y}}^F \log \bar{\mathbf{y}}^F_{hr}\right], \tag{5}$$

where $\bar{\mathbf{y}}^S_{hr}$ and $\bar{\mathbf{y}}^T_{hr}$ are high resolution semantic predictions. $\mathbf{y}^S$ is the one-hot semantic label for the source domain and $\hat{\mathbf{y}}^F$ is the one-hot pseudo label for the depth-aware fused domain.

Adaptive Feature Optimization. In addition to the high resolution encoder, We use another cross-task encoder to encode input images which are shared for both tasks. Depth maps are rich in spatial depth information, but a naive concatenation of depth information directly to visual information causes some interference, e.g. categories at similar depth positions are already well distinguished by visual information, and attention mechanisms can help the network to select the crucial part of the multitask information. In the proposed multi-task learning framework, the visual semantic feature and depth feature is generated by a visual head and a depth head, respectively. As shown in Figure 3, after applying batch normalization, an Adaptive Feature Optimization module then concatenates the normalized input visual feature and the input depth feature to create a fused multi-task feature by concatenation as $f^{in}_{fuse} = \text{CONCAT}\left(f^{in}_{vis}, f^{in}_{depth}\right)$. The fused feature is then fed into a series of transformer blocks to capture the key information between the two tasks. The attention mechanism adaptively adjusts the extent to which depth features are embedded in visual features:

$$f^{out}_{fuse} = \mathcal{W}_{Trans}\left(f^{in}_{fuse}\right), \tag{6}$$

where $\mathcal{W}_{Trans}$ is the transformer parameter. The learned output of the transformer blocks is a weight map γ which is multiplied back to the input visual feature and depth feature resulting in an

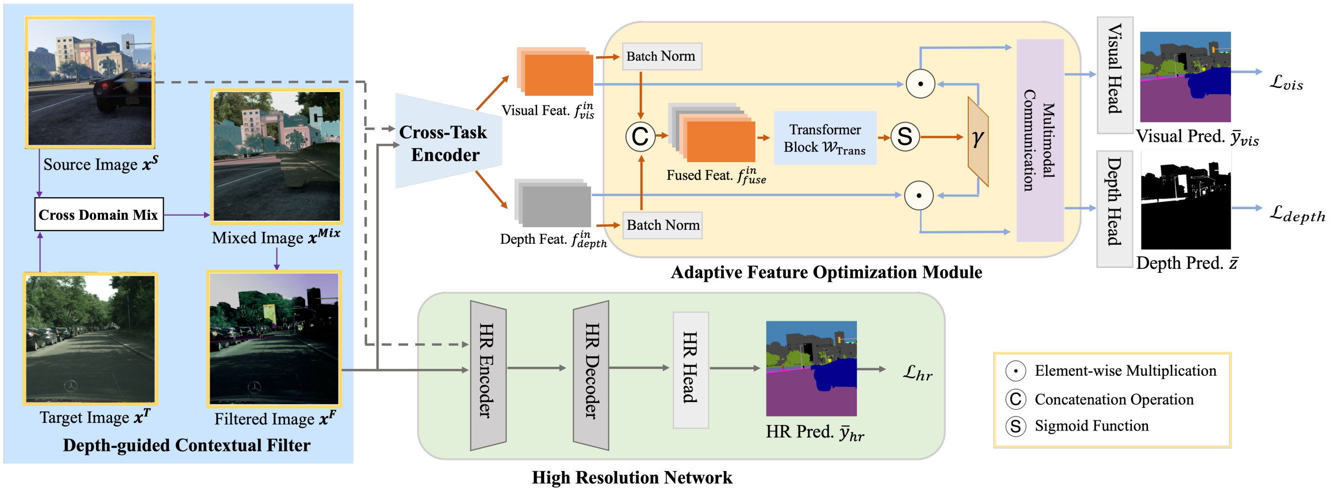

Figure 3: The proposed multi-task learning framework. The input images x^F are mixed from the source image x^S and target domain x^T according to the depth (Please refer to Figure 2). Then we are fed x^S and x^F into the high resolution encoder to generate high resolution predictions. To enhance multi-modal learning, the visual and depth feature created by the cross-task encoder are fused and fed into the proposed Adaptive Feature Optimization module (AFO) for multimodal communication. Finally, the multimodal communication via several transformer blocks incorporates and optimizes the fusion of depth information, improving the final visual predictions.

optimized feature as:

$$\boldsymbol{\gamma} = \boldsymbol{\sigma}\left(\mathcal{W}_{Conv} \otimes f^{out}_{fuse}\right), \tag{7}$$

where $\mathcal{W}_{Conv}$ denotes the convolution parameter, $\otimes$ denotes the convolution operation and $\boldsymbol{\sigma}$ represents the sigmoid function. The weight matrix $\boldsymbol{\gamma}$ performs adaptive optimization of the muti-task features. Then, the fused feature f^{out}_{fuse} is fed into different decoders for predicting different final tasks, *i.e.*, the visual and the depth task. The output features are essentially multimodal features containing crucial depth information:

$$f^{out}_{vis} = f^{in}_{vis} \odot \boldsymbol{\gamma}, \quad f^{out}_{depth} = f^{in}_{depth} \odot \boldsymbol{\gamma}, \tag{8}$$

where $\odot$ represents element-wise multiplication. The optimized visual and depth feature is then fed into the multimodal communication module for further processing. The multimodal communication module refines the learning of key information between two tasks by iterative use of transformer blocks. the inference is merely based on the visual input when the feature optimization is fished. The final semantic prediction $\bar{\mathbf{y}}^S_{vis}$ and depth prediction $\bar{\mathbf{z}}^S$ can be generated from the final visual feature f^{final}_{vis} and depth feature f^{final}_{depth} by visual head and depth head. Similar to high resolution predictions, we use the cross entropy loss for the semantic loss calculation:

$$\mathcal{L}^S_{vis} = \mathbb{E}\left[-\mathbf{y}^S \log \bar{\mathbf{y}}^S_{vis}\right], \quad \mathcal{L}^F_{vis} = \mathbb{E}\left[-\hat{\mathbf{y}}^F \log \bar{\mathbf{y}}^F_{vis}\right]. \tag{9}$$

We also employ berHu loss for depth regression at source domain:

$$\mathcal{L}^S_{depth} = \mathbb{E}\left[\text{berHu}\left(\bar{\mathbf{z}}^S - \mathbf{z}^S\right)\right], \tag{10}$$

where $\bar{z}$ and z are predicted and ground truth semantic maps. Following [55, 66], we deploy the reversed Huber criterion [30], which is defined as :

$$\text{ber}\,Hu\left(e_z\right) = \begin{cases} \left|e_z\right|, & \left|e_z\right| \leq H \\ \frac{(e_z)^2 + H^2}{2H}, & \left|e_z\right| > H \end{cases} \tag{11}$$
$$H = 0.2 \max\left(\left|e_z\right|\right),$$

where H is a positive threshold and we set it to 0.2 of the maximum depth residual. Finally, the overall loss function is:

$$\mathcal{L} = \mathcal{L}^S_{hr} + \mathcal{L}^S_{vis} + \lambda_{depth}\mathcal{L}^S_{depth} + \mathcal{L}^F_{hr} + \mathcal{L}^F_{vis}, \tag{12}$$

where hyperparameter λ_{depth} is the loss weight. Considering that our main task is semantic segmentation and the depth estimation is the auxiliary task, we empirically $\lambda_{depth} = \lambda_{depth} = 1 \times 10^{-3}$. We also designed the ablation studies to change the weight of depth task λ_{depth} to the level of 10^{-1} or 10^{-3}.

4 Experiment

4.1 Implementation Details

Datasets. We evaluate the proposed framework on two scene adaptation settings, *i.e.*, GTA → Cityscapes and SYNTHIA → Cityscapes, following common protocols [1, 22–24, 61, 69]. Particularly, the GTA5 dataset [53] is the synthetic dataset collected from a video game, which contains 24,966 images annotated by 19 classes. Following [69], we adopt depth information generated by Monodepth2 [18] model which is trained merely on GTA image sequences. SYNTHIA [54] is a synthetic urban scene dataset with 9,400 training images and 16 classes. Simulated depth information provided by SYNTHIA is adopted. GTA and SYNTHIA serve as source domain datasets. The target domain dataset is Cityscapes, which is collected from real-world street-view images. Cityscapes contains 2,975 unlabeled training images and 500 validation images. The resolution of

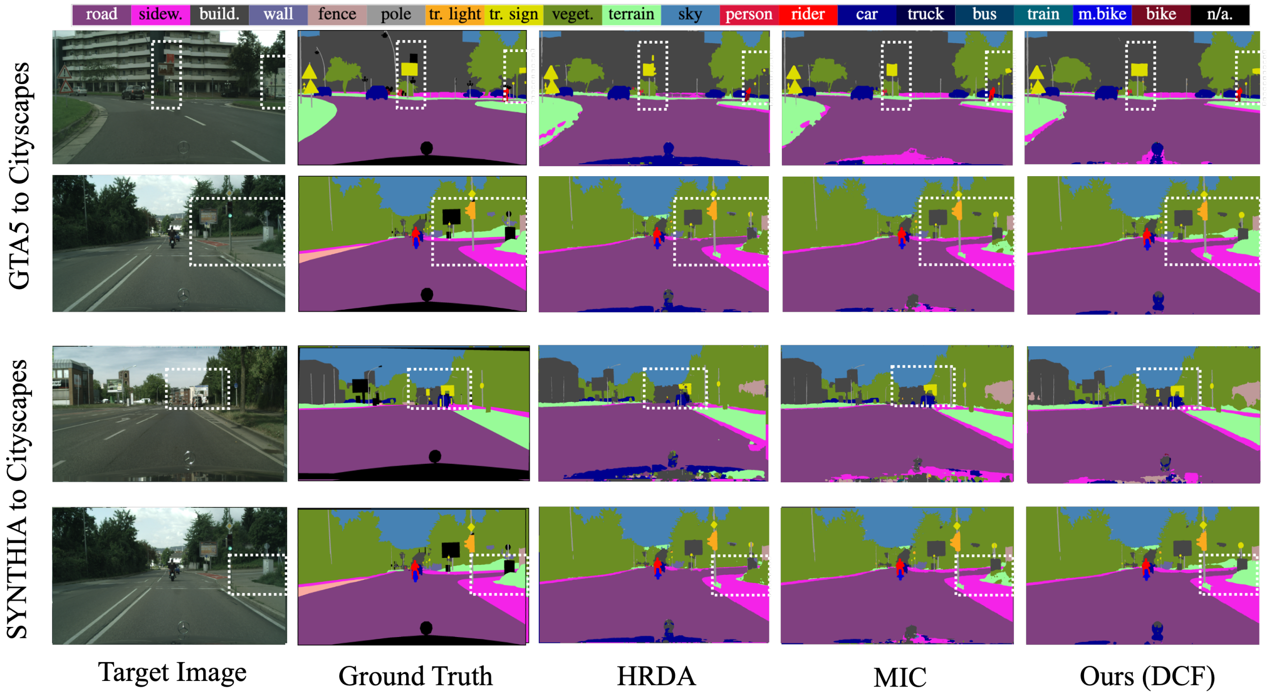

Figure 4: Qualitative results. From left to right: Target Image, Ground Truth, the visual results predicted by HRDA, MIC and Ours. We highlight prediction differences in white dash boxes and it is observed that the proposed method predicts clear edges.

Cityscapes is 2048 × 1024 and the common protocol downscales the size to 1024 × 512 to save memory. Following [69], the stereo depth estimation from [56] is used. We leverage the Intersection Over Union (IoU) for per-class performance and the mean Intersection over Union (mIoU) over all classes to report the result. The code is based on Pytorch [51].

Experimental Setup. We adopt DAFormer [22] network with MiT-B5 backbone [79] for the high resolution encoder and DeepLabV2 network with ResNet-101 backbone for the cross-task encoder to reduce the memory consumption. All backbones are initialized with ImageNet pretraining. Our training procedure is based on self-training methods with cross-domain mixing [22–24, 61] and enhanced by our proposed Depth-guided Contextual Filter. Following [23, 61], the input image resolution is half of the full resolution for the cross-task encoder and full resolution for high resolution encoder. We utilize the same data augmentation, *e.g.*, color jitter and Gaussian blur and empirically set pseudo labels threshold 0.968 following [61]. We train the network with batch size 2 for 40k iterations on a Tesla V100 GPU.

Data Resolution. Our proposed depth-aware multi-task framework contains a high resolution encoder and a cross-task encoder with an adaptive feature optimization module (AFO). Previous works [38, 61, 63] downsample Cityscapes to 1024 × and GTA to 1280 × 720. Following [23], for the high resolution encoder, we resize GTA to 2560 × 1440 and SYNTHIA to 2560 x 1520. Then the crop size is 1024 × 1024. In addition, SegFormer [79] MLP decoder with an embedding dimension of 256 is used for the high resolution branch. For the cross-task encoder branch, we follow common UDA methods [22, 61] to adopt 1024 × 512 pixels (half of the full resolution) for Cityscapes, 1280 × 760 for SYNTHIA and 1280 × 720 for GTA. In addition, a 512 × 512 random crop is extracted.

4.2 Comparison with SOTA

Results on GTA→Cityscapes. We show our results on GTA → Cityscapes in Table 1 and highlight the best results in bold. Our method yields significant performance improvement over the state-of-the-art method MIC [24] from 75.9 mIoU to 77.7 mIoU. Usually, classes that occupy a small number of pixels are difficult to adapt and have a comparably low IoU performance. However, our method demonstrates competitive IoU improvement in most categories especially on small objects such as +5.7 on "Rider", +5.4 on "Fence", +5.2 on "Wall", +4.4 on "Traffic Sign" and +3.4 on "Pole". The result shows the effectiveness of the proposed contextual filter and cross-task learning framework in contextual learning. Our method also increases the mIoU performance of classes that aggregate and occupy a large amount of pixels in an image by a smaller margin such as +1.8 on "Pedestrain" and +1.1 on "Bike", probably because the rich texture and color information contained in the visual feature already has the ability to recognize these relatively easier classes. The above observations are also qualitatively reflected in Figure 4, where we visualize the segmentation results of the proposed method and the comparison with previous strong transformer-based methods

Table 1: Quantitative comparison with previous UDA methods on GTA → Cityscapes. We present pre-class IoU and mIoU. The best accuracy in every column is in bold. Our results are averaged over 3 random seeds.

Method	Road	SW	Build	Wall	Fence	Pole	TL	TS	Veg.	Terrain	Sky	PR	Rider	Car	Truck	Bus	Train	Motor	Bike	mIoU
AdaptSegNet [62]	86.5	36.0	79.9	23.4	23.3	23.9	35.2	14.8	83.4	33.3	75.6	58.5	27.6	73.7	32.5	35.4	3.9	30.1	28.1	42.4
CyCADA [21]	86.7	35.6	80.1	19.8	17.5	38.0	39.9	41.5	82.7	27.9	73.6	64.9	19.0	65.0	12.0	28.6	4.5	31.1	42.0	42.7
CLAN [47]	87.0	27.1	79.6	27.3	23.3	28.3	35.5	24.2	83.6	27.4	74.2	58.6	28.0	76.2	33.1	36.7	6.7	31.9	31.4	43.2
SP-Adv [58]	86.2	38.4	80.8	25.5	20.5	32.8	33.4	28.2	85.5	36.1	80.2	60.3	28.6	78.7	27.3	36.1	4.6	31.6	28.4	44.3
MaxSquare [9]	88.1	27.7	80.8	28.7	19.8	24.9	34.0	17.8	83.6	34.7	76.0	58.6	28.6	84.1	37.8	43.1	7.2	32.3	34.2	44.3
AdvEnt [65]	89.4	33.1	81.0	26.6	26.8	27.2	33.5	24.7	83.9	36.7	78.8	58.7	30.5	84.8	38.5	44.5	1.7	31.6	32.4	45.5
MRNet [95]	89.1	23.9	82.2	19.5	20.1	33.5	42.2	39.1	85.3	33.7	76.4	60.2	33.7	86.0	36.1	43.3	5.9	22.8	30.8	45.5
APODA [82]	85.6	32.8	79.0	29.5	25.5	26.8	34.6	19.9	83.7	40.6	77.9	59.2	28.3	84.6	34.6	49.2	8.0	32.6	39.6	45.9
CBST [102]	91.8	53.5	80.5	32.7	21.0	34.0	28.9	20.4	83.9	34.2	80.9	53.1	24.0	82.7	30.3	35.9	16.0	25.9	42.8	45.9
MRKLD [103]	91.0	55.4	80.0	33.7	21.4	37.3	32.9	24.5	85.0	34.1	80.8	57.7	24.6	84.1	27.8	30.1	26.9	26.0	42.3	47.1
FADA [68]	91.0	50.6	86.0	43.4	29.8	36.8	43.4	25.0	86.8	38.3	87.4	64.0	38.0	85.2	31.6	46.1	6.5	25.4	37.1	50.1
Uncertainty [96]	90.4	31.2	85.1	36.9	25.6	37.5	48.8	48.5	85.3	34.8	81.1	64.4	36.8	86.3	34.9	52.2	1.7	29.0	44.6	50.3
FDA [84]	92.5	53.3	82.4	26.5	27.6	36.4	40.6	38.9	82.3	39.8	78.0	62.6	34.4	84.9	34.1	53.1	16.9	27.7	46.4	50.5
Adaboost [97]	90.7	35.9	85.7	40.1	27.8	39.0	49.0	48.4	85.9	35.1	85.1	63.1	34.4	86.8	38.3	49.5	0.2	26.5	45.3	50.9
DACS [61]	89.9	39.7	87.9	30.7	39.5	38.5	46.4	52.8	88.0	44.0	88.8	67.2	35.8	84.5	45.7	50.2	0.0	27.3	34.0	52.1
BAPA [43]	94.4	61.0	88.0	26.8	39.9	38.3	46.1	55.3	87.8	46.1	89.4	68.8	40.0	90.2	60.4	59.0	0.0	45.1	54.2	57.4
ProDA [87]	87.8	56.0	79.7	46.3	44.8	45.6	53.5	53.5	88.6	45.2	82.1	70.7	39.2	88.8	45.5	59.4	1.0	48.9	56.4	57.5
CaCo [26]	93.8	64.1	85.7	43.7	42.2	46.1	50.1	54.0	88.7	47.0	86.5	68.1	2.9	88.0	43.4	60.1	31.5	46.1	60.9	58.0
DAFormer [22]	95.7	70.2	89.4	53.5	48.1	49.6	55.8	59.4	89.9	47.9	92.5	72.2	44.7	92.3	74.5	78.2	65.1	55.9	61.8	68.3
CAMix [99]	96.0	73.1	89.5	53.9	50.8	51.7	58.7	64.9	90.0	51.2	92.2	71.8	44.0	92.8	78.7	82.3	70.9	54.1	64.3	70.0
HRDA [23]	96.4	74.4	91.0	61.6	51.5	57.1	63.9	69.3	91.3	48.4	94.2	79.0	52.9	93.9	84.1	85.7	75.9	63.9	67.5	73.8
MIC [24]	97.4	80.1	91.7	61.2	56.9	59.7	66.0	71.3	91.7	51.4	**94.3**	79.8	56.1	94.6	85.4	**90.3**	80.4	64.5	68.5	75.9
CorDA† [69]	94.7	63.1	87.6	30.7	40.6	40.2	47.8	51.6	87.6	47.0	89.7	66.7	35.9	90.2	48.9	57.5	0.0	39.8	56.0	56.6
FAFS† [3]	93.4	60.7	88.0	43.5	32.1	40.3	54.3	53.0	88.2	44.5	90.0	69.5	35.8	88.7	34.1	53.9	41.3	51.7	54.7	58.8
DBST† [3]	94.3	60.0	87.9	50.5	43.0	42.6	50.8	51.3	88.0	45.9	89.7	68.9	41.8	88.0	45.8	63.8	0.0	50.0	55.8	58.8
Ours†	**97.5**	**80.7**	**92.1**	**66.4**	**62.3**	**63.1**	**67.7**	**75.7**	**91.8**	**52.4**	93.9	**81.6**	**61.8**	**94.7**	**88.3**	90.0	**81.2**	**65.8**	**69.6**	**77.7**

†: Training with depth data.

HRDA [23], and MIC [24]. The qualitative results highlighted by white dash boxes show that the proposed method largely improved the prediction quality of challenging small object "Traffic Sign" and large category "Terrain".

Results on Synthia→Cityscapes. We show our results on SYNTHIA → Cityscapes in Table 1 and the results show the consistent performance improvement of our method, increasing from 67.3 to 69.3 (+2.0 mIoU) compared to the state-of-the-art method MIC [24]. Especially, our method significantly increases the IoU performance of the challenging class "SideWalk" from 50.5 to 63.1 (+12.6 mIoU). It is also noticeable that our method remains competitive in segmenting most individual classes and yields a significant increase of +6.8 on "Road", +6.6 on "Bus", +3.9 on "Pole", +3.7 on "Road", +3.2 on "Wall" and +2.9 on "Truck".

4.3 Ablation Study and Further Disccussion

Ablation Study on Different Scene Adaptation Frameworks. We combine our method with different scene adaptation architectures on GTA→Cityscapes. Table 4 shows that our method achieves consistent and significant improvements across different methods with different network architectures. Firstly, our method improves the state-of-the-art performance by +1.8 mIoU. Then we evaluate the proposed method on two strong methods based on transformer backbone, yielding +3.2 mIoU and +2.3 mIoU performance increase on DAFormer [22] and HRDA [23], respectively. Secondly, we evaluate our method on DeepLabV2 [6] architecture with ResNet-101 [20] backbone. We show that we improve the performance of the CNN-based cross-domain mixing method, *i.e.*, DACS by +4.1 mIoU. The ablation study verifies the effectiveness of our method in leveraging depth information to enhance cross-domain mixing not only on Transformer-based networks but also on CNN-based architecture.

Ablation Study on Different Components of the Proposed Method. In order to verify the effectiveness of our proposed components, we train four different models from M1 to M4 and show the result in Table 3. "ST Base" means the self training baseline with semantic segmentation branch and depth regression branch. "Naive Mix" denotes the cross-domain mixing strategy. "DCF" represents the proposed depth-aware mixing (Depth-guided Contextual Filter). "AFO" denotes the proposed Adaptive Feature Optimization module and we used two different method to perform AFO. Firstly, we leverage channel attention (CA) that could select useful information along the channel dimension to perform the feature optimization. In this method, the fused feature is adaptively optimized by SENet [25], the output is a weighted vector which is multiplied back to the visual and depth feature. We leavrage "AFO (CA)" to denote this method. Secondly, we leverage the iterative use of transformer block to adaptively optimize the multi-task feature. In this case, the output of the transformer block is a weighted map. The Multimodal Communication (MMC) module is then used to incorporate rich knowledge from the depth prediction. We denote this method as "AFO (Trans + MMC)". M1 is the self-training baseline with depth regression based on DAFormer architecture. M2 adds the cross-domain mixing strategy for improvement and shows a competitive result of 76.0 mIoU. M3 is the model with the Depth-guided Contextual Filter, increasing the performance from 76.0 to 77.1 mIoU (+1.1 mIoU), which demonstrates the effectiveness of transferring the mixed training images to real-world layout with the help of the depth information. M4 adds the multi-task framework that leverages Channel Attention (CA) mechanism to fuse the discriminative

Table 2: Quantitative comparison with previous UDA methods on SYNTHIA → Cityscapes. We present pre-class IoU, mIoU and mIoU*. mIoU and mIoU* are averaged over 16 and 13 categories, respectively. The best accuracy in every column is in bold.Our results are averaged over 3 random seeds.

Method	Road	SW	Build	Wall*	Fence*	Pole*	TL	TS	Veg.	Sky	PR	Rider	Car	Bus	Motor	Bike	mIoU*	mIoU
SIBAN [46]	82.5	24.0	79.4	–	–	–	16.5	12.7	79.2	82.8	58.3	18.0	79.3	25.3	17.6	25.9	46.3	–
PatchAlign [63]	82.4	38.0	78.6	8.7	0.6	26.0	3.9	11.1	75.5	84.6	53.5	21.6	71.4	32.6	19.3	31.7	46.5	40.0
AdaptSegNet [62]	84.3	42.7	77.5	–	–	–	4.7	7.0	77.9	82.5	54.3	21.0	72.3	32.2	18.9	32.3	46.7	–
CLAN [47]	81.3	37.0	80.1	–	–	–	16.1	13.7	78.2	81.5	53.4	21.2	73.0	32.9	22.6	30.7	47.8	–
CBST [102]	68.0	29.9	76.3	10.8	1.4	33.9	22.8	29.5	77.6	78.3	60.6	28.3	81.6	23.5	18.8	39.8	48.9	42.6
MRNet [95]	82.0	36.5	80.4	4.2	0.4	33.7	18.0	13.4	81.1	80.8	61.3	21.7	84.4	32.4	14.8	45.7	50.2	43.2
MRKLD [103]	67.7	32.2	73.9	10.7	1.6	37.4	22.2	31.2	80.8	80.5	60.8	29.1	82.8	25.0	19.4	45.3	50.1	43.8
CCM [33]	79.6	36.4	80.6	13.3	0.3	25.5	22.4	14.9	81.8	77.4	56.8	25.9	80.7	45.3	29.9	52.0	52.9	45.2
Uncertainty [96]	87.6	41.9	83.1	14.7	1.7	36.2	31.3	19.9	81.6	80.6	63.0	21.8	86.2	40.7	23.6	53.1	54.9	47.9
BL [38]	86.0	46.7	80.3	–	–	–	14.1	11.6	79.2	81.3	54.1	27.9	73.7	42.2	25.7	45.3	51.4	–
DT [75]	83.0	44.0	80.3	–	–	–	17.1	15.8	80.5	81.8	59.9	33.1	70.2	37.3	28.5	45.8	52.1	–
IAST [48]	81.9	41.5	83.3	17.7	4.6	32.3	30.9	28.8	83.4	85.0	65.5	30.8	86.5	38.2	33.1	52.7	49.8	-
DAFormer [22]	84.5	40.7	88.4	41.5	6.5	50.0	55.0	54.6	86.0	89.8	73.2	48.2	87.2	53.2	53.9	61.7	67.4	60.9
CAMix [99]	87.4	47.5	88.8	–	–	–	55.2	55.4	87.0	91.7	72.0	49.3	86.9	57.0	57.5	63.6	69.2	–
HRDA [23]	85.2	47.7	88.8	49.5	4.8	57.2	65.7	60.9	85.3	92.9	79.4	52.8	89.0	64.7	63.9	64.9	72.4	65.8
MIC [24]	86.6	50.5	89.3	47.9	7.8	59.4	66.7	63.4	87.1	**94.6**	**81.0**	**58.9**	90.1	61.9	**67.1**	64.3	74.0	67.3
DADA[†] [66]	89.2	44.8	81.4	6.8	0.3	26.2	8.6	11.1	81.8	84.0	54.7	19.3	79.7	40.7	14.0	38.8	49.8	42.6
CorDA[†] [69]	93.3	61.6	85.3	19.6	5.1	37.8	36.6	42.8	84.9	90.4	69.7	41.8	85.6	38.4	32.6	53.9	62.8	55.0
Ours[†]	**93.4**	**63.1**	**89.8**	**51.1**	**9.1**	**61.4**	**66.9**	**64.0**	**88.0**	94.5	80.9	56.6	**90.9**	**68.5**	63.7	**66.6**	**75.9**	**69.3**

[†]: Training with depth data.

Table 3: Ablation study of different components of our proposed framework on GTA→Cityscapes. The results are averaged over 3 random seeds.

Method	ST Base.	Naive Mix.	DCF.	AFO. (CA)	AFO. (Trans + MMC)	mIoU↑
M1	✓					73.1
M2	✓	✓				76.0
M3	✓	✓	✓			77.1
M4	✓	✓	✓	✓		77.3
M5	✓	✓	✓		✓	77.7

Table 4: Compatibility of the proposed method on different UDA methods and backbones on GTA→Cityscapes. Our results are averaged over 3 random seeds.

Backbone	UDA Method	w/o	w/	Diff.
DeepLabV2 [6]	DACS [61]	52.1	56.2	+4.1
DAFormer [22]	DAFormer [22]	68.3	71.5	+3.2
DAFormer [22]	HRDA [23]	73.8	76.1	+2.3
DAFormer [22]	MIC [24]	75.9	77.7	+1.8

depth feature into the visual feature. The segmentation result is increased by a small margin (+0.2 mIoU), which means CA could help the network to adaptively learn to focus or to ignore information from the auxiliary task to some extent. M5 is our proposed depth-aware multi-task model with both Depth-guided Contextual Filter and Adaptive Feature Optimization (AFO) module. Compared to M3, M5 has a mIoU increase of +0.6 from 77.1 to 77.7, which shows the effectiveness of multi-modal feature optimization using transformers to facilitate contextual learning.

Table 5: Quantitative results on GTA+SYNTHIA → Cityscapes. Here we use the vanilla backbone for a fair comparison.

Method	mIoU (%)	Δ mIoU (%)
Baseline (Single Source)	52.1	-
Multi Source	54.2	+2.1
Adaboost [97]	50.8	-
Multi Source + Depth	56.7	+4.6

Ablation study on GTA+SYNTHIA → Cityscapes. We evaluate the proposed method on multi-source domains setting and report the quantitative result on GTA+SYNTHIA → Cityscapes. With multi-source domain data, the model can be trained more robust to the unlabelled target environment. We adopt DACS [61] as our baseline with 52.1 mIoU (Only GTA) performance shown in Table 5. With more source-domain data, the model yields a better result of 54.2 mIoU. Then, we can observe that our method yields a larger improvement from 54.2 to 56.7 mIoU, demonstrating that the proposed model could adapt multi-domain depth to the target domain and hence increase performance.

5 Conclusion

In this work, we introduce a new depth-aware scene adaptation framework that effectively leverages the guidance of depth to enhance data augmentation and contextual learning. The proposed framework not only explicitly refines the cross-domain mixing by stimulating real-world layouts with the guidance of depth distributions of objects, but also introduced a cross-task encoder that adaptively optimizes the multi-task feature and focused on the discriminative depth feature to help contextual learning. By integrating our depth-aware framework into existing self-training methods based on either transformer or CNN, we achieve state-of-the-art performance on two widely used benchmarks and a significant improvement on small-scale categories. Extensive experimental results verify our motivation to transfer the training images to real-world layouts and demonstrate the effectiveness of our multi-task framework in improving scene adaptation performance.

Acknowledgement

The paper is supported by Start-up Research Grant at the University of Macau (SRG2024-00002-FST), and Institute of Collaborative Innovation, University of Macau.

References

[1] Nikita Araslanov and Stefan Roth. 2021. Self-supervised augmentation consistency for adapting semantic segmentation. In *CVPR*.

[2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence* 39, 12 (2017), 2481–2495.

[3] Adriano Cardace, Luca De Luigi, Pierluigi Zama Ramirez, Samuele Salti, and Luigi Di Stefano. 2022. Plugging self-supervised monocular depth into unsupervised domain adaptation for semantic segmentation. In *CVPR*.

[4] Lin Chen, Huaian Chen, Zhixiang Wei, Xin Jin, Xiao Tan, Yi Jin, and Enhong Chen. 2022. Reusing the task-specific classifier as a discriminator: Discriminator-free adversarial domain adaptation. In *CVPR*.

[5] Lin Chen, Zhixiang Wei, Xin Jin, Huaian Chen, Miao Zheng, Kai Chen, and Yi Jin. 2022. Deliberated Domain Bridging for Domain Adaptive Semantic Segmentation. In *NeurIPS*.

[6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence* 40, 4 (2017), 834–848.

[7] Mu Chen, Minghan Chen, and Yi Yang. 2024. UAHOI: Uncertainty-aware robust interaction learning for HOI detection. *Computer Vision and Image Understanding* (2024), 104091.

[8] Mu Chen, Liulei Li, Wenguan Wang, Ruijie Quan, and Yi Yang. 2024. General and Task-Oriented Video Segmentation. In *ECCV*.

[9] Minghao Chen, Hongyang Xue, and Deng Cai. 2019. Domain adaptation for semantic segmentation with maximum squares loss. In *ICCV*.

[10] Mu Chen, Zhedong Zheng, Yi Yang, and Tat-Seng Chua. 2023. Pipa: Pixel-and patch-wise self-supervised learning for domain adaptative semantic segmentation. In *ACM MM*. 1905–1914.

[11] Yuhua Chen, Wen Li, Xiaoran Chen, and Luc Van Gool. 2019. Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. In *CVPR*.

[12] Yangming Cheng, Liulei Li, Yuanyou Xu, Xiaodi Li, Zongxin Yang, Wenguan Wang, and Yi Yang. 2023. Segment and track anything. *arXiv preprint arXiv:2305.06558* (2023).

[13] Jaehoon Choi, Taekyung Kim, and Changick Kim. 2019. Self-ensembling with gan-based data augmentation for domain adaptation in semantic segmentation. In *ICCV*.

[14] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The cityscapes dataset for semantic urban scene understanding. In *CVPR*.

[15] Yuhang Ding, Liulei Li, Wenguan Wang, and Yi Yang. 2024. Clustering propagation for universal medical image segmentation. In *CVPR*.

[16] Hao Feng, Minghao Chen, Jinming Hu, Dong Shen, Haifeng Liu, and Deng Cai. 2021. Complementary Pseudo Labels for Unsupervised Domain Adaptation On Person Re-Identification. *IEEE Transactions on Image Processing* 30 (2021), 2898–2907.

[17] Behnam Gholami, Pritish Sahu, Ognjen Rudovic, Konstantinos Bousmalis, and Vladimir Pavlovic. 2020. Unsupervised multi-target domain adaptation: An information theoretic approach. *IEEE Transactions on Image Processing* 29 (2020), 3993–4002.

[18] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. 2019. Digging into self-supervised monocular depth estimation. In *ICCV*.

[19] Shaohua Guo, Qianyu Zhou, Ye Zhou, Qiqi Gu, Junshu Tang, Zhengyang Feng, and Lizhuang Ma. 2021. Label-free regional consistency for image-to-image translation. In *ICME*.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*.

[21] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. 2018. Cycada: Cycle-consistent adversarial domain adaptation. In *ICML*.

[22] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. 2022. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *CVPR*.

[23] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. 2022. HRDA: Context-Aware High-Resolution Domain-Adaptive Semantic Segmentation. In *ECCV*.

[24] Lukas Hoyer, Dengxin Dai, Haoran Wang, and Luc Van Gool. 2023. MIC: Masked Image Consistency for Context-Enhanced Domain Adaptation. In *CVPR*.

[25] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *CVPR*.

[26] Jiaxing Huang, Dayan Guan, Aoran Xiao, Shijian Lu, and Ling Shao. 2022. Category contrast for unsupervised domain adaptation in visual tasks. In *CVPR*.

[27] Zhengkai Jiang, Yuxi Li, Ceyuan Yang, Peng Gao, Yabiao Wang, Ying Tai, and Chengjie Wang. 2022. Prototypical Contrast Adaptation for Domain Adaptive Segmentation. In *ECCV*.

[28] Menelaos Kanakis, David Bruggemann, Suman Saha, Stamatios Georgoulis, Anton Obukhov, and Luc Van Gool. 2020. Reparameterizing convolutions for incremental multi-task learning without task interference. In *ECCV*.

[29] Myeongjin Kim and Hyeran Byun. 2020. Learning texture invariant representation for domain adaptation of semantic segmentation. In *CVPR*.

[30] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. 2016. Deeper depth prediction with fully convolutional residual networks. In *3DV*.

[31] Kuan-Hui Lee, German Ros, Jie Li, and Adrien Gaidon. 2018. Spigan: Privileged adversarial learning from simulation. *arXiv:1810.03756* (2018).

[32] Seunghun Lee, Wonhyeok Choi, Changjae Kim, Minwoo Choi, and Sunghoon Im. 2022. ADAS: A Direct Adaptation Strategy for Multi-Target Domain Adaptive Semantic Segmentation. In *CVPR*.

[33] Guangrui Li, Guoliang Kang, Wu Liu, Yunchao Wei, and Yi Yang. 2020. Content-Consistent Matching for Domain Adaptive Semantic Segmentation. In *ECCV*.

[34] Liulei Li, Wenguan Wang, and Yi Yang. 2023. Logicseg: Parsing visual semantics with neural logic learning and reasoning. In *ICCV*.

[35] Liulei Li, Wenguan Wang, Tianfei Zhou, Ruijie Quan, and Yi Yang. 2023. Semantic hierarchy-aware segmentation. *IEEE TPAMI* (2023).

[36] Liulei Li, Tianfei Zhou, Wenguan Wang, Jianwu Li, and Yi Yang. 2022. Deep hierarchical semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*

[37] Ruihuang Li, Shuai Li, Chenhang He, Yabin Zhang, Xu Jia, and Lei Zhang. 2022. Class-balanced pixel-level self-labeling for domain adaptive semantic segmentation. In *CVPR*.

[38] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. 2019. Bidirectional learning for domain adaptation of semantic segmentation. In *CVPR*.

[39] Qing Lian, Fengmao Lv, Lixin Duan, and Boqing Gong. 2019. Constructing self-motivated pyramid curriculums for cross-domain semantic segmentation: A non-adversarial approach. In *ICCV*.

[40] Chen Liang, Wenguan Wang, Jiaxu Miao, and Yi Yang. 2022. Gmmseg: Gaussian mixture based generative semantic segmentation models. In *NeurIPS*.

[41] James Chenhao Liang, Tianfei Zhou, Dongfang Liu, and Wenguan Wang. 2023. CLUSTSEG: Clustering for Universal Segmentation. In *ICML*.

[42] Jinliang Liu, Zhedong Zheng, Zongxin Yang, and Yi Yang. 2024. High Fidelity Makeup via 2D and 3D Identity Preservation Net. *ACM Transactions on Multimedia Computing, Communications and Applications* (2024).

[43] Yahao Liu, Jinhong Deng, Xinchen Gao, Wen Li, and Lixin Duan. 2021. Bapanet: Boundary adaptation and prototype alignment for cross-domain semantic segmentation. In *ICCV*.

[44] Yahao Liu, Jinhong Deng, Jiale Tao, Tong Chu, Lixin Duan, and Wen Li. 2022. Undoing the damage of label shift for cross-domain semantic segmentation. In *CVPR*.

[45] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *CVPR*.

[46] Yawei Luo, Ping Liu, Tao Guan, Junqing Yu, and Yi Yang. 2019. Significance-aware Information Bottleneck for Domain Adaptive Semantic Segmentation. In *ICCV*.

[47] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. 2019. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *CVPR*.

[48] Ke Mei, Chuang Zhu, Jiaqi Zou, and Shanghang Zhang. 2020. Instance adaptive self-training for unsupervised domain adaptation. In *ECCV*.

[49] Luke Melas-Kyriazi and Arjun K Manrai. 2021. Pixmatch: Unsupervised domain adaptation via pixelwise consistency training. In *CVPR*.

[50] Viktor Olsson, Wilhelm Tranheden, Juliano Pinto, and Lennart Svensson. 2021. Classmix: Segmentation-based data augmentation for semi-supervised learning. In *WACV*.

[51] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. (2017).

[52] Harsh Rangwani, Sumukh K Aithal, Mayank Mishra, Arihant Jain, and Venkatesh Babu Radhakrishnan. 2022. A closer look at smoothness in domain adversarial training. In *ICML*.

[53] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. 2016. Playing for data: Ground truth from computer games. In *ECCV*.

[54] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. 2016. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *CVPR*.

[55] Suman Saha, Anton Obukhov, Danda Pani Paudel, Menelaos Kanakis, Yuhua Chen, Stamatios Georgoulis, and Luc Van Gool. 2021. Learning to relate depth and semantics for unsupervised domain adaptation. In *CVPR*.

[56] Christos Sakaridis, Dengxin Dai, Simon Hecker, and Luc Van Gool. 2018. Model adaptation with synthetic and real data for semantic dense foggy scene understanding. In *ECCV*.

[57] Antoine Saporta, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. 2021. Multi-target adversarial frameworks for domain adaptation in semantic segmentation. In *ICCV*.

[58] Yuhu Shan, Chee Meng Chew, and Wen Feng Lu. 2020. Semantic-aware short path adversarial training for cross-domain semantic segmentation. *Neurocomputing* 380 (2020), 125–132. https://doi.org/10.1016/j.neucom.2019.11.008

[59] Trevor Standley, Amir Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. 2020. Which tasks should be learned together in multi-task learning?. In *ICML*.

[60] Guolei Sun, Wenguan Wang, Jifeng Dai, and Luc Van Gool. 2020. Mining cross-image semantics for weakly supervised semantic segmentation. In *ECCV*.

[61] Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. 2021. Dacs: Domain adaptation via cross-domain mixed sampling. In *WACV Workshop*.

[62] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. 2018. Learning to adapt structured output space for semantic segmentation. In *CVPR*.

[63] Yi-Hsuan Tsai, Kihyuk Sohn, Samuel Schulter, and Manmohan Chandraker. 2019. Domain adaptation for structured output via discriminative patch representations. In *ICCV*.

[64] Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. 2020. Mti-net: Multi-scale task interaction networks for multi-task learning. In *ECCV*.

[65] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. 2019. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *CVPR*.

[66] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. 2019. Dada: Depth-aware domain adaptation in semantic segmentation. In *ICCV*.

[67] Chao Wang, Zhedong Zheng, Ruijie Quan, and Yi Yang. 2024. Depth-aware blind image decomposition for real-world adverse weather recovery. In *ACM MM*.

[68] Haoran Wang, Tong Shen, Wei Zhang, Ling-Yu Duan, and Tao Mei. 2020. Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation. In *ECCV*.

[69] Qin Wang, Dengxin Dai, Lukas Hoyer, Luc Van Gool, and Olga Fink. 2021. Domain adaptive semantic segmentation with self-supervised depth estimation. In *ICCV*.

[70] Shanshan Wang, ALuSi, Xun Yang, Ke Xu, Huibin Tan, and Xingyi Zhang. 2024. Dual-stream Feature Augmentation for Domain Generalization. In *ACM MM*.

[71] Shanshan Wang, Yiyang Chen, Zhenwei He, Xun Yang, Mengzhu Wang, Quanzeng You, and Xingyi Zhang. 2023. Disentangled representation learning with causality for unsupervised domain adaptation. In *ACM MM*.

[72] Wenguan Wang, Xiankai Lu, Jianbing Shen, David J Crandall, and Ling Shao. 2019. Zero-shot video object segmentation via attentive graph neural networks. In *ICCV*.

[73] Wenguan Wang, Yi Yang, and Yunhe Pan. 2024. Visual Knowledge in the Big Model Era: Retrospect and Prospect. *arXiv preprint arXiv:2404.04308* (2024).

[74] Yaxiong Wang, Yunchao Wei, Xueming Qian, Li Zhu, and Yi Yang. 2021. AINet: Association Implantation for Superpixel Segmentation. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 7058–7067. https://doi.org/10.1109/ICCV48922.2021.00699

[75] Zhonghao Wang, Mo Yu, Yunchao Wei, Rogerio Feris, Jinjun Xiong, Wen-mei Hwu, Thomas S Huang, and Honghui Shi. 2020. Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation. In *CVPR*.

[76] Quanliang Wu and Huajun Liu. 2022. Unsupervised Domain Adaptation for Semantic Segmentation using Depth Distribution. In *NeurIPS*.

[77] Zuxuan Wu, Xin Wang, Joseph E Gonzalez, Tom Goldstein, and Larry S Davis. 2019. Ace: Adapting to changing environments for semantic segmentation. In *ICCV*.

[78] Binhui Xie, Shuang Li, Mingjia Li, Chi Harold Liu, Gao Huang, and Guoren Wang. 2022. SePiCo: Semantic-Guided Pixel Contrast for Domain Adaptive Semantic Segmentation. *arXiv:2204.08808* (2022).

[79] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *NeurIPS* (2021).

[80] Dan Xu, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. 2018. Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In *CVPR*.

[81] Jinyu Yang, Weizhi An, Sheng Wang, Xinliang Zhu, Chaochao Yan, and Junzhou Huang. 2020. Label-driven reconstruction for domain adaptation in semantic segmentation. In *ECCV*.

[82] Jihan Yang, Ruijia Xu, Ruiyu Li, Xiaojuan Qi, Xiaoyong Shen, Guanbin Li, and Liang Lin. 2020. An Adversarial Perturbation Oriented Domain Adaptation Approach for Semantic Segmentation. In *AAAI*.

[83] Xiangpeng Yang, Linchao Zhu, Xiaohan Wang, and Yi Yang. 2024. DGL: Dynamic Global-Local Prompt Tuning for Text-Video Retrieval. In *AAAI*.

[84] Yanchao Yang and Stefano Soatto. 2020. Fda: Fourier domain adaptation for semantic segmentation. In *CVPR*.

[85] Zongxin Yang, Guikun Chen, Xiaodi Li, Wenguan Wang, and Yi Yang. 2024. Doraemongpt: Toward understanding dynamic scenes with large language models (exemplified as a video agent). In *ICML*.

[86] Xiangyu Yue, Yang Zhang, Sicheng Zhao, Alberto Sangiovanni-Vincentelli, Kurt Keutzer, and Boqing Gong. 2019. Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In *ICCV*.

[87] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. 2021. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *CVPR*.

[88] Qiming Zhang, Jing Zhang, Wei Liu, and Dacheng Tao. 2019. Category anchor-guided unsupervised domain adaptation for semantic segmentation. In *NeurIPS*.

[89] Xinyu Zhang, Dongdong Li, Zhigang Wang, Jian Wang, Errui Ding, Javen Qinfeng Shi, Zhaoxiang Zhang, and Jingdong Wang. 2022. Implicit sample extension for unsupervised person re-identification. In *CVPR*.

[90] Xuanmeng Zhang, Zhedong Zheng, Minyue Jiang, and Xiaoqing Ye. 2024. Self-ensembling depth completion via density-aware consistency. *Pattern Recognition* 154 (2024), 110618.

[91] Yurong Zhang, Liulei Li, Wenguan Wang, Rong Xie, Li Song, and Wenjun Zhang. 2023. Boosting video object segmentation via space-time correspondence learning. In *CVPR*.

[92] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Zequn Jie, Xiang Li, and Jian Yang. 2018. Joint task-recursive learning for semantic segmentation and depth estimation. In *ECCV*.

[93] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Yan Yan, Nicu Sebe, and Jian Yang. 2019. Pattern-affinitive propagation across depth, surface normal and semantic segmentation. In *CVPR*.

[94] Yuyang Zhao, Zhun Zhong, Zhiming Luo, Gim Hee Lee, and Nicu Sebe. 2022. Source-free open compound domain adaptation in semantic segmentation. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 10 (2022), 7019–7032.

[95] Zhedong Zheng and Yi Yang. 2020. Unsupervised Scene Adaptation with Memory Regularization in vivo. In *IJCAI*.

[96] Zhedong Zheng and Yi Yang. 2021. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *International Journal of Computer Vision* 129, 4 (2021), 1106–1120.

[97] Zhedong Zheng and Yi Yang. 2022. Adaptive Boosting for Domain Adaptation: Toward Robust Predictions in Scene Segmentation. *IEEE Transactions on Image Processing* 31 (2022), 5371–5382. https://doi.org/10.1109/TIP.2022.3195642

[98] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2019. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision* 127 (2019), 302–321.

[99] Qianyu Zhou, Zhengyang Feng, Qiqi Gu, Jiangmiao Pang, Guangliang Cheng, Xuequan Lu, Jianping Shi, and Lizhuang Ma. 2022. Context-aware mixup for domain adaptive semantic segmentation. *IEEE Transactions on Circuits and Systems for Video Technology* (2022).

[100] Tianfei Zhou and Wenguan Wang. 2024. Prototype-based semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).

[101] Tianfei Zhou, Wenguan Wang, Yazhou Yao, and Jianbing Shen. 2020. Target-aware adaptive tracking for unsupervised video object segmentation. In *CVPR Workshop*.

[102] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. 2018. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *ECCV*.

[103] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. 2019. Confidence regularized self-training. In *ICCV*.