EDAPS: Enhanced Domain-Adaptive Panoptic Segmentation

Suman Saha*
ETH Zurich

Lukas Hoyer*
ETH Zurich

Anton Obukhov
ETH Zurich

Dengxin Dai
Huawei Zurich
Research Center

Luc Van Gool
ETH Zurich,
KU Leuven

Abstract

With autonomous industries on the rise, domain adaptation of the visual perception stack is an important research direction due to the cost savings promise. Much prior art was dedicated to domain-adaptive semantic segmentation in the synthetic-to-real context. Despite being a crucial output of the perception stack, panoptic segmentation has been largely overlooked by the domain adaptation community. Therefore, we revisit well-performing domain adaptation strategies from other fields, adapt them to panoptic segmentation, and show that they can effectively enhance panoptic domain adaptation. Further, we study the panoptic network design and propose a novel architecture (EDAPS) designed explicitly for domain-adaptive panoptic segmentation. It uses a shared, domain-robust transformer encoder to facilitate the joint adaptation of semantic and instance features, but task-specific decoders tailored for the specific requirements of both domain-adaptive semantic and instance segmentation. As a result, the performance gap seen in challenging panoptic benchmarks is substantially narrowed. EDAPS significantly improves the state-of-the-art performance for panoptic segmentation UDA by a large margin of 20% on SYNTHIA-to-Cityscapes and even 72% on the more challenging SYNTHIA-to-Mapillary Vistas. The implementation is available at https://github.com/susaha/edaps.*

1. Introduction

Panoptic segmentation [30] of images is a core computer vision task that jointly solves two related problems – semantic segmentation and instance segmentation. With the rise of robotics and emerging autonomous driving markets, efficient visual perception stacks are in high demand. However, large-scale supervised learning of panoptic segmentation [7, 30, 29, 49, 70, 44] is prohibitively expensive as it requires dense annotations for both semantics and instances, which require time-consuming manual labeling.

Figure 1: EDAPS is an architecture and a collection of recipes, designed specifically for Domain-Adaptive Panoptic Segmentation. It demonstrates a significant improvement over the prior art on the challenging synthetic-to-real benchmarks. As shown above, it is better than CVRN [25] on SYNTHIA → Cityscapes by 9 mPQ.

A promising alternative to circumvent this issue is to learn from abundantly available photo-realistic synthetic images [51, 50] as their ground truth annotations can be automatically generated by the rendering engine. However, often models trained on synthetic data (*source domain*) fail to generalize well on the real data (*target domain*) due to differences in data distribution, known as the *domain gap*.

A common remedy to this problem is to minimize the domain gap using Unsupervised Domain Adaptation (UDA). This field is actively studied for image classification [38, 14, 39, 54, 48], object detection [4, 53, 71, 6, 33], and semantic segmentation [19, 61, 18, 32, 52, 60, 20, 21, 22]. However, UDA for panoptic segmentation is often overlooked and there are only two works, namely CVRN [25] and UniDAPS [75], which address this problem from the synthetic-to-real point of view. Compared to the related semantic segmentation UDA, these approaches achieve only subpar performance. Specifically, the relative performance of the best UDA and fully-supervised learning approaches to panoptic segmentation (64% in UniDAPS [75])

---

*These authors contributed equally to this work.

is much smaller than that of semantic segmentation (88% in DAFormer [20]).

To understand the root cause of the identified performance gap, we revisit the progress of UDA in semantic segmentation, lift the well-performing UDA strategies to panoptic segmentation, and show that they can effectively enhance panoptic segmentation UDA.

Further, we revisit the panoptic network design and conduct a study of principal architecture designs for panoptic segmentation with respect to their UDA capabilities. We show that previous UDA methods took sub-optimal design choices. While separating the networks for both tasks prevents the network from jointly adapting task-shared features from the source to the target domain (see Fig. 2 a), a shared encoder-decoder cannot accommodate the different needs (such as task-specific knowledge or architecture design) when adapting both tasks (see Fig. 2 b).

To address these problems, we propose EDAPS, a network architecture that is particularly designed for domain-adaptive panoptic segmentation. It uses a shared transformer [63] encoder to facilitate the joint adaptation of semantic and instance features, but task-specific decoders tailored for the specific requirements of both domain-adaptive semantic segmentation and domain-adaptive instance segmentation (see Fig. 2 c). Specifically, separate decoders can learn task-specific parameters, mitigating the issue of conflicting gradients [74], and their architectures can be independently chosen to best suit semantic or instance segmentation. Even though this design was used in the supervised setting, we are the first to identify its strong potential specifically for panoptic UDA.

Utilizing the enhanced panoptic UDA and the enhanced domain-adaptive panoptic network design, EDAPS shows significant performance gains over the prior works on a number of standard perception benchmarks (Fig. 1). EDAPS improves the state-of-the-art mPQ from 33.0 to 41.2 on On SYNTHIA → Cityscapes and from 21.3 to 36.6 on SYNTHIA → Mapillary Vistas. EDAPS trains semantic segmentation and instance segmentation tasks together using a joint training optimization. Therefore, EDAPS can be trained end-to-end in a single stage and the training only requires 21 hours on a single RTX 2080 Ti GPU, which improves its applicability for the community.

The main contribution of EDAPS is the novel combination of network components and UDA strategies on a system level, which results in a *major* relative gain in the state-of-the-art panoptic segmentation UDA performance of 20% on SYNTHIA-to-Cityscapes and even 72% on the more challenging SYNTHIA-to-Mapillary Vistas. In particular, we carefully study various principal panoptic architectures for their UDA capability and identify a network design that is particularly suited for panoptic UDA. It improves the UDA performance over strong baselines, while
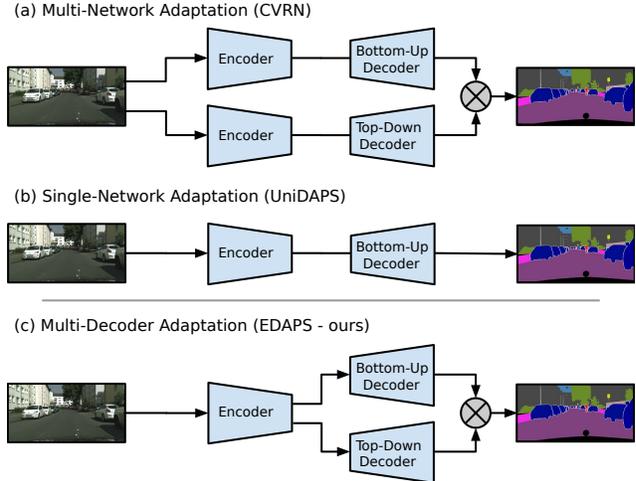

Figure 2: Overview of network architectures employed in prior panoptic UDA works and the proposed architecture. CVRN [25] (a) does not share parameters between panoptic branches, whereas UniDAPS [75] (b) resorts to the opposite extreme and shares everything. With EDAPS (c), we propose to share the encoder but to use task-specific decoders, which facilitates panoptic domain adaptation.

being parameter-efficient and fast at inference. Further, this is the first paper that lifts recent UDA techniques to panoptic segmentation and systematically studies their effectiveness for panoptic UDA.

2. Related Work

We outline two cornerstone classes of works related to UDA in Semantic and Panoptic Segmentation settings.

UDA for Semantic Segmentation Methods in this class take input images from source and target domains along with the source semantic ground truth label. A supervised cross-entropy loss is computed on the source image semantic prediction. A UDA semantic loss (e.g., adversarial [64] or pseudo-label-based self-training [80] loss) is used for domain alignment that operates on the semantic feature space. UDA for semantic segmentation is one of the most popular tasks in dense prediction, and a large number of works in this category exist in the literature [64, 61, 18, 32, 28, 79, 60, 43, 9, 1, 76, 80, 73, 68, 41, 13, 5, 19, 27, 77, 37, 52]. The first approach of using an adversarial loss helps to align the source and target domain distributions at the input [18, 16], at the feature-level [61, 19], at the output [61, 64], or at the patch level [62]. In the second approach of using self-training [80, 32, 77, 81], pseudo-labels are generated on the unannotated target domain images either offline [56, 73, 80, 81] or online [76, 66, 60, 79, 20, 21]. The pseudo-labels can be stabilized with consistency regularization [59, 57] based on pseudo-label prototypes [76], differ-

ent data augmentation schemes [1, 9, 43, 23], cross-domain mixup strategies [60, 79], and multiple resolutions [21, 22].

UDA for Panoptic Segmentation   Considering the used network architectures, there are distinct shortcomings in both CVRN [25] and UniDAPS [75]. On the one side, CVRN requires the training of two separate networks for semantic segmentation and instance segmentation, which is time-consuming, parameter-inefficient, and slow during inference (Fig. 2). Furthermore, no task-agnostic knowledge can be shared during the adaptation process by separating the semantic from the instance network. The approach relies on expensive multi-stage training (i.e., training, pseudo-label generation, and retraining with fused labels). On the other side, UniDAPS proposed to adapt panoptic segmentation within a unified network [2], which directly predicts panoptic segments instead of separate semantic and instance masks. While this simple concept is intriguing and performs well in a supervised setting [2, 8, 34], unified network architectures are not inherently suited for UDA as shown in [75]. Even with specific UDA strategies to compensate for that, UniDAPS achieves only a small improvement over CVRN.

The authors of [20, 22] pointed out that architectures and training schemes that work for supervised learning might not work as desirably for UDA, and special care is required (in network design and training recipes) to address the domain shift problem effectively. Outside of the synthetic-to-real domain, the authors of [35] proposed a panoptic UDA approach to microscopy image analysis. In this paper, we propose a domain-adaptive panoptic segmentation framework that carefully selects the network design and training recipes tailored explicitly for UDA.

3. Method

In this section, we first recap panoptic image segmentation. Second, we present our enhanced panoptic segmentation UDA pipeline. Third, we define principal panoptic architectures for a systematic analysis of their UDA capabilities. And finally, we introduce our EDAPS network architecture, specifically designed for panoptic UDA.

3.1. Supervised Panoptic Segmentation

Panoptic segmentation is commonly approached by decomposing the task into a semantic segmentation and an instance segmentation component so that the panoptic segmentation loss $\mathcal{L}_{PS}$ is composed of a semantic and instance loss. $\mathcal{L}_{PS} = \mathcal{L}_{sem} + \mathcal{L}_{inst}$. The semantic segmentation loss typically uses a pixel-wise cross-entropy loss to assign each pixel of an image to one class from a pre-defined set. Instance segmentation further distinguishes instances within classes with countable entities, the so-called thing-classes such as *car* or *person* (as opposed to uncountable stuff-classes such as *road* or *sky*). Instance segmentation can be
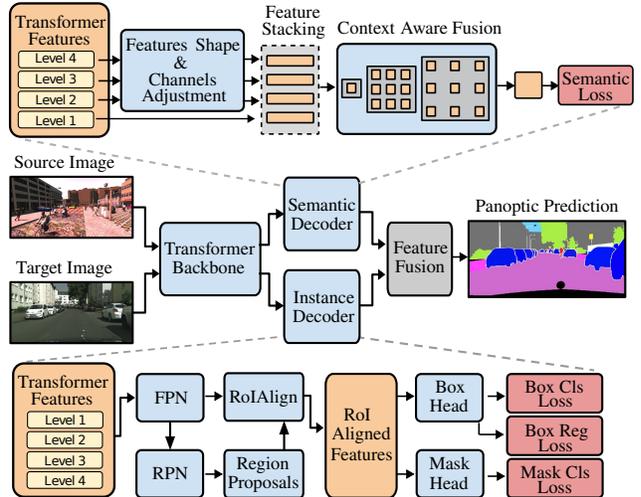

Figure 3: The proposed EDAPS (Enhanced Domain-Adaptive Panoptic Segmentation) network architecture. EDAPS is built with many design choices tailored to UDA in mind. It achieves competitive results on challenging synthetic-to-real panoptic segmentation benchmarks.

approached in two ways. In top-down approaches such as Mask R-CNN [17], instances are predicted based on proposals in the form of bounding boxes and instances masks. In bottom-up approaches such as Panoptic-DeepLab [7], the instances are grouped on a pixel-level without proposals, for example using pixel-wise instance center and offset heatmaps. During inference, semantic and instance segmentations are deterministically fused into a panoptic segmentation representation.

3.2. Enhanced Panoptic UDA

In the UDA setting, a neural network $\mathcal{F}_\theta$ is trained on annotated source domain images $\mathcal{X}^{(s)} = \{x_i^{(s)}\}_{i=1}^{N^{(s)}}$ and unannotated target domain images $\mathcal{X}^{(t)} = \{x_i^{(t)}\}_{i=1}^{N^{(t)}}$ with $x_i^{(s)}, x_i^{(t)} \in \mathbb{R}^{H \times W \times 3}$ with the objective of achieving good performance on the target domain. As panoptic segmentation ground truth $\mathcal{Y}^{(s)} = \{y_i^{(s)}\}_{i=1}^{N_S}$ is only available on the source domain, the panoptic model can only be trained with source data in a supervised fashion. In particular, the source loss term $\mathcal{L}^{(s)} = \mathcal{L}_{PS}(\hat{y}^{(s)}, y^{(s)})$.

Models trained using the supervised loss $\mathcal{L}^{(s)}$ on the source domain often exhibit poor generalization on the target domain due to the "domain gap" between the distributions of source and target data. To adapt the model to the target domain, an additional unsupervised loss term $\mathcal{L}^{(t)}$ is computed on the target domain. The overall loss combines both the source and target loss terms $\mathcal{L}_{UDA} = \mathcal{L}^{(s)} + \mathcal{L}^{(t)}$.

Various unsupervised target loss terms were proposed in the literature, mostly based on adversarial training [61, 62,

65] or self-training [80, 77, 42, 60, 76, 24].

Self-Training on Target Domain In this work, we resort to self-training for adapting the network to the target domain similar to previous panoptic UDA methods [71, 75]. In self-training, the model $\mathcal{F}_\theta$ is trained with high-confidence pseudo-labels [31] on the target domain using a weighted cross-entropy loss

$$\mathcal{L}^{(t)} = -\sum_{i,j,c} q_{i,j}^{(t)} \left(p^{(t)} \log(\hat{y}^{(t)}) \right)_{i,j,c} \quad (1)$$

which helps the network gradually adapt to the target domain. Here, $p^{(t)}$ denotes the pseudo-label and $q^{(t)}$ its confidence estimate.

The pseudo-labels can be generated using the predictions of a teacher network $\mathcal{T}_\phi$. As the output of $\mathcal{T}_\phi$ is a map of per-pixel probabilities, a single class assignment is done using the mode of categorical distribution, which could be converted back to one-hot categorical form via the Iverson bracket $[\cdot]$:

$$p_{i,j,c}^{(t)} = \left[c = \arg\max_{c'} \mathcal{T}_\phi(x^{(t)})_{i,j,c'} \right] . \quad (2)$$

Since the pseudo-labels are the predictions of a network, they are not always correct, and their quality can be estimated by a confidence estimate $q^{(t)}$ [80, 42, 60, 24]

$$q^{(t)} = \frac{\sum_{i,j} \left[\max_{c'} \mathcal{T}_\phi(x^{(t)})_{(ijc')} > \tau \right]}{H \cdot W} . \quad (3)$$

To stabilize the quality of the pseudo-labels, we resort to the mean teacher framework [59], which is commonly used in semantic segmentation UDA [1, 60, 36, 20]. The parameters of the teacher network $\mathcal{T}_\phi$ are updated with the exponential moving average of the parameters of the student network $\mathcal{F}_\theta$ at every training step t:

$$\phi_{t+1} \leftarrow \alpha\phi_t + (1 - \alpha)\theta_t . \quad (4)$$

Alongside self-training, we adopt a consistency training [55, 59, 57], in which the student network $\mathcal{F}_\theta$ is trained on augmented target images, whereas, the mean teacher network $\mathcal{F}_\phi$ predicts pseudo-labels for actual target images. The augmented images are generated by mixing pixels of the source and target images. The mixing is done following the Class-Mix strategy [47], which has been successfully applied to semantic segmentation UDA [60, 36, 20]: first, we randomly select $N/2$ semantic classes among N classes present in the source image. Next, the pixels belonging to these selected $N/2$ semantic classes are pasted into the target image, resulting in a new augmented image.

Furthermore, we leverage the recent findings in semantic segmentation UDA [24] to further enhance the performance of the panoptic UDA. More specifically, we adopt rare class sampling and a feature regularization loss based on ImageNet features to learn domain-invariant representations for panoptic UDA.

Rare Class Sampling (RCS) Most datasets are imbalanced, so certain classes are underrepresented. It can hurt the adaptation process if a certain class is sampled with a low frequency [20]. Therefore, we follow DAFormer [20] and sample images with rare classes more frequently. The sampling probability of class c is

$$P(c) = \frac{e^{(1-f_c)/T}}{\sum_{c'=1}^{C} e^{(1-f_{c'})/T}} . \quad (5)$$

where f_c denotes the class frequency in the source dataset and T is a temperature parameter. Given a sampled class $c \sim P$, a source sample is drawn from the subset containing class c: $x^{(s)} \sim \text{uniform}(\mathcal{X}^{(s,c)})$.

ImageNet Feature Distance (FD) The encoder network in panoptic segmentation UDA is usually pre-trained using ImageNet classification. Given that ImageNet is a real-world dataset, its features can be valuable when adapting to a real-world target domain. However, it was observed that some relevant features could be forgotten during the adaptation process [20]. Therefore, it can be useful to regularize the adaptation process with an ImageNet feature distance loss term [20]

$$\mathcal{L}_{FD} = \left\| \mathcal{F}_{\text{ImageNet}}(x^{(s)}) - \mathcal{F}_\theta(x^{(s)}) \right\|_2 , \quad (6)$$

where F_{ImageNet} denotes the frozen ImageNet model. As ImageNet annotates thing classes, it is beneficial to constraint the feature distance to image regions that are labeled as thing classes [20].

3.3. Principal Panoptic Architectures for UDA

To further facilitate panoptic UDA, we aim to design a network architecture particularly tailored for domain-adaptive panoptic segmentation. Several different network architectures have been proposed for panoptic segmentation in the supervised setting. However, the effect of the different designs on the performance in a domain adaptation setting is not well studied. Therefore, we analyze four principal panoptic architectures in a fair comparison using the same network building blocks and the same enhanced panoptic UDA strategy in §4.3. The systematic comparison is the foundation for our EDAPS architecture.

M-Net uses two separate encoder-decoder networks for semantic and instance segmentation (see Fig. 2 a). This design was deployed in CVRN [25]. The two encoders utilize the same architecture but the decoder architecture are task-specific. A bottom-up decoder is used for semantic segmentation while a top-down decoder is used for instance segmentation. The design of the network building blocks is further detailed in §3.4. A potential disadvantage of M-Net is the use of separate encoders, which prevents the network from jointly adapting task-shared features from the source to the target domain.

S-Net shares the encoder and decoder for both tasks (see Fig. 2 b), which is similar to the architecture of UniDAPS [75]. As the decoder is shared, it is not possible to have task-specific designs, which can be disadvantageous if different decoder architectures have better domain adaptation properties for semantic and instance segmentation. Also, shared decoders can favor negative transfer of task-specific knowledge.

M-Dec-BU shares the encoder but splits the decoders to enable task-specific knowledge. Both decoders use a bottom-up design.

M-Dec-TD (EDAPS) shares the encoder but splits the decoders to enable task-specific knowledge. The instance decoder uses a top-down approach in contrast to the bottom-up semantic decoder (see Fig. 2 c). We adopt this design in our EDAPS architecture. In the following section, we motivate and detail the design choices of the EDAPS architecture.

3.4. EDAPS Network Architecture

An overview of the proposed EDAPS architecture is presented in Fig. 3. On the one side, EDAPS utilizes a shared transformer backbone, which has an improved domain-robustness compared to CNNs [45, 20]. Sharing the encoder is advantageous compared the separate networks as it can facilitate the joint adaptation of semantic and instance features from the source to the target domain. On the other side, EDAPS uses task-specific decoders with different design tailored for the specific requirements of domain-adaptive semantic segmentation and domain-adaptive instance segmentation. In that way, EDAPS combines the strength of previous methods (see Fig. 2).

Shared Hierarchical Transformer Encoder Transformers have shown domain-robust properties beneficial for UDA [20, 72, 58]. Therefore, EDAPS uses a Mix Transformer (MiT-B5) [69] as its encoder network, which is tailored for dense predictions by generating coarse (high-resolution) and fine-grained (low-resolution) features at different levels. Unlike 16 patches of ViT [12], MiT divides an image into relatively smaller patches of shape 4×4; this facilitates preserving finer details helpful in improving semantic segmentation. The patches are processed with four transformer blocks into multi-level features at $\{1/4, 1/8, 1/16, 1/32\}$ scales of the original image shapes. Moreover, an efficient self-attention is used to cope with the computation bottleneck of transformer encoders, i.e., the computational complexity of the self-attention operation is minimized by reducing the input sequence length $(H \times W)$ using a reduction ratio as in [67]. Overlapping patch merging is used to downsample feature maps which helps preserve local continuity. These properties of MiT make it possible to learn robust domain-invariant features

for UDA panoptic segmentation using limited computing resources (i.e., on a single GPU).

Bottom-Up Semantic Decoder We adopt a domain-robust context-aware feature fusion decoder [24] as our semantic decoder, which has two main benefits over traditional DeepLab decoders [3, 15] or MLP-based decoders [67, 69, 78]. Firstly, it allows exploiting context information alongside local information in the decoder, which is beneficial for domain-robust semantic segmentation [26]. Secondly, instead of only using the context information coming from the bottleneck features [3, 15], it fuses context encoded by features at different levels of the backbone network. The high-resolution features from the earlier layers encode low-level concepts helpful in better semantic segmentation. The multi-level context-aware feature fusion is learned using a series of 3×3 depthwise separable convolutions [10] with different dilation rates.

Top-Down Instance Decoder Based on the findings of the network study in §4, EDAPS resorts to a proposal-based top-down instance decoder, which exhibits better domain-adaptive properties compared to a bottom-up decoder.

Specifically, we follow Mask R-CNN [17] and use a Region Proposal Network (RPN) to predict candidate object bounding boxes. It is trained using a binary cross-entropy (CE) loss for bounding box classification (object or no object), and an L_1 loss for box regression. Given the bounding boxes predicted by RPN, a Region-of-Interest Alignment (RoIAlign) layer is used to extract features from each box. The extracted features are used as inputs to the box and mask heads. The box head is trained to predict the bounding boxes and class labels, and the mask head is trained to predict the class-agnostic binary instance masks. The box head is trained using CE loss for object classification and L_1 loss for box regression over all thing classes. The mask head predictions are penalized using the binary CE loss.

Bottom-Up Instance Decoder (Baseline) For the network studies with a bottom-up instance decoder, we adapt the strong domain-robust decoder from [24] from semantic segmentation to instance segmentation by replacing the classification head (for semantic prediction) with two 1×1 convolutional layers that predict the instance centers and offsets. It is trained using the losses proposed in [7].

Feature Fusion We follow Panoptic-DeepLab [7] to fuse semantic and instance predictions into the final panoptic segmentation upon inference. Class-agnostic instance segmentation maps are generated by selecting the top-k instance predictions with a detection score above a certain threshold. The resulting instance segmentation is fused with the predicted semantic map by a majority-voting rule to generate the final panoptic segmentation.
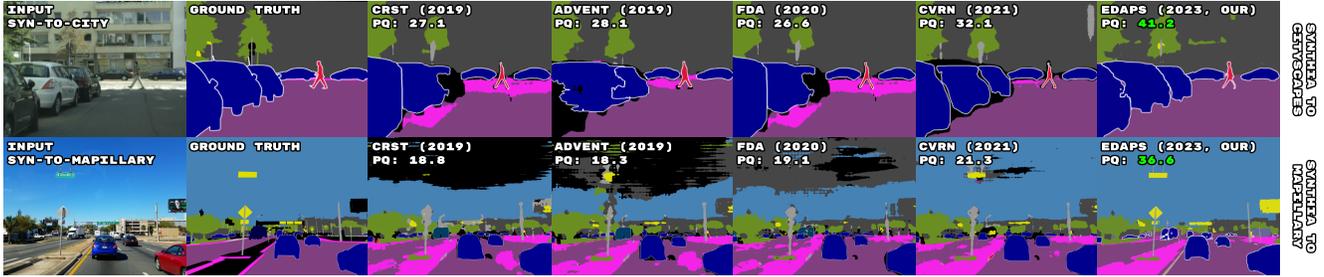
Figure 4: Visual comparison of the panoptic segmentation quality of EDAPS (our method) with the prior works CRST [81], FDA [73], AdvEnt [64], and CVRN [25] on the two UDA benchmarks: SYNTHIA → Cityscapes (top) and SYNTHIA → Mapillary Vistas (bottom).

4. Experiments

4.1. Implementation Details

Datasets We evaluate EDAPS for the common setting of synthetic-to-real adaptation. As a source dataset, we use Synthia [51], which contains 9,400 synthetic images and panoptic labels. As target datasets, we use Cityscapes [11] and Mapillary Vistas [46]. Cityscapes [11] consists of 2,975 training and 500 validation images of European street scenes. Mapillary Vistas [46] is a large-scale autonomous driving dataset consisting of 18,000 training and 2,000 validation images.

Training We follow DAFormer [20] and train EDAPS with AdamW [40] for 40k iterations, a batch size of 2, a crop size of 512×512, a learning rate of 6×10^{-5} with a linear warmup for 1.5k iterations and polynomial decay afterward, and a weight decay of 0.01. The experiments are conducted on a RTX 2080 Ti GPU with 11 GB memory. We use the same data augmentation parameters as in DACS [60]. To encourage the under-represented classes (in the source domain) to be sampled more frequently, we use RCS [20] and set the RCS temperature to 0.01. Following DAFormer [20], we also use the ImageNet feature distance loss to preserve information about certain thing classes encoded in the ImageNet features. We set this loss weight $\lambda_{FD} = 0.005$. For the bottom-up instance head, we use the center loss weight $\lambda_{\text{heatmap}} = 10.0$, offset loss weight $\lambda_{\text{offs}} = 0.1$. For the top-down instance head, we set the following loss weights to 1.0: $\lambda_{\text{RPN-cls}}$, $\lambda_{\text{RPN-box}}$, $\lambda_{\text{RoI-cls}}$, $\lambda_{\text{RoI-box}}$, $\lambda_{\text{RoI-mask}}$.

Evaluation Metrics For evaluating panoptic segmentation, we use the panoptic quality (PQ) metric [30] that captures performance for all classes (including stuff and things) in an interpretable and unified manner. PQ can be seen as the multiplication of a segmentation quality (SQ) term and a recognition quality (RQ) term, i.e., PQ = SQ × RQ. We report PQ for each category. Mean SQ (mSQ), mean RQ (mSQ), and mean PQ (mPQ) are average scores over all categories. For evaluating instance and semantic segmentation, we use the mIoU and mAP metrics respectively.

4.2. Comparison with the State of the Art

First, we compare our EDAPS with the state-of-the-art methods for panoptic segmentation UDA on SYNTHIA → Cityscapes (Tab. 1), SYNTHIA → Mapillary (Tab. 2), Cityscapes → Foggy Cityscapes, and Cityscapes → Mapillary Vistas (Tab. 3). It can be seen that EDAPS consistently outperforms previous methods with a large margin in all aggregated metrics (mSQ, mRQ, mPQ). Specifically, EDAPS improves the mPQ from 34.2 to 41.2 on SYNTHIA → Cityscapes and from 21.3 to 36.6 SYNTHIA → Mapillary, which is a respective improvement of remarkable 20% and 72% over previous works. Tab. 1 shows that EDAPS significantly improves both the recognition (+9.3 mRQ) as well as the segmentation quality (+5.8 mSQ) over previous SOTA. Both can be also observed qualitatively in Fig. 4. EDAPS particularly improves the performance of classes that previous methods struggled with, such as wall, pole, traffic light, traffic sign, person, rider on Cityscapes (Tab. 1), and road, sidewalk, building, wall, pole, traffic light, traffic sign, vegetation, sky, bus on Mapillary (Tab. 2). Additional benchmarks for other domain gaps are presented in Tab. 3. EDAPS also improves the SOTA on Cityscapes → Foggy Cityscapes and Cityscapes → Mapillary Vistas with respective improvements of 47% and 23% over UniDAPS and CVRN with fewer parameters and faster inference speed (Tab. 5). As the different panoptic UDA methods resort to different network architectures, it has to be considered that they have a varying inherent capability to learn panoptic segmentation. Therefore, we compare both the UDA performance and the supervised performance (as indicators of the capability of a network architecture in Tab. 4. Further, we normalize the UDA performance by the supervised performance (mPQ$_{\text{Rel}}$) for a more fair comparison. It can be seen that even though UniDAPS uses a more powerful network than CVRN, as indicated by the higher supervised performance, this does not translate well to UDA performance.

Table 1: Comparison with state-of-the-art methods on SYNTHIA → Cityscapes benchmark for UDA panoptic segmentation. For clarity, per class PQs are reported. The results of EDAPS are averaged over 3 random seeds.

UDA Method	road	sidewalk	building	wall	fence	pole	light	sign	veg	sky	person	rider	car	bus	m.bike	bike	mSQ	mRQ	mPQ
FDA [73]	79.0	22.0	61.8	1.1	0.0	5.6	5.5	9.5	51.6	70.7	23.4	16.3	34.1	31.0	5.2	8.8	65.0	35.5	26.6
CRST [81]	75.4	19.0	70.8	1.4	0.0	7.3	0.0	5.2	74.1	69.2	23.7	19.9	33.4	26.6	2.4	4.8	60.3	35.6	27.1
AdvEnt [64]	87.1	32.4	69.7	1.1	0.0	3.8	0.7	2.3	71.7	72.0	28.2	17.7	31.0	21.1	6.3	4.9	65.6	36.3	28.1
CVRN [25]	86.6	33.8	74.6	3.4	0.0	10.0	5.7	13.5	80.3	76.3	26.0	18.0	34.1	37.4	7.3	6.2	66.6	40.9	32.1
UniDAPS [75]	73.7	26.5	71.9	1.0	0.0	7.6	9.9	12.4	81.4	77.4	27.4	23.1	**47.0**	**40.9**	12.6	15.4	64.7	42.2	33.0
↳ w/ CVRN Net [75]	**87.7**	34.0	73.2	1.3	0.0	8.1	9.9	6.7	78.2	74.0	37.6	25.3	40.7	37.4	15.0	**18.8**	66.9	44.3	34.2
EDAPS (Ours)	77.5	**36.9**	**80.1**	**17.2**	**1.8**	**29.2**	**33.5**	**40.9**	**82.6**	**80.4**	**43.5**	**33.8**	45.6	35.6	**18.0**	2.8	**72.7**	**53.6**	**41.2**

Table 2: Comparison with state-of-the-art methods on SYNTHIA → Mapillary Vistas benchmark for UDA Panoptic Segmentation. For clarity, per class PQs are reported. The results of EDAPS are averaged over 3 random seeds.

UDA Method	road	sidewalk	building	wall	fence	pole	light	sign	veg	sky	person	rider	car	bus	m.bike	bike	mSQ	mRQ	mPQ
FDA [73]	44.1	7.1	26.6	1.3	0.0	3.2	0.2	5.5	45.2	61.3	30.1	13.9	39.4	12.1	8.5	7.0	63.8	26.1	19.1
CRST [81]	36.0	6.4	29.1	0.2	0.0	2.8	0.5	4.6	47.7	68.9	28.3	13.0	42.4	13.6	5.1	2.0	63.9	25.2	18.8
AdvEnt [64]	27.7	6.1	28.1	0.3	0.0	3.4	1.6	5.2	48.1	66.5	28.4	13.4	40.5	14.6	5.2	3.3	63.6	24.7	18.3
CVRN [25]	33.4	7.4	32.9	1.6	0.0	4.3	0.4	6.5	50.8	76.8	30.6	15.2	44.8	18.8	7.9	**9.5**	65.3	28.1	21.3
EDAPS (Ours)	**77.5**	**25.3**	**59.9**	**14.9**	0	**27.5**	**33.1**	**37.1**	**72.6**	**92.2**	**32.9**	**16.4**	**47.5**	**31.4**	**13.9**	3.7	**71.7**	**46.1**	**36.6**

Table 3: Comparison on Cityscapes → Foggy Cityscapes and Cityscapes → Mapillary Vistas.

Cityscapes → Foggy Cityscapes			Cityscapes → Mapillary Vistas		
UDA Method	#Params	mPQ	UDA Method	#Params	mPQ
UniDAPS [75]	58.7 M	37.6	CVRN [25]	185.5 M	33.5
EDAPS-MiTB2	47.6 M	55.1	EDAPS-MiTB5	104.9 M	41.2

Table 5: Efficiency comparison on an RTX 2080 Ti.

Network architecture	#Parameters	Inference Speed	mPQ$_{SYN→CS}$
CVRN [25]	185.5 M	0.36 fps	32.1
UniDAPS [75]	58.7 M	7.24 fps	33.0
EDAPS w/ MiT-B2	47.6 M	9.03 fps	37.1
EDAPS w/ MiT-B3	67.5 M	7.65 fps	38.9
EDAPS w/ MiT-B5	104.9 M	5.84 fps	41.2

Table 4: Comparison of the mPQ for source-only training, UDA (SYNTHIA → Cityscapes), and supervised oracle training (Cityscapes) along with the relative UDA performance mPQ$_{Rel}$=mPQ$_{UDA}$/mPQ$_{Sup}$.

UDA Method	mPQ$_{Src-Only}$	mPQ$_{UDA}$	mPQ$_{Sup}$	mPQ$_{Rel}$
CVRN [25]	20.1	32.1	47.7	67.3%
UniDAPS [75]	18.3	33.0	51.9	63.6%
EDAPS (Ours)	**21.6**	**41.2**	**56.6**	**72.7%**

Therefore, UniDAPS has a significantly lower relative UDA performance than CVRN. The lower relative performance of UniDAPS could be caused by the additional transformer encoder-decoder architecture added to the backbone, which is not pre-trained on ImageNet and might overfit to the source domain more easily. In contrast, our method can increase the relative UDA performance alongside with the best UDA and the best supervised performance. This means that EDAPS can effectively narrow the domain gap to supervised learning.

We compare EDAPS with previous works from an efficiency viewpoint in Tab. 5. EDAPS with a MiT-B5 backbone (default) increases the inference speed by a factor of 16 compared to CVRN, demonstrating the efficiency of EDAPS. Compared to UniDAPS, EDAPS w/ MiT-B5 reaches 80% of its inference speed, which is due to the additional instance decoder of EDAPS. However, given the large gains in the quality of the adapted panoptic segmentation, this is an acceptable trade-off. Even with a smaller MiT-B2 backbone, EDAPS outperforms UniDAPS by +4.1 mPQ on SYNTHIA → Cityscapes while requiring fewer parameters and with faster inference time (9.0 vs. 7.2 fps).

4.3. Study of Panoptic Architectures for UDA

To gain insights into the influence of the network architecture on the UDA performance, we compare the four principal architectures from §3.3. For all network variants, we use the enhanced network components of EDAPS, i.e., the transformer encoder and the domain-robust decoder as well

Table 6: Network topology study as detailed in §3.3. Mean and standard deviation are computed over 3 random seeds.

Network	mAP	mIoU	mSQ	mRQ	mPQ
S-Net	$7.0_{\pm0.4}$	$57.8_{\pm1.1}$	$72.0_{\pm0.4}$	$43.8_{\pm0.8}$	$34.0_{\pm0.7}$
M-Net	$23.3_{\pm0.3}$	$56.6_{\pm0.6}$	$71.7_{\pm0.3}$	$50.0_{\pm0.5}$	$38.1_{\pm0.3}$
M-Dec-BU	$17.6_{\pm1.5}$	$60.3_{\pm0.6}$	$73.9_{\pm0.4}$	$49.9_{\pm0.3}$	$39.0_{\pm0.2}$
M-Dec-TD (EDAPS)	$34.4_{\pm0.5}$	$57.5_{\pm0.0}$	$72.7_{\pm0.2}$	$53.6_{\pm0.5}$	$41.2_{\pm0.4}$

Table 7: Class-wise PQ comparison of EDAPS and M-Net.

Network	road	sidew.	build.	wall	fence	pole	light	sign	veg.	sky	pers.	rider	car	bus	m.bike	bike
M-Net	75.2	37.5	80.4	15.7	2.0	27.1	36.2	38.0	79.8	80.0	38.2	25.8	38.9	28.0	6.1	0.3
EDAPS	77.5	36.9	80.1	17.2	1.8	29.2	33.5	40.9	82.6	80.4	43.5	33.8	45.6	35.6	18.0	2.8
Gain	+2.3	-0.6	-0.3	+1.5	-0.2	+2.1	-2.7	+2.9	+2.8	+0.4	**+5.3**	**+8.0**	**+6.7**	**+7.6**	**+11.9**	+2.5

as the same enhanced UDA strategy for a fair comparison.

Tab. 6 shows that S-Net achieves the lowest mPQ of 34.0. In contrast, M-Net achieves 38.1 mPQ, which is +4.1 mPQ better. This demonstrates the importance of task-specific networks (components) to learning task-specific knowledge in the UDA setting. Still, sharing features across tasks in a common encoder for UDA is useful, as shown by M-Dec-BU, which gains a performance improvement of +0.9 mPQ over M-Net. However, the symmetric architecture with bottom-up decoders for both semantics and instances is sub-optimal, as can be seen when comparing M-Dec-BU with M-Dec-TD. With a top-down decoder for instance segmentation, a significant improvement of +2.2 mPQ can be achieved over M-Dec-BU. In particular, the instance mAP improves by +16.8. This shows that the top-down approach is more domain-robust for instance segmentation than the bottom-up approach.

The network study in Tab. 6 also allows a fair comparison of EDAPS and CVRN as M-Net uses the CVRN architecture with the enhanced components of EDAPS, i.e., Transformer encoder, Mean Teacher, RCS, and FD. Therefore, M-Net has a higher 38.1 mPQ compared to the original CVRN 32.1 mPQ. Still, EDAPS outperforms M-Net by +3.1 mPQ. These improvements come from the joint training of EDAPS for semantic and instance segmentation (compared to disjoint M-Net training), allowing the shared encoder to learn better domain-robust instance and semantic features. In particular, EDAPS improves the instance mAP by +11.1 over M-Net demonstrating that particularly the instance head benefits from the shared features. The class-wise comparison in Tab. 7 reveals that EDAPS mostly benefits difficult thing-classes such as rider or motorbike, showing that shared features are particularly important for instances that are difficult to adapt.

Table 8: Ablation study of the UDA strategies Self-Training (Self-Tr.), Mean Teacher (MT), ImageNet Feature Distance (FD), and Rare Class Sampling (RCS). Mean and standard deviation are provided over 3 random seeds.

Self-Tr.	MT	FD	RCS	mAP	mIoU	mSQ	mRQ	mPQ
				$22.0_{\pm1.0}$	$33.0_{\pm1.8}$	$60.9_{\pm2.6}$	$29.3_{\pm1.7}$	$21.6_{\pm1.2}$
✓				$34.7_{\pm0.9}$	$54.0_{\pm0.1}$	$71.5_{\pm0.4}$	$49.1_{\pm0.4}$	$37.5_{\pm0.2}$
✓	✓			$35.7_{\pm1.7}$	$56.3_{\pm0.9}$	$69.3_{\pm1.6}$	$50.7_{\pm1.5}$	$38.9_{\pm1.2}$
✓	✓	✓		$35.2_{\pm0.9}$	$56.7_{\pm0.9}$	$70.7_{\pm1.8}$	$52.1_{\pm1.1}$	$39.7_{\pm0.9}$
✓	✓	✓	✓	$34.4_{\pm0.5}$	$57.5_{\pm0.0}$	$72.7_{\pm0.2}$	$53.6_{\pm0.5}$	$41.2_{\pm0.4}$

4.4. UDA Ablation Study

To better understand the influence of the UDA components that we newly introduced to panoptic UDA, we ablate them in Tab. 8. While the source-only model performs at 21.6 mPQ, the UDA baseline with basic self-training achieves a performance of 37.5 mPQ. This is already higher than previous state-of-the-art methods, emphasizing the strength of the proposed EDAPS architecture. The self-training can be improved by +1.4 mPQ using an EMA teacher for pseudo-label generation. ImageNet Thing-Class Feature Distance (FD) further increases performance by +0.8 mPQ. When further integrating Rare Class Sampling (RCS), the performance gains another +1.5 mPQ. This shows that the EMA teacher, FD and RCS are valuable components for panoptic UDA. Empirically, we found that additional instance pseudo-labels do not improve the performance of EDAPS. We assume that semantic self-training is sufficient to bridge the domain gap at the level of the shared encoder features. For simplicity, we therefore only use semantic pseudo-labels.

5. Conclusions

Previous approaches to domain-adaptive panoptic segmentation either follow inefficient and expensive adaptation techniques, or their network architecture and training strategies are influenced by supervised learning. In this work, we addressed these issues by carefully selecting the network design and training schemes tailored for UDA and proposed EDAPS, an efficient domain-adaptive panoptic segmentation network that surpasses prior art by a large margin. Furthermore, we provided a detailed analysis of the various design choices for enhancing the panoptic UDA performance. When compared with previous methods on UDA panoptic segmentation, EDAPS shows significant performance gains of 20% on SYNTHIA-to-Cityscapes and even 72% on the more challenging SYNTHIA-to-Mapillary Vistas. We believe that EDAPS as a network architecture will facilitate benchmarking future UDA strategies in panoptic segmentation, making domain-adaptive panoptic segmentation more usable in practice.

References

[1] Nikita Araslanov and Stefan Roth. Self-supervised augmentation consistency for adapting semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15384–15394, 2021. 2, 3, 4

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 3

[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2017. 5

[4] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3339–3348, 2018. 1

[5] Yuhua Chen, Wen Li, and Luc Van Gool. Road: Reality oriented adaptation for semantic segmentation of urban scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7892–7901, 2018. 2

[6] Yuhua Chen, Haoran Wang, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Scale-aware domain adaptive faster r-cnn. *International Journal of Computer Vision*, 129(7):2223–2243, 2021. 1

[7] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12475–12485, 2020. 1, 3, 5

[8] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021. 3

[9] Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-ensembling with gan-based data augmentation for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6830–6840, 2019. 2, 3

[10] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1251–1258, 2017. 5

[11] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016. Dataset URL: https://www.cityscapes-dataset.com/, Dataset License: https://www.cityscapes-dataset.com/license/. 6

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*, 2021. 5

[13] Liang Du, Jingang Tan, Hongye Yang, Jianfeng Feng, Xiangyang Xue, Qibao Zheng, Xiaoqing Ye, and Xiaolin Zhang. Ssf-dan: Separated semantic feature based domain adaptation network for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 982–991, 2019. 2

[14] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016. 1

[15] Liang-Chieh Chen Yukun Zhu George, Papandreou Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *Proceedings of the European Conference on Computer Vision*, 2018. 5

[16] Rui Gong, Wen Li, Yuhua Chen, Dengxin Dai, and Luc Van Gool. Dlow: Domain flow and applications. *International Journal of Computer Vision*, 129(10):2865–2888, 2021. 2

[17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2961–2969, 2017. 3, 5

[18] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International Conference on Machine Learning*, pages 1989–1998. PMLR, 2018. 1, 2

[19] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*, 2016. 1, 2

[20] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. 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*, 2022. 1, 2, 3, 4, 5, 6

[21] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. HRDA: Context-aware high-resolution domain-adaptive semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 1, 2, 3

[22] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Domain adaptive and generalizable network architectures and training strategies for semantic image segmentation. *arXiv preprint arXiv:2304.13615*, 2023. 1, 3

[23] Lukas Hoyer, Dengxin Dai, Haoran Wang, and Luc Van Gool. MIC: Masked image consistency for context-enhanced domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 3

[24] Lukas Hoyer, Dengxin Dai, Qin Wang, Yuhua Chen, and Luc Van Gool. Improving semi-supervised and domain-adaptive semantic segmentation with self-supervised depth estimation. *International Journal of Computer Vision*, 2021. 4, 5

[25] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Cross-view regularization for domain adaptive panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10133–10144, 2021. 1, 2, 3, 4, 6, 7

[26] Christoph Kamann and Carsten Rother. Benchmarking the robustness of semantic segmentation models with respect to common corruptions. *International Journal of Computer Vision*, 129(2):462–483, 2021. 5

[27] Guoliang Kang, Yunchao Wei, Yi Yang, Yueting Zhuang, and Alexander Hauptmann. Pixel-level cycle association: A new perspective for domain adaptive semantic segmentation. *Advances in Neural Information Processing Systems*, 33:3569–3580, 2020. 2

[28] Myeongjin Kim and Hyeran Byun. Learning texture invariant representation for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12975–12984, 2020. 2

[29] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6399–6408, 2019. 1

[30] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9404–9413, 2019. 1, 6

[31] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Int. Conf. Mach. Learn. Worksh.*, 2013. 4

[32] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6936–6945, 2019. 1, 2

[33] Yu-Jhe Li, Xiaoliang Dai, Chih-Yao Ma, Yen-Cheng Liu, Kan Chen, Bichen Wu, Zijian He, Kris Kitani, and Peter Vajda. Cross-domain adaptive teacher for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7581–7590, 2022. 1

[34] Zhiqi Li, Wenhai Wang, Enze Xie, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, Ping Luo, and Tong Lu. Panoptic segformer: Delving deeper into panoptic segmentation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1280–1289, 2022. 3

[35] Dongnan Liu, Donghao Zhang, Yang Song, Fan Zhang, Lauren O'Donnell, Heng Huang, Mei Chen, and Weidong Cai. Pdam: A panoptic-level feature alignment framework for unsupervised domain adaptive instance segmentation in microscopy images. *IEEE Transactions on Medical Imaging*, 40(1):154–165, 2021. 3

[36] Yahao Liu, Jinhong Deng, Xinchen Gao, Wen Li, and Lixin Duan. Bapa-net: Boundary adaptation and prototype alignment for cross-domain semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8801–8811, 2021. 4

[37] Yunan Liu, Shanshan Zhang, Yang Li, and Jian Yang. Learning to adapt via latent domains for adaptive semantic segmentation. *Advances in Neural Information Processing Systems*, 34, 2021. 2

[38] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*, pages 97–105. PMLR, 2015. 1

[39] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *Neural Information Processing Systems*, 31, 2018. 1

[40] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. 6

[41] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. 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*, pages 2507–2516, 2019. 2

[42] Ke Mei, Chuang Zhu, Jiaqi Zou, and Shanghang Zhang. Instance adaptive self-training for unsupervised domain adaptation. In *Proceedings of the European Conference on Computer Vision*, pages 415–430. Springer, 2020. 4

[43] Luke Melas-Kyriazi and Arjun K Manrai. Pixmatch: Unsupervised domain adaptation via pixelwise consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12435–12445, 2021. 2, 3

[44] Rohit Mohan and Abhinav Valada. Efficientps: Efficient panoptic segmentation. *International Journal of Computer Vision*, 129(5):1551–1579, 2021. 1

[45] Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing Properties of Vision Transformers. In *Neural Information Processing Systems*, 2021. 5

[46] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE international conference on computer vision*, pages 4990–4999, 2017. 6

[47] Viktor Olsson, Wilhelm Tranheden, Juliano Pinto, and Lennart Svensson. Classmix: Segmentation-based data augmentation for semi-supervised learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1369–1378, 2021. 4

[48] Yingwei Pan, Ting Yao, Yehao Li, Yu Wang, Chong-Wah Ngo, and Tao Mei. Transferrable prototypical networks for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2239–2247, 2019. 1

[49] Lorenzo Porzi, Samuel Rota Bulo, Aleksander Colovic, and Peter Kontschieder. Seamless scene segmentation. In *Pro-*

ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8277–8286, 2019. 1

[50] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *European Conference on Computer Vision*, pages 102–118. Springer, 2016. 1

[51] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3234–3243, 2016. Dataset URL: http://synthia-dataset.net/, Dataset License: CC BY-NC-SA 3.0. 1, 6

[52] Suman Saha, Anton Obukhov, Danda Pani Paudel, Menelaos Kanakis, Yuhua Chen, Stamatios Georgoulis, and Luc Van Gool. Learning to relate depth and semantics for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8197–8207, 2021. 1, 2

[53] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6956–6965, 2019. 1

[54] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3723–3732, 2018. 1

[55] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Advances in Neural Information Processing Systems*, 29, 2016. 4

[56] Christos Sakaridis, Dengxin Dai, Simon Hecker, and Luc Van Gool. Model adaptation with synthetic and real data for semantic dense foggy scene understanding. In *Proceedings of the European Conference on Computer Vision*, pages 687–704, 2018. 2

[57] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems*, 33:596–608, 2020. 2, 4

[58] Tao Sun, Cheng Lu, Tianshuo Zhang, and Haibin Ling. Safe self-refinement for transformer-based domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7191–7200, 2022. 5

[59] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in Neural Information Processing Systems*, 30, 2017. 2, 4

[60] Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. Dacs: Domain adaptation via cross-domain mixed sampling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1379–1389, 2021. 1, 2, 3, 4, 6

[61] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7472–7481, 2018. 1, 2, 4

[62] Yi-Hsuan Tsai, Kihyuk Sohn, Samuel Schulter, and Manmohan Chandraker. Domain adaptation for structured output via discriminative patch representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1456–1465, 2019. 2, 4

[63] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017. 2

[64] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2517–2526, 2019. 2, 6, 7

[65] Haoran Wang, Tong Shen, Wei Zhang, Ling-Yu Duan, and Tao Mei. Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 642–659. Springer, 2020. 4

[66] Qin Wang, Dengxin Dai, Lukas Hoyer, Luc Van Gool, and Olga Fink. Domain adaptive semantic segmentation with self-supervised depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8515–8525, 2021. 2

[67] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021. 5

[68] Zhonghao Wang, Mo Yu, Yunchao Wei, Rogerio Feris, Jinjun Xiong, Wen-mei Hwu, Thomas S Huang, and Honghui Shi. Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12635–12644, 2020. 2

[69] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 5

[70] Yuwen Xiong, Renjie Liao, Hengshuang Zhao, Rui Hu, Min Bai, Ersin Yumer, and Raquel Urtasun. Upsnet: A unified panoptic segmentation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8818–8826, 2019. 1

[71] Minghao Xu, Hang Wang, Bingbing Ni, Qi Tian, and Wenjun Zhang. Cross-domain detection via graph-induced prototype alignment. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition, pages 12355–12364, 2020. 1, 4

[72] Tongkun Xu, Weihua Chen, Pichao Wang, Fan Wang, Hao Li, and Rong Jin. Cdtrans: Cross-domain transformer for unsupervised domain adaptation. *arXiv preprint arXiv:2109.06165*, 2021. 5

[73] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4085–4095, 2020. 2, 6, 7

[74] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in Neural Information Processing Systems*, 33:5824–5836, 2020. 2

[75] Jingyi Zhang, Jiaxing Huang, and Shijian Lu. Hierarchical mask calibration for unified domain adaptive panoptic segmentation. *arXiv preprint arXiv:2206.15083*, 2022. 1, 2, 3, 4, 5, 7

[76] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. 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*, pages 12414–12424, 2021. 2, 4

[77] Qiming Zhang, Jing Zhang, Wei Liu, and Dacheng Tao. Category anchor-guided unsupervised domain adaptation for semantic segmentation. *Advances in neural information processing systems*, 32, 2019. 2, 4

[78] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6881–6890, 2021. 5

[79] Qianyu Zhou, Zhengyang Feng, Qiqi Gu, Jiangmiao Pang, Guangliang Cheng, Xuequan Lu, Jianping Shi, and Lizhuang Ma. Context-aware mixup for domain adaptive semantic segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022. 2, 3

[80] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European Conference on Computer Vision*, pages 289–305, 2018. 2, 4

[81] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5982–5991, 2019. 2, 6, 7