MICDrop: Masking Image and Depth Features via Complementary Dropout for Domain-Adaptive Semantic Segmentation

Linyan Yang[1,3], Lukas Hoyer[2], Mark Weber[1,3], Tobias Fischer[2], Dengxin Dai[2], Laura Leal-Taixé[4], Marc Pollefeys[2,5], Daniel Cremers[1,3], and Luc Van Gool[2]

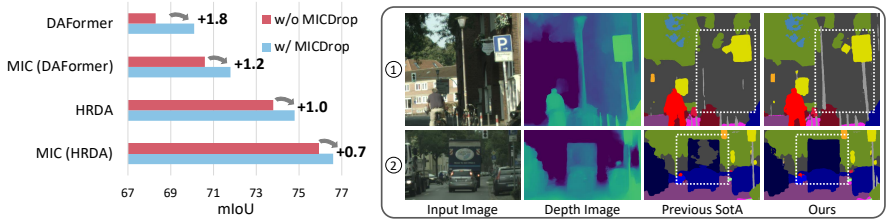
[1] TU Munich, [2] ETH Zurich, [3] Munich Center for Machine Learning, [4] NVIDIA, [5] Microsoft

Abstract. Unsupervised Domain Adaptation (UDA) is the task of bridging the domain gap between a labeled source domain, e.g., synthetic data, and an unlabeled target domain. We observe that current UDA methods show inferior results on fine structures and tend to oversegment objects with ambiguous appearance. To address these shortcomings, we propose to leverage geometric information, i.e., depth predictions, as depth discontinuities often coincide with segmentation boundaries. We show that naively incorporating depth into current UDA methods does not fully exploit the potential of this complementary information. To this end, we present MICDrop, which learns a joint feature representation by masking image encoder features while inversely masking depth encoder features. With this simple yet effective complementary masking strategy, we enforce the use of both modalities when learning the joint feature representation. To aid this process, we propose a feature fusion module to improve both global as well as local information sharing while being robust to errors in the depth predictions. We show that our method can be plugged into various recent UDA methods and consistently improve results across standard UDA benchmarks, obtaining new state-of-the-art performances. Project Page: `https://github.com/ly-muc/MICDrop`

Keywords: Domain Adaptation · Semantic Segmentation · Depth Guidance

1 Introduction

The computer vision community has seen tremendous success in recognition tasks over the years, yet the issue of efficiently sourcing large volumes of labeled images for supervised training of neural networks persists. This problem is especially pronounced in semantic segmentation, where manually creating labels is particularly labor-intensive [7, 43]. Alternatively, images can be obtained at a large scale from a simulator, which can also easily generate the corresponding



a) Quantitative improvement b) Qualitative Examples

Fig. 1: Previous UDA methods such as MIC [23] struggle with the segmentation of fine structures (top row) and oversegmentation of difficult objects (bottom row). Therefore, we propose MICDrop to improve semantic segmentation UDA with depth estimates, which can capture fine structures and are consistent within object boundaries. We apply MICDrop to four different methods on the GTA→Cityscapes benchmark and show consistent improvements.

segmentation labels. In that scenario, models are often trained on synthetic datasets and later applied to real-world data. This transition frequently results in a noticeable performance decline due to the variance in data distribution between the synthetic source and real-world target sets (*e.g.* appearance of objects), a phenomenon known as domain shift. Therefore, conventional training mostly considers datasets where the training and test data are drawn from the same distribution. However, this assumption often breaks in real-world applications under domain shifts. Recognizing these challenges, researchers have been exploring ways to minimize or eliminate the need for annotated data from the target domain. This paper focuses on Unsupervised Domain Adaptation (UDA), where a model is trained using labeled data from a synthetic source domain and unlabeled data from a real-world target domain.

Current challenges in UDA. Recent UDA methods [4, 20, 21, 23, 56] are able to significantly reduce the gap to methods trained in a fully supervised fashion on the target domain. However, state-of-the-art methods struggle with two main aspects shown in Fig. 1: (1) Despite using high-resolution strategies such as HRDA [21], they still face problems with fine structures and high-frequency details. (2) UDA methods are prone to oversegmentation when visual appearance clues are ambiguous. These issues motivate us to look into scene representations that are more robust to appearance changes and provide precise boundaries to strengthen the existing UDA models.

Complementary representation. An appearance-based image representation is essential to our task, however, a *geometric representation* could provide complementary cues when it comes to segmentation. In particular, the correlation of depth and segmentation boundaries can help to address challenges (1) and (2), as shown in Fig. 1. First, a pole might blend with a building behind it in color, but its depth profile is distinct, simplifying its segmentation. Second, the back of the truck might have visual features that could also be part of a building.

However, the depth is smooth within the boundaries of the truck, suggesting that the semantic class should be consistent. While actively measured depth might not be available, advances in image-based depth estimation [10, 60] enable us to explore the task in a general setting. Previous works [13, 27, 53] in UDA have focused on improving the learning process via an auxiliary depth prediction task. In such multi-task learning settings, a network is trained to predict both depth and semantics from RGB inputs. However, multi-task learning adds additional complexity, including balancing multiple network branches and their corresponding losses, to the already challenging UDA setting.

Contributions. We propose a more streamlined approach: Instead of engaging in multi-task learning, we treat it as a modality fusion problem. Rather than producing multiple outputs from a single input (one-to-many), we redefine and simplify the task as a many-to-one prediction problem. With semantic segmentation as our output and readily available depth estimates, we study two research questions: First, what is the most effective method to *fuse features from two modalities* in a UDA context? Second, how can we *utilize existing work* and design our method as a *plugin network* that seamlessly integrates with pretrained models, thereby eliminating the need for extensive retraining?

We integrate our findings into MICDrop, a novel framework for leveraging depth in domain-adaptive semantic segmentation. Our framework is based on a novel *cross-modality complementary dropout* technique along with a tailored masking schedule. Our masking strategy mitigates the tendency of the network to underutilize additional depth features, as is prevalent in multi-modal learning, and becomes more pronounced with pretrained networks. In particular, we foster cross-modal feature learning by strategically corrupting both RGB and depth features in a complementary manner, enforcing the utilization of the different modalities to fill in masked information. To integrate information from both modalities effectively, we also propose a *cross-modality feature fusion* module. It is designed to integrate global and local cues from one modality to the other. First, it computes depth feature similarities to aggregate RGB features based on the resulting attention map, aiding the RGB feature aggregation with global depth cues. This is particularly beneficial for segmenting objects that the RGB encoder struggles to represent accurately but have a smooth depth profile. Second, it applies local self-attention to depth features, leveraging the discontinuity in local depth for describing boundaries, a critical factor in identifying thin structures. This approach yields significant improvements over various recent UDA methods on two standard benchmarks while only requiring the training of a light-weight plugin network for a low number of iterations. Thus, MICDrop (w/ DAFormer) can be trained within 11 hours on a single GTX Titan X GPU (12 GB).

In summary, our key contributions are:

- A **complementary feature masking strategy** for depth and RGB, fostering cross-modal feature learning.
- A **cross-modality fusion module** to improve segmentation based on depth by using global and local cues.

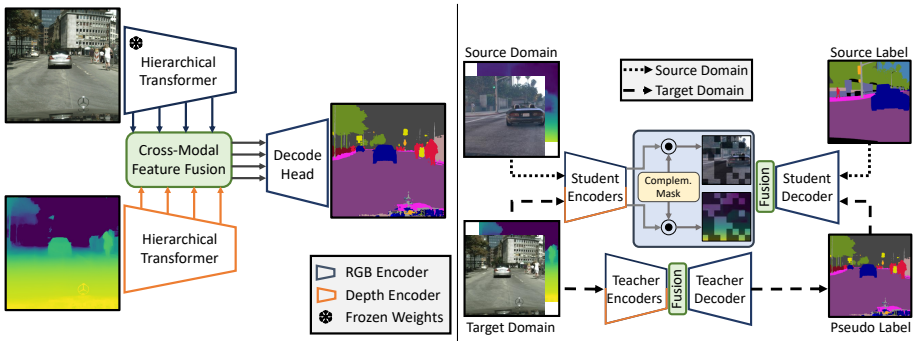
– Comprehensive ablations demonstrating MICDrop's efficacy, with **improvements ranging from 0.7 to 1.8 mIoU** across four recent UDA methods on the GTA→Cityscapes benchmark.

By showing that complementary geometric information even improves modern high-resolution, Transformer-based UDA methods, we hope to lay the foundation for future research exploring the merits of auxiliary modalities for semantic segmentation UDA.

2 Related Work

Unsupervised Domain Adaptation (UDA). In UDA, methods have access to labeled source and unlabeled target data at training time and can mostly be categorized into two primary groups. The first one utilizes a Generative Adversarial Network (GAN) [11] to align input images [17], image features [18], or output features [42,48,51] across domains. The second stream of works are built on self-training [12,28]. Here, pseudo labels are created using a teacher network [1,45]. These labels can be further refined using confidence thresholds [35, 65, 68] and pseudo label prototypes [36, 63, 64]. The student model then receives an image version with cross-domain class mix [46, 67] and color augmentations [1]. The self-training can further strengthened by domain-robust Transformers [20, 40, 61], class-balanced sampling [20, 68], multi-resolution adaptation [21], or contrastive learning [4, 56]. Our proposed MICDrop builds on the self-training paradigm.

Depth in Semantic Segmentation. Several works in semantic segmentation have shown the merits of leveraging geometric cues. In one branch of work [19, 29, 41, 49, 52, 53], depth estimation is only used as an auxiliary task. Different from that, some methods [32, 59, 66] explore multi-task learning from RGB input, in which depth is another output target. Similar multi-task studies [13, 24, 27, 53] have also been made in the context of UDA. In both cases, this requires a bi-directional feature exchange across modalities. Our method, however, is more closely related to RGB-D semantic segmentation [5, 26], in which both RGB and depth are input, while semantic segmentation is the only output, and hence we focus on a *uni-directional* feature refinement from depth to RGB. In our case, depth also serves the purpose of reducing the domain gap further. While some methods [26, 44, 59] uses variants and extensions of the Squeeze-and-Excitation Block [25] for cross-modal feature fusion, more recent methods [5,32,54,62] propose softmax attention-based aggregation. Inspired by their success, we propose a combination of an excitation block for local windows and a cross-attention block for global reasoning. Crucially, we use a *depth-guided attention map*, thereby enhancing uni-directional guidance using geometric data. In contrast to previous RGB-D works such as [5, 31, 62], we leverage both local and global dependencies for domain-robust depth-to-segmentation refinement. We show in Tab. 3b that leveraging geometric cues is not trivial in the context of UDA and conventional cross-attention fails here.



a) Architecture b) Training Scheme

Fig. 2: Method overview. Our proposed architecture is visualized on the left side. We use a light-weight hierarchical depth encoder and process the features in our proposed cross-modal feature fusion module. On the right side, we illustrate our training pipeline, in which source and target images are fed through the student encoders. Then, our proposed cross-modality complementary dropout is applied to the corresponding features on each feature resolution. Finally, we feed them through our fusion block, followed by the decoder, to make a final prediction.

Masked Image Modeling (MIM). MIM is a powerful method for self-supervised pretraining. In this approach, information is withheld in order to train the network to recover certain targets. Such reconstruction targets can range from RGB inputs [14, 58], to HOG features [55], to visual tokens [3, 8]. MultiMAE [2] shows the benefits of using masking of input patches in *supervised* multi-task learning by using a shared encoder and modality-specific decoder. In contrast to their work, we propose complementary masking in a UDA setup on a *multi-resolution feature level* in separate encoders (instead of input masking), enabling the use of pretrained RGB encoders. MIC [23] applies MIM to UDA to improve context reasoning. Different from MIC, we propose a novel complementary multi-modal feature dropout to facilitate cross-modality learning instead of only masking RGB inputs for context enhancement. In Sec. 4.2, we show that complementary feature dropout is orthogonal and further boosts networks trained with MIC.

3 Method

Overview. In Fig. 2, we present our method, featuring two novel modules that can be plugged into various UDA methods to leverage geometric cues. Our feature fusion module (Sec. 3.1) integrates auxiliary inputs, *e.g.*, depth, into RGB features. It fuses global and local information via attention-based aggregation. Our masking module (Sec. 3.2) ensures balanced input use, avoiding pure reliance on a single input modality such as RGB or depth. We outline UDA training essentials before diving into the details of the multi-modal feature fusion module and the masking strategy.

Problem Definition. We tackle the problem of unsupervised domain adaptation, in which we have access to labeled source data ($\mathbf{X}_s$, $\mathbf{Y}_s$) and unlabeled target data ($\mathbf{X}_t$) to train a neural network. The goal is to bridge the domain gap between $\mathbf{X}_s$ and $\mathbf{X}_t$. The performance is measured on a labeled hold-out validation set of the target domain. In this work, we focus on RGB images and depth images as input and semantic segmentation as output.

Preliminaries. Training a network on a source domain typically follows standard supervised methods. However, overcoming the domain gap with the target domain requires leveraging the unlabeled target data. Recent approaches, such as those in [1, 20, 21, 23, 46, 47, 56], adopt a student-teacher framework. In this framework, the teacher network is updated each training iteration as an exponential-moving average (EMA) [45] of the student network. This EMA teacher generates pseudo labels on the target images, which in turn act as a supervisory signal ($\mathbf{X}_t$, $\tilde{\mathbf{Y}}_t$) to the student. We follow standard practice [46] and present the student with a heavily augmented view while presenting the teacher with a weakly augmented view of an image. Additionally, we note that most methods use hierarchical encoders to produce multi-resolution feature maps, enhancing fine-grained segmentation. We study the effectiveness of our proposed method by extending existing pretrained hierarchical encoders [20, 21] to leverage depth. The depth estimates are obtained from RGB images. If stereo image pairs are available, we utilize UniMatch [60]. If not, we use the monocular method MonoDepth2 [10].

3.1 Multi-Modal Feature Fusion

First, we study the fusion of features from different modalities. Our goal is to have a *light-weight* training pipeline, which can make use of already *existing work* in UDA. For that purpose, we construct a multi-modality encoder that contains two individual encoders, one for RGB features and one for depth features. The depth features come from a newly trained, light-weight depth encoder, while the RGB features come from a pretrained RGB feature encoder. As state-of-the-art encoders typically output multi-scale feature levels, we perform feature fusion separately on each level. Different from multi-task learning, in which features from different modalities are *all* refined, our goal is *solely* to improve semantic segmentation. Thus, we focus on a unidirectional refinement, *i.e.*, the depth features are used to enrich the RGB features but not vice-versa. As can be seen in Fig. 3, we divide our feature fusion block into (1) *global* depth-guided cross-attention, (2) *local* self-attention and (3) final *residual fusion*.

Global Depth-Guided Cross-Attention. Intuitively, similarities in depth features can provide a strong cue towards the same semantic class. For example, large objects like bus or train exhibit similar gradual changes within their object, while thin structures such as pole or sign typically exhibit rapid depth changes relative to their surroundings. Such additional cues could serve as a *correctional and complementary signal* to the RGB features when predicting the semantic class. Thus, the purpose of this branch is to aggregate RGB features globally based on

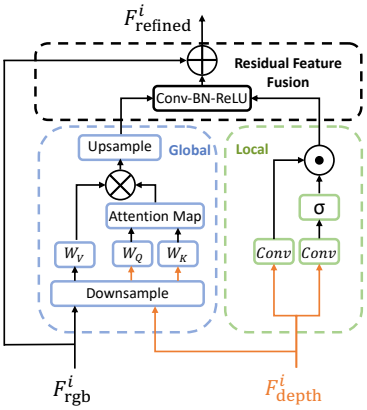


Fig. 3: Feature fusion of RGB and depth. The presented method comprises two key components: a global and a local attention module. The local attention module refines information coming from depth within a local window by using sigmoid gates. In contrast to that, the global attention module aggregates image features based on similarity in their corresponding depth features, and thus providing more global context. Finally, the residual feature fusion block fuses all features.

their corresponding depth feature similarity. For such a global aggregation across different tasks, one natural choice would be cross-attention.

However, directly using (global) attention usually exhibits problematic scaling behavior. Given an input $x \in \mathbb{R}^{H\times W\times C}$, the standard attention [50] has a complexity of $\mathcal{O}((HW)^2C + HWC^2)$ making it computationally infeasible for our case. We, therefore, bilinearly downsample high-resolution feature maps to reduce the spatial dimensions before applying cross-attention. During training, we sample feature maps with a pooling factor of {4, 2, 1, 1} for low- and high-level features, respectively. Conversely, during inference and pseudo-label generation, this pooling is adjusted to {2, 1, 1, 1} and thus only applied to low-level features.

Given potentially downscaled depth features $\mathbf{F}^i_{\text{depth}}$ at level i, we obtain depth-based queries $\mathbf{Q}_{\text{depth}}$ and keys $\mathbf{K}_{\text{depth}}$ by using projection weights $\mathbf{W}^i_q$ and $\mathbf{W}^i_k$. The corresponding RGB features $\mathbf{F}^i_{\text{rgb}}$ are downscaled in similar fashion and serve as values $\mathbf{V}_{\text{rgb}}$ after being projected by $\mathbf{W}^i_k$. Formally, the cross-attention for the aggregation is:

$$\mathbf{F}^i_{\text{global}} = \text{softmax}\left(\frac{\mathbf{Q}_{\text{depth}}\mathbf{K}^T_{\text{depth}},}{\sqrt{d_k}}\right)\mathbf{V}_{\text{rgb}} \tag{1}$$

Local Self-Attention. Modeling global interactions on a downsized resolution might not be enough to capture the fine details of objects like sign or pole. Keeping the same computational complexity problem for the self-attention of depth features in mind, we argue that many important interactions for segmentation happen within a *local* window. Specifically, *depth discontinuities* provide strong cues for *boundary regions* among semantic classes, while *smooth and continuous depth* indicate *no change* in semantics. Thus, we hypothesize that restricting the self-attention to a local window would still capture important *complementary signals* to the global information.

To model such dynamics, we draw inspiration from earlier work [16, 59], in which sigmoid gates were used successfully to control the local flow of information

without adding a large computational overhead. In particular, we use two 3×3 convolutions. The first convolution outputs features into a sigmoid function σ to obtain a *local* attention map. We note that *no pooling* has been used, enabling the network to model a precise local control flow. The second convolution is used to refine the depth features, which are fed into a pointwise multiplication together with the local attention map. As this branch is used to model *complementary* features, we exclusively use depth features. Formally, we compute the local self-attention as:

$$\mathbf{F}^i_{\text{local}} = \sigma\left(\text{Conv}_{3\times3}\left(\mathbf{F}^i_{\text{depth}}\right)\right) \odot \text{Conv}_{3\times3}\left(\mathbf{F}^i_{\text{depth}}\right) \tag{2}$$

Residual Feature Fusion. After aggregating global and local features, we propose a simple two-step feature fusion block to fuse all aggregated features as well as the original RGB features. At first, the depth-guided global features $\mathbf{F}^i_{\text{global}}$ and local features $\mathbf{F}^i_{\text{local}}$ are concatenated ($||$) and fused through a 1×1-Conv-BN-ReLU block. After that, the original RGB features $\mathbf{F}_{\text{rgb}}$ are added, resulting in the refined features:

$$\mathbf{F}^i_{\text{refined}} = \mathbf{F}^i_{\text{rgb}} + \text{ReLU}(\text{BN}(\text{Conv}(\mathbf{F}^i_{\text{global}}||\mathbf{F}^i_{\text{local}}))) \tag{3}$$

The refined features are then fed to a DAFormer head [20] for the final predictions.

3.2 Complementary Feature Masking

During initial experiments, we observe that simply providing estimated depth and RGB images to the network does not enable the network to leverage the full potential of all provided information. We refer the reader to Sec. 4.2 for details of that analysis. We hypothesize that the network grows too confident in the RGB encoder and thus dismisses complementary information from the depth encoder, which limits adaptability to the target domain. To improve cross-modal information exchange, we therefore introduce a *cross-modal masking strategy*. In contrast to Masked Image Modeling in UDA [23], our method involves masking the learned representation of *different modalities* on *feature-level* rather than a single modality on input-level. Moreover, our masking strategy and schedule are specifically designed to improve redundancy across modalities and prevent getting stuck in local minima due to one encoder already being pretrained.

Complementary Dropout. For this, we propose using *blockwise dropout* [9] to generate masked features. When masking only individual pixels, the information could be easily restored from the neighborhood in the same modality without requiring the other modality. When masking larger blocks, the network has to understand the semantics of the other modality to recover the missing information. We, therefore, opt to mask out whole blocks within the feature map according to a predefined schedule. Furthermore, we hypothesize that learning complementary features across modalities can be achieved best by masking the feature maps of different modalities in *complementary* fashion. For example, if we mask 70% of the RGB features, we would mask the remaining 30% in the depth features. This

intuitively corrupts the information across modalities and forces the network to rely on features from all modalities. Fig. 2 illustrates the idea of complementary dropout. Formally, we define the complementary masking as:

$$\mathbf{M}_{\text{rgb}}(u, v) = [\gamma > m_r^t], \quad \gamma \sim \text{Uniform}(0, 1) \tag{4}$$

$$\mathbf{M}_{\text{depth}}(u, v) = 1 - \mathbf{M}_{\text{rgb}}(u, v) \tag{5}$$

where m_r^t denotes the masking ratio at iteration t and (u, v) the block index of the i-th feature map. To fulfill our goal of true cross-modal complementary masking, we use the same masking across all feature map levels and experimentally validate that design in Sec. 4.2. Conceptually, this avoids the recovery of features within the feature pyramid of the same modality. Therefore, our method is designed to foster the *transfer of complementary information* and to promote the *learning of potentially redundant information*, which in turn increases robustness and reduces sensitivity to domain-specific appearance changes.

Masking Schedule. Prior studies [9, 30] have highlighted the limitations of a static masking ratio. In response, we adopt a dynamic masking ratio schedule for RGB and depth features. This approach is particularly effective when using a pretrained encoder for one modality and an untrained encoder for the other, as it compensates for the initial disparity in feature quality. At the beginning of the training, we keep a high proportion of depth features to accelerate the training of the depth encoder and improve its feature quality. As training progresses, the schedule is adjusted to gradually reduce depth feature retention, thereby increasing the reliance on the RGB encoder. We note that this masking is only applied during training but not during inference. This method not only promotes an efficient exchange of information between modalities but also capitalizes on the depth data to bolster semantic learning in the early stages of training.

4 Experiments

Datasets. We perform our experimental evaluation on two widely used UDA benchmarks. The first one uses synthetic source data from the GTA [38] dataset, which contains 24,966 images with a resolution of 1914×1052. The second benchmark uses SYNTHIA [39], which consists of 9,400 synthetic images with a resolution of 1280×760. In both cases, the target dataset is Cityscapes [7], which includes 500 validation images, each having a resolution of 2048×1024.

Depth Estimates. We obtain depth estimations for the source domain from MonoDepth2 [10] via self-supervised monocular depth estimation trained on image sequences from VIPER(GTA) [37] and SYNTHIA-SEQ. For Cityscapes, we obtain disparity estimations from UniMatch [60] using stereo images trained on a large synthetic dataset [34].

Metrics. Following previous studies, we report the mean Intersection over Union (mIoU) in % over the 19 common categories shared by GTA and Cityscapes and the 16 common categories shared by SYNTHIA and Cityscapes.

Network Architecture. We use DAFormer [20] as our baseline model for ablation studies, as it achieves a strong performance at high training and inference speed. As depth feature extractor, we use the light-weight MiT-B3 [57]. To demonstrate the *plugin* capability of our method, we additionally apply MICDrop to the state-of-the-art methods HRDA [21] and MIC [23].

Training details. We use an AdamW [33] optimizer with a learning rate of 6×10^{-5} for the depth encoder and 6×10^{-4} for the decode head and feature fusion module. To address the limited scale of SYNTHIA, we align the learning rate for all modules to the depth encoder. As learning rate schedule, we use linear warm-up in the first 1.5k iterations and polynomial decay with factor 0.9 afterward. The EMA [45] teacher is updated with a momentum of α=0.999 at each step. Following prior works [20, 23, 46], the batch size is set to 2, with data augmentations such as color jitter, Gaussian blur, and cross-domain class mixing. We initialize the RGB encoder and decode head with the publicly available pretrained weights [20, 21, 23] for our experiments. The depth encoder is initialized with ImageNet weights. We keep the RGB encoder frozen and train the rest of the network 20k iterations on both GTA [38] and SYNTHIA [39]. We use a cross-entropy loss for both source and target images. We additionally apply a forward pass without masking to reduce the feature distribution shift between training and (unmasked) inference time. MICDrop can be trained in 11 hours using a single GTX Titan X GPU (12 GB) with DAFormer and in 17 hours on two Titan GPUs with HRDA.

4.1 Main Results

To validate the effectiveness of MICDrop and its capabilities as a *plugin*, we evaluate its performance across the three state-of-the-art methods DAFormer [20], HRDA [21], and MIC [23]. The results are shown in Tab. 1.

Starting with applying MICDrop to DAFormer [20] on GTA, the results improve by 1.8 mIoU. Using the recent MIC pretrained model, we obtain improvements by 1.2 mIoU. Remarkably, our method still improves over the strong HRDA method by 1.0 mIoU. When we build on top of the currently best performing model MIC_{HRDA}, we can further boost results by 0.7 mIoU, setting a new state of the art in UDA semantic segmentation. Considering that the improvement is on top of the best-performing SOTA approach on a saturating benchmark (94% of the oracle performance), this gain can be considered significant. By plugging our light-weight modules into each of these architectures and adding complementary dropout, we achieve consistent improvements, clearly showing that leveraging depth helps in closing the domain gap. Furthermore, the comparison to MIC supports our hypothesis that our contributions are orthogonal to the successes of input masking (MIM) in UDA.

Diving into the details of these improvements, we notice predominantly gains in two types of objects. First, we see consistent improvements in classes of thin structures such as poles, signs, or motorbikes. This is enabled by our design of aggregating local depth features without using any pooling, as these local depth

Method	*mIoU*	*Road*	*S.walk*	*Build.*	*Wall*	*Fence*	*Pole*	*Tr.Light*	*Sign*	*Vege.*	*Terrain*	*Sky*	*Person*	*Rider*	*Car*	*Truck*	*Bus*	*Train*	*M.bike*	*Bike*
Synthetic-to-Real: GTA→Cityscapes (Val.)																				
AdaptSeg [48]	41.4	86.5	25.9	79.8	22.1	20.0	23.6	33.1	21.8	81.8	25.9	75.9	57.3	26.2	76.3	29.8	32.1	7.2	29.5	32.5
ADVENT [51]	45.5	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
DACS [46]	52.1	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
CorDA [53]	56.6	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
ProDA [63]	57.5	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
DAFormer [20]	54.2	85.7	66.8	81.5	27.3	20.4	46.4	53.2	63.0	84.5	32.1	72.9	71.9	45.0	90.5	60.7	58.8	0.1	23.2	46.4
+ *MICDrop*	58.3	95.2	69.1	88.1	26.0	27.7	48.8	55.2	63.6	89.6	49.5	90.3	72.0	45.4	91.4	63.3	61.1	0.0	23.8	46.7
AdaptSeg† [48]	47.8	85.2	20.4	85.5	38.2	30.9	34.5	43.0	26.2	87.4	40.3	86.4	63.6	23.7	88.6	48.5	50.6	5.8	33.1	16.2
DACS† [46]	58.2	88.9	50.0	88.4	46.4	43.9	43.1	53.4	54.8	89.9	51.2	92.8	64.2	9.4	91.4	77.3	63.3	0.0	47.4	49.8
DAFormer [20]	68.3	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
+ *MICDrop*	70.1	96.0	71.8	90.3	53.3	46.4	54.8	57.8	66.7	90.0	49.2	92.2	73.6	46.3	92.8	78.1	80.6	70.7	57.5	63.2
MIC$_{DAFormer}$ [23]	70.6	96.7	75.0	90.0	58.2	50.4	51.1	56.7	62.1	90.2	51.3	92.9	72.4	47.1	92.8	78.9	83.4	75.6	54.2	62.6
+ *MICDrop*	71.8	96.5	74.2	90.8	60.5	52.0	55.8	59.9	65.6	90.3	51.8	93.0	73.1	46.9	93.4	82.0	85.8	74.3	56.6	62.8
HRDA [21]	73.8	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
+ *MICDrop*	74.8	95.8	71.1	91.5	**62.8**	55.0	*60.8*	*64.0*	*73.4*	91.3	49.1	94.0	79.2	54.6	94.4	84.8	88.5	79.0	**65.9**	65.5
MIC$_{HRDA}$ [23]	*75.9*	*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*	**95.6**	*85.4*	*90.3*	*80.4*	64.5	**68.5**
+ *MICDrop*	**76.6**	**97.6**	**81.5**	**92.0**	**62.8**	**59.4**	**62.6**	62.9	**73.6**	*91.6*	**52.6**	94.1	**80.2**	**57.0**	*94.8*	**87.4**	**90.7**	**81.6**	*65.3*	*67.8*
Synthetic-to-Real: Synthia→Cityscapes (Val.)																				
ADVENT [51]	41.2	85.6	42.2	79.7	8.7	0.4	25.9	5.4	8.1	80.4	–	84.1	57.9	23.8	73.3	–	36.4	–	14.2	33.0
DACS [46]	48.3	80.6	25.1	81.9	21.5	2.9	37.2	22.7	24.0	83.7	–	90.8	67.6	38.3	82.9	–	38.9	–	28.5	47.6
CorDA [53]	55.0	**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
ProDA [63]	55.5	*87.8*	45.7	84.6	37.1	0.6	44.0	54.6	37.0	*88.1*	–	84.4	74.2	24.3	88.2	–	51.1	–	40.5	45.6
DACS† [46]	52.2	58.0	46.0	84.8	37.7	5.2	38.6	20.9	47.3	85.9	–	81.6	73.0	43.9	86.9	–	55.6	–	51.1	18.6
DAFormer [20]	61.3	82.2	37.2	88.6	42.9	8.5	50.1	55.1	54.5	85.7	–	88.0	73.6	48.6	87.6	–	62.8	–	53.1	62.4
+ *MICDrop*	62.4	81.0	37.1	89.4	45.7	*9.5*	51.8	57.3	58.0	86.7	–	85.0	73.6	50.4	88.2	–	64.7	–	56.8	62.8
HRDA [21]	65.8	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
+ *MICDrop*	66.8	86.3	49.6	*89.3*	**53.7**	5.1	57.6	*66.4*	*63.8*	86.1	–	94.1	79.1	56.0	87.8	–	*65.0*	–	64.2	*65.0*
MIC$_{HRDA}$ [23]	*67.3*	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
+ *MICDrop*	**67.9**	82.8	42.6	**90.5**	*51.6*	**9.6**	**61.0**	65.7	**65.0**	**89.1**	–	**95.0**	**81.1**	**59.7**	**90.6**	–	**68.3**	–	**67.4**	**66.5**

Table 1: Comparison of MICDrop with state-of-the-art UDA methods. The performance is reported as IoU in %. We group methods based on ResNet [15] and Segformer [57] backbones. † denotes results obtained with a Segformer backbone from [22]. On both GTA and SYNTHIA, MICDrop achieves consistent improvements, demonstrating the effectiveness of our masking strategy and fusion module.

continuities at boundary regions serve as a strong cue. Second, larger classes of lower prevalence in the dataset, such as truck, bus, or train show generally improved performance when adding MICDrop. In these cases, both global as well as local depth features can help. Due to their size, global reasoning can improve the consistency of their segmentation, but also the locally smooth, continuous depth lower the likelihood of changes in the semantics within a local window.

We also benchmark MICDrop with a ResNet-101 architecture in the DAFormer framework in Tab. 1. It shows a significant gain of 4.1 mIoU over the baseline and outperforms the previous SOTA depth-guided UDA method CorDA [53].

Tab. 1 further provides results on SYNTHIA→Cityscapes. Also here, MICDrop achieves consistent improvements over its baselines, *i.e.* 1.1 mIoU for DAFormer, 1.0 mIoU for HRDA, and 0.6 mIoU for MIC$_{HRDA}$. The improvements are slightly

Method	mIoU	Road	S.walk	Build.	Wall	Fence	Pole	Tr.Light	Sign	Vege.	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	M.bike	Bike
MIC$_{HRDA}$ [23]	52.0	41.9	59.1	54.0	36.6	31.7	58.1	53.3	56.1	64.9	34.6	66.6	63.3	44.3	72.5	49.2	61.0	49.4	41.2	50.3
+ *MICDrop*	53.6	41.1	60.2	58.2	36.9	33.8	61.0	51.9	59.9	65.3	35.0	66.3	65.6	46.6	73.6	54.1	64.1	49.3	43.6	52.3
Δ	+1.6	-0.8	+1.1	**+4.2**	+0.3	+2.1	**+2.9**	-1.4	**+3.8**	+0.4	+0.4	-0.3	+2.3	+2.3	+1.1	**+4.9**	**+3.1**	-0.1	+2.4	+2.0

Table 2: Boundary IoU on GTA→Cityscapes with a dilation factor of 0.005.

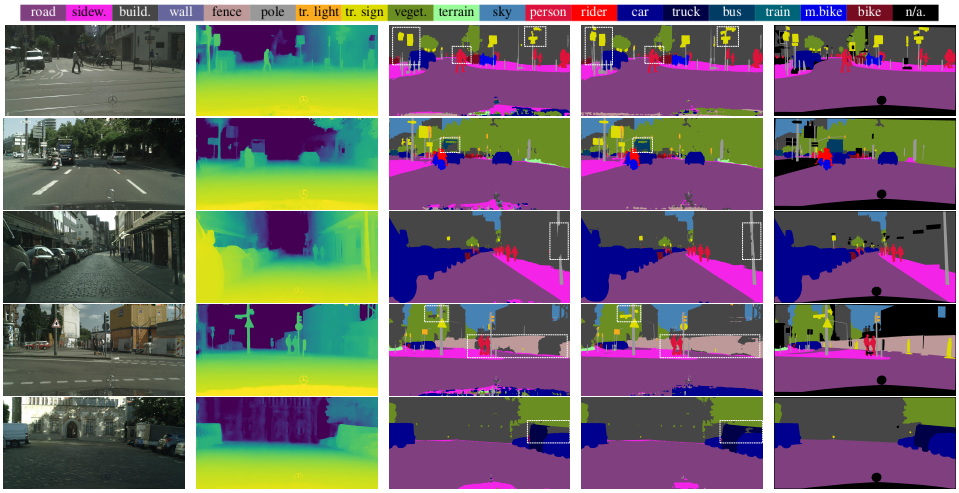


Fig. 4: Qualitative results. These results show the improvements of MICDrop in comparison to MIC (HRDA). We highlight improvements on thin structures, such as pole and traffic sign, as well as on larger objects like trucks, busses and fences. In rows 1, 3, and 4, we can see that thin structures have a distinct depth profile, which helps in predicting accurate boundaries. In rows 2, 4, and 5, we observe that the depth region for the fence, bus, and truck is smooth, improving the consistency of the predicted segmentation.

smaller than for GTA→Cityscapes, which might be caused by the smaller dataset size of SYNTHIA, resulting in overfitting issues.

Boundary Analysis. Tab. 2 additionally studies the boundary IoU [6]. Compared to the default IoU it improves by a significantly larger margin (1.6 vs 0.7), supporting our hypothesis that MICDrop particularly improves segmentation boundaries. The class-wise boundary IoUs further demonstrate that both classes with fine structures (*e.g.* pole or sign) and classes that are prone to oversegmentation (*e.g.* truck and building) are improved, quantitatively supporting our motivation in Fig 1.

Qualitative Analysis. In Fig. 4, we showcase a qualitative comparison with the current state-of-the-art model. We note that the estimated depth exhibits sharp discontinuities, providing strong cues for thin structures such as poles or

traffic signs (*cf.* row 1, 3, and 4). Moreover, these examples demonstrate how the piecewise smooth depth can help to mitigate oversegmentation of larger objects by guiding the network to predict more consistent semantic segmentation within depth contours, as can be seen for the truck, the bus, and fence (*cf.* row 2, 4, and 5). Further qualitative comparisons with other methods are provided in the supplementary material.

4.2 Ablation Studies

We start the experimental validation of our design choices by ablating our dropout strategy. After that, we compare different operations for the task of feature fusion. For a fair comparison, we also finetune the pretrained baseline model without any changes using the same hyper-parameter described before but did not observe any performance improvements (68.3 ±0.2 mIoU).

Cross-Modal Complementary Dropout. The ablation study in Tab. 3a explores the impact of various masking strategies. All experiments use our proposed feature fusion module. Adding depth information to our baseline without masking increases the mIoU from 68.3 to 69.1 on the GTA dataset, showing the promise of depth. However, we show experimentally that depth features are not fully utilized by the decoder by testing simple masking strategies first.

When masking RGB features, the network can leverage depth information marginally better by 0.2 percentage points, indicating that feature corruption, when done right, could enhance cross-modal feature integration. However, applying independent masking to both RGB and depth features simultaneously does not show improvements over no masking. As evident from a significantly higher standard deviation, strategies in which the same regions in depth and RGB might be masked make the training more unstable.

Notably, using the same *complementary masking across all levels* leads to a substantial gain: an increase of 1.0 mIoU over the baseline with depth (with or without independent masking) and 1.8 mIoU over the DAFormer baseline. Furthermore, we show that true complementary masking is essential for effective learning. For that ablation, we allow the network to recover masked features from other feature levels as we apply complementary masking independently

Masking Strategy	Masking RGB	Masking Depth	mIoU (↑)
Baseline (w/o Depth)	✗	✗	68.3 ±0.5
Baseline (w/ Depth)	✗	✗	69.1 ±0.2
Only RGB	✓	✗	69.3 ±0.1
Independent	✓	✓	69.1 ±0.6
Complementary (ours)	✓	✓	**70.1** ±0.1
- Different per Level	✓	✓	69.7 ±0.1

(a) Dropout strategy ablation.

Fusion Operation	mIoU (↑)
Baseline (no Depth)	68.3 ±0.5
Add	69.3 ±0.4
CMX [62]	68.6 ±0.3
Local Self-Attn	69.7 ±0.1
Global Cross-Attn	68.1 ±0.8
Local+Global (ours)	**70.1** ±0.1

(b) Feature Fusion ablation.

Table 3: Ablation study. We use DAFormer [20] trained on GTA as our baseline model. In (a), we study different dropout strategies. In (b), we ablate different designs to fuse RGB and depth features. Mean and std. deviation are reported over 3 seeds.

at each feature level, resulting in a 0.4 mIoU decrease. These findings support that complementary masking plays a crucial role in effectively leveraging depth information for semantic segmentation in UDA, as it achieves a great balance between geometric and visual scene information.

Feature Fusion. The fusion of depth and RGB features is the essential block in our RGB-D semantic segmentation. In Tab. 3b, we compare our proposed module with different feature fusion operations. To best utilize both modalities, we deploy our proposed complementary masking strategy across all tested feature fusion operations. We first explore one simple fusion technique, namely feature addition. The scores show that a naive feature fusion technique exhibits suboptimal performance in our context. We further examine the SOTA RGB-D method CMX [62], which fuses features at various encoder stages using cross-attention. However, CMX only obtains a marginal improvement of 0.3 mIoU in the UDA setting.

Turning our focus to the individual efficacy of our proposed global and local feature fusion blocks, we observed distinct outcomes. The local self-attention block, employed independently, outperformed our naive addition baseline, indicating its effectiveness in contextual feature integration. In contrast, the global depth-guided cross-attention block, when used alone, failed to demonstrate improvement and exhibited significant training instability, as evidenced by a high standard deviation of 0.8 in mIoU. Analogous to the results observed with CMX, we conjecture that these findings underscore the significance of controlling the flow of local information in UDA. However, it is crucial to note that these blocks were designed to *complement* each other. When combined, their synergy becomes clear, validating our hypothesis that both local and global attention mechanisms are indispensable for optimal performance. This combination led to a notable improvement of additional 0.4 mIoU over local self-attention, achieving an overall gain of 1.8 mIoU over the baseline [20]. In summary, our fusion module effectively harnesses both *global and local cues*, significantly enhancing the overall effectiveness of our RGB and depth feature fusion task.

5 Conclusion

We present a novel complementary dropout method specifically tailored for UDA. Coupled with our cross-modal fusion module that combines RGB and depth features, our approach consistently improves various recent UDA methods, achieving state-of-the-art results. In particular, on both GTA and SYNTHIA, MICDrop achieves a boost of 0.7 to 1.8 mIoU, depending on the method used for encoding RGB features. Thus, MICDrop demonstrates the effectiveness of utilizing depth in UDA without the need for retraining existing encoders, achieved by adopting a many-to-one prediction framework rather than traditional multi-task learning or auxiliary predictions. The plugin design of MICDrop is intended to facilitate ease of integration into future domain-adaptive semantic segmentation methods. We hope that our simple but effective approach inspires further research into leveraging complementary cues in UDA.

References

1. Araslanov, N., Roth, S.: Self-supervised augmentation consistency for adapting semantic segmentation. In: CVPR (2021)
2. Bachmann, R., Mizrahi, D., Atanov, A., Zamir, A.: Multimae: Multi-modal multi-task masked autoencoders. In: ECCV (2022)
3. Bao, H., Dong, L., Piao, S., Wei, F.: Beit: Bert pre-training of image transformers. In: ICLR (2021)
4. Chen, M., Zheng, Z., Yang, Y., Chua, T.S.: Pipa: Pixel-and patch-wise self-supervised learning for domain adaptative semantic segmentation. ACM Multimedia (2023)
5. Chen, X., Lin, K.Y., Wang, J., Wu, W., Qian, C., Li, H., Zeng, G.: Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation. In: ECCV (2020)
6. Cheng, B., Girshick, R., Dollár, P., Berg, A.C., Kirillov, A.: Boundary IoU: Improving object-centric image segmentation evaluation. In: CVPR (2021)
7. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR (2016)
8. Dong, X., Bao, J., Zhang, T., Chen, D., Zhang, W., Yuan, L., Chen, D., Wen, F., Yu, N., Guo, B.: Peco: Perceptual codebook for bert pre-training of vision transformers. In: AAAI (2023)
9. Ghiasi, G., Lin, T.Y., Le, Q.V.: Dropblock: A regularization method for convolutional networks. NeurIPS (2018)
10. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: ICCV (2019)
11. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. NeurIPS (2014)
12. Grandvalet, Y., Bengio, Y.: Semi-supervised learning by entropy minimization. NeurIPS **17** (2004)
13. Guizilini, V., Li, J., Ambruș, R., Gaidon, A.: Geometric unsupervised domain adaptation for semantic segmentation. In: ICCV (2021)
14. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: CVPR (2022)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
16. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. (1997)
17. Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A., Darrell, T.: CyCADA: Cycle-consistent adversarial domain adaptation. In: ICML (2018)
18. Hoffman, J., Wang, D., Yu, F., Darrell, T.: Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. arXiv preprint arXiv:1612.02649 (2016)
19. Hoyer, L., Dai, D., Chen, Y., Koring, A., Saha, S., Van Gool, L.: Three ways to improve semantic segmentation with self-supervised depth estimation. In: CVPR. pp. 11130–11140 (2021)
20. Hoyer, L., Dai, D., Van Gool, L.: Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In: CVPR (2022)
21. Hoyer, L., Dai, D., Van Gool, L.: Hrda: Context-aware high-resolution domain-adaptive semantic segmentation. In: ECCV (2022)
22. Hoyer, L., Dai, D., Van Gool, L.: Domain adaptive and generalizable network architectures and training strategies for semantic image segmentation. IEEE TPAMI **46**(1), 220–235 (2024)

23. Hoyer, L., Dai, D., Wang, H., Van Gool, L.: Mic: Masked image consistency for context-enhanced domain adaptation. In: CVPR (2023)
24. Hoyer, L., Dai, D., Wang, Q., Chen, Y., Van Gool, L.: Improving semi-supervised and domain-adaptive semantic segmentation with self-supervised depth estimation. IJCV (2023)
25. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: CVPR (2018)
26. Hu, X., Yang, K., Fei, L., Wang, K.: Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation. In: ICIP (2019)
27. Jaritz, M., Vu, T.H., de Charette, R., Wirbel, E., Pérez, P.: xMUDA: Cross-modal unsupervised domain adaptation for 3D semantic segmentation. In: CVPR (2020)
28. Lee, D.H., et al.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on challenges in representation learning, ICML. p. 896 (2013)
29. Lee, K.H., Ros, G., Li, J., Gaidon, A.: Spigan: Privileged adversarial learning from simulation. arXiv preprint arXiv:1810.03756 (2018)
30. Li, B., Hu, Y., Nie, X., Han, C., Jiang, X., Guo, T., Liu, L.: Dropkey for vision transformer. In: CVPR (2023)
31. Liu, N., Zhang, N., Han, J.: Learning selective self-mutual attention for rgb-d saliency detection. In: CVPR (2020)
32. Lopes, I., Vu, T.H., de Charette, R.: Cross-task attention mechanism for dense multi-task learning. In: WACV (2023)
33. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
34. Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: CVPR (2016)
35. Mei, K., Zhu, C., Zou, J., Zhang, S.: Instance adaptive self-training for unsupervised domain adaptation. In: ECCV (2020)
36. Pan, Y., Yao, T., Li, Y., Wang, Y., Ngo, C.W., Mei, T.: Transferrable prototypical networks for unsupervised domain adaptation. In: CVPR (2019)
37. Richter, S.R., Hayder, Z., Koltun, V.: Playing for benchmarks. In: ICCV (2017)
38. Richter, S.R., Vineet, V., Roth, S., Koltun, V.: Playing for data: Ground truth from computer games. In: ECCV (2016)
39. Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.M.: The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In: CVPR (2016)
40. Saha, S., Hoyer, L., Obukhov, A., Dai, D., Van Gool, L.: Edaps: Enhanced domain-adaptive panoptic segmentation. In: ICCV (2023)
41. Saha, S., Obukhov, A., Paudel, D.P., Kanakis, M., Chen, Y., Georgoulis, S., Van Gool, L.: Learning to relate depth and semantics for unsupervised domain adaptation. In: CVPR (2021)
42. Saito, K., Watanabe, K., Ushiku, Y., Harada, T.: Maximum classifier discrepancy for unsupervised domain adaptation. In: CVPR (2018)
43. Sakaridis, C., Dai, D., Van Gool, L.: Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In: ICCV (2021)
44. Sodano, M., Magistri, F., Guadagnino, T., Behley, J., Stachniss, C.: Robust double-encoder network for rgb-d panoptic segmentation. In: ICRA (2023)
45. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: NeurIPS (2017)

46. Tranheden, W., Olsson, V., Pinto, J., Svensson, L.: Dacs: Domain adaptation via cross-domain mixed sampling. In: WACV (2021)
47. Truong, T.D., Le, N., Raj, B., Cothren, J., Luu, K.: Fredom: Fairness domain adaptation approach to semantic scene understanding. In: CVPR (2023)
48. Tsai, Y.H., Hung, W.C., Schulter, S., Sohn, K., Yang, M.H., Chandraker, M.: Learning to adapt structured output space for semantic segmentation. In: CVPR (2018)
49. Vandenhende, S., Georgoulis, S., Van Gool, L.: Mti-net: Multi-scale task interaction networks for multi-task learning. In: ECCV (2020)
50. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017)
51. Vu, T.H., Jain, H., Bucher, M., Cord, M., Pérez, P.: Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In: CVPR (2019)
52. Vu, T.H., Jain, H., Bucher, M., Cord, M., Pérez, P.: Dada: Depth-aware domain adaptation in semantic segmentation. In: CVPR (2019)
53. Wang, Q., Dai, D., Hoyer, L., Van Gool, L., Fink, O.: Domain adaptive semantic segmentation with self-supervised depth estimation. In: ICCV (2021)
54. Wang, Y., Chen, X., Cao, L., Huang, W., Sun, F., Wang, Y.: Multimodal token fusion for vision transformers. In: CVPR (2022)
55. Wei, C., Fan, H., Xie, S., Wu, C.Y., Yuille, A., Feichtenhofer, C.: Masked feature prediction for self-supervised visual pre-training. In: CVPR. pp. 14668–14678 (2022)
56. Xie, B., Li, S., Li, M., Liu, C.H., Huang, G., Wang, G.: Sepico: Semantic-guided pixel contrast for domain adaptive semantic segmentation. IEEE TPAMI (2023)
57. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. NeurIPS (2021)
58. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: Simmim: A simple framework for masked image modeling. In: CVPR (2022)
59. Xu, D., Ouyang, W., Wang, X., Sebe, N.: Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In: CVPR (2018)
60. Xu, H., Zhang, J., Cai, J., Rezatofighi, H., Yu, F., Tao, D., Geiger, A.: Unifying flow, stereo and depth estimation. IEEE TPAMI (2023)
61. Xu, T., Chen, W., Wang, P., Wang, F., Li, H., Jin, R.: Cdtrans: Cross-domain transformer for unsupervised domain adaptation. arXiv preprint arXiv:2109.06165 (2021)
62. Zhang, J., Liu, H., Yang, K., Hu, X., Liu, R., Stiefelhagen, R.: Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. IEEE Transactions on Intelligent Transportation Systems (2023)
63. Zhang, P., Zhang, B., Zhang, T., Chen, D., Wang, Y., Wen, F.: Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In: CVPR (2021)
64. Zhang, Q., Zhang, J., Liu, W., Tao, D.: Category anchor-guided unsupervised domain adaptation for semantic segmentation. NeurIPS (2019)
65. Zhang, W., Ouyang, W., Li, W., Xu, D.: Collaborative and adversarial network for unsupervised domain adaptation. In: CVPR (2018)
66. Zhang, Z., Cui, Z., Xu, C., Yan, Y., Sebe, N., Yang, J.: Pattern-affinitive propagation across depth, surface normal and semantic segmentation. In: CVPR (2019)
67. Zhou, Q., Feng, Z., Gu, Q., Pang, J., Cheng, G., Lu, X., Shi, J., Ma, L.: Context-aware mixup for domain adaptive semantic segmentation. IEEE Transactions on Circuits and Systems for Video Technology (2022)

68. Zou, Y., Yu, Z., Kumar, B., Wang, J.: Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In: ECCV (2018)