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

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

[1] Technical University of Munich [2] ETH Zurich [3] NVIDIA [4] 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 depth predictions, as depth discontinuities often coincide with segmentation boundaries. We show that naively incorporating depth does not fully exploit its potential. To this end, we present MICDrop, which learns a joint feature representation by masking image encoder features by 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. We further propose a feature fusion module to improve both global and local information sharing. MICDrop can be plugged into various recent UDA methods and consistently improves results across standard UDA benchmarks, obtaining new state-of-the-art performances. Project Page:* `https://github.com/ly-muc/MICDrop`

1. Introduction

The computer vision community has seen tremendous success in recognition tasks over the years, yet the issue of labor-intensive labeled images [7, 43] for supervised training of neural networks persists. Alternatively, a simulator can easily obtain images and corresponding segmentation labels at a large scale. However, models trained on synthetic datasets experience a noticeable performance decline when applied to real-world data due to the variance in data distributions (*e.g.*, the appearance of objects), a phenomenon known as domain shift. This paper focuses on Unsupervised Domain Adaptation (UDA), where a model is trained on labeled synthetic source and unlabeled real-world target domain data.

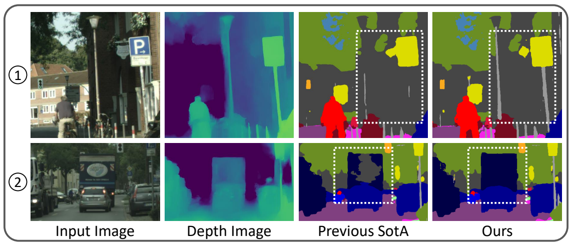

Figure 1. **Qualitative Examples.** Previous UDA methods, *e.g.*, MIC [21] struggle with the segmentation of fine structures (top) and oversegmentation of difficult objects (bottom). MICDrop improves semantic segmentation UDA with depth estimates, capturing fine structures and consistency within object boundaries.

Current challenges in UDA: Recent UDA methods [4, 19–21, 24, 27, 55] significantly reduce the gap to a fully supervised method, yet they still struggle with two main aspects: (1) Fine structures and high-frequency details, despite using high-resolution strategies such as HRDA [20]. (2) Oversegmentation when visual appearance clues are ambiguous. Motivated by these issues, we recognize that *geometric representation* could provide complementary cues 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 visual features of the back of the truck resemble a building, however, the smooth depth within the boundaries suggests a consistent semantic class. While measured depth might not be available, advances in image-based depth estimation [10, 59] enable us to explore the task in a general setting.

Contributions: We propose a streamlined approach for leveraging depth in UDA, using a novel *cross-modality complementary dropout* technique along with a tailored masking schedule. Our masking strategy fosters 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. We also propose a *cross-modality feature fusion* module, 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 segmentation with global depth cues. Second, it applies local self-attention to depth features, leveraging the discontinuity in local depth for describing boundaries thus identifying thin structures. In summary, our key contributions are: (1) A **complementary feature masking strategy** for depth and RGB, fostering cross-modal feature learning. (2) A **cross-modality fusion module** to improve segmentation based on depth by using global and local cues. (3) 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 improves modern UDA methods, we hope to lay the foundation for future research exploring the merits of auxiliary modalities for UDA.

2. Related Work

Unsupervised Domain Adaptation (UDA): In UDA, methods can mostly be categorized into adversarial training [16, 17, 42, 47, 51] and self-training [11, 28]. In self-training, pseudo labels are created by a teacher network [1, 35, 36, 62–64, 67]. The student model then receives an augmented [1, 45, 66] image version. Self-training can further be strengthened by domain-robust Transformers [19, 23, 41, 48, 60], class-balanced sampling [19, 68], multi-resolution adaptation [20], or contrastive learning [4, 55]. 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 as an auxiliary task [12, 18, 22, 26, 29, 32, 40, 49, 52, 53, 58, 65]. Our method, however, is more closely related to RGB-D semantic segmentation [5, 25]. In contrast to previous RGB-D works such as [5, 31, 61], we leverage both local and global dependencies for domain-robust segmentation and show in Tab. 3b that leveraging geometric cues is not trivial in the context of UDA.

Masked Image Modeling (MIM): In MIM [2, 3, 8, 14, 54, 57], information is withheld to train the network to recover certain targets. Different from MIC [21], we propose a novel complementary multi-modal feature dropout to facilitate cross-modality learning. In Sec. 4, we show that our method is orthogonal to MIC and can further improve it.

3. Method

Preliminaries. In Figs. 2 and 3, we present our architecture and training scheme, featuring two novel modules that can be plugged into various UDA methods to leverage geometric cues. Our feature fusion module integrates auxiliary inputs via global and local attention-based aggregation. The masking module ensures balanced input usage, avoiding pure reliance on a single input modality. 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 ($\mathbf{X}_t$) during training. 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. The network is typically trained in a supervised manner on a source domain. To leverage unlabeled target data, we follow recent approaches by adopting a student-teacher [1, 19–21, 45, 46, 55] framework. Here, we present the student with a heavily augmented view [45], while the teacher receives a weakly augmented image. We study the effectiveness of our proposed method by extending pretrained encoders [19, 20].

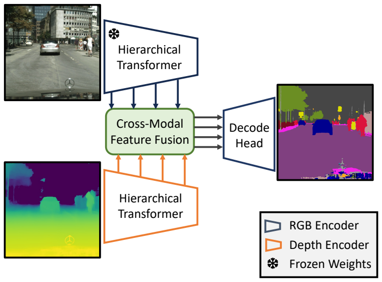

Figure 2. **Architecture.** We use a light-weight depth encoder and process the features in our cross-modal feature fusion module.

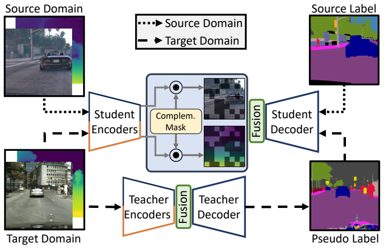

Figure 3. **Training Scheme.** In our training pipeline, source and target images are fed through the student encoders, which apply our proposed cross-modality complementary dropout. Following our feature fusion block, the decoder makes the final prediction.

Multi-Modal Feature Fusion: To achieve a *light-weight* training pipeline, we construct a multi-modality encoder that contains two individual encoders. A newly trained light-weight depth encoder to produce depth features and a trained RGB multi-scale feature encoder. As seen in Fig. 4, our feature fusion block is divided into (1) *global* depth-guided cross-attention, (2) *local* self-attention and (3) *residual fusion*. 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. Thus, the purpose of the global branch is to aggregate RGB features globally based on their corresponding

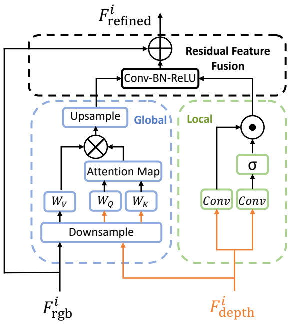

Figure 4. **Feature fusion of RGB and depth.** The local attention module refines depth information within a local window. The global attention module aggregates image features based on similarity in their corresponding depth features. Finally, the residual feature fusion block fuses all features.

depth feature similarity. To address the problematic scaling behavior of (global) cross-attention [50], we downsample low- and high-level feature maps with a pooling factor of $\{4, 2, 1, 1\}$ during inference and $\{2, 1, 1, 1\}$ during training. 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 $\mathbf{W}^i_k$. However, modeling global interactions on a downsized resolution might not be enough to capture the fine details of objects like sign or pole. Specifically, local *depth discontinuities* provide strong cues for *boundary regions* among semantic classes, while *smooth and continuous depth* indicate *no change* in semantics. Keeping the same computational complexity problem in mind, we draw inspiration from earlier work [15, 58] and restrict the self-attention to a local window without pooling, using two 3×3 convolutions. As this branch is used to model *complementary* features, we exclusively use depth features. Formally, we compute the local self-attention as: $\mathbf{F}^i_{\text{local}} = \sigma\left(\text{Conv}_{3\times3}\left(\mathbf{F}^i_{\text{depth}}\right)\right) \odot \text{Conv}_{3\times3}\left(\mathbf{F}^i_{\text{depth}}\right)$. After aggregating global and local features, we use a simple two-step feature fusion block to obtain the refined features: $\mathbf{F}^i_{\text{refined}} = \mathbf{F}^i_{\text{rgb}} + \text{ReLU}(\text{BN}(\text{Conv}(\mathbf{F}^i_{\text{global}}||\mathbf{F}^i_{\text{local}})))$. The refined features are fed to the decode head for final predictions.

Complementary Feature Masking: During initial experiments, we observed 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. For this, we propose using *blockwise dropout* [9] to generate masked features. Masking larger blocks prevents easy recovery from the neighborhood in the same modality and requires understanding the semantics of the other modality. Furthermore, we hypothesize that learning features across modalities can be achieved best by masking the feature maps of different modalities in a *complementary* fashion, as illustrated in Fig. 2. We experimentally validate that design in Sec. 4. Formally, we define complementary masking as:

$$\mathbf{M}_{\text{rgb}}(u, v) = [\gamma > m^t_r], \quad \gamma \sim \text{Uniform}(0, 1) \qquad (1)$$

$$\mathbf{M}_{\text{depth}}(u, v) = 1 - \mathbf{M}_{\text{rgb}}(u, v) \qquad (2)$$

where m^t_r denotes the masking ratio at iteration t and (u, v) the block index of the i-th feature map. 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. Similar to prior studies [9, 30], 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. In early training stages, we keep a high proportion of depth features to accelerate the depth encoder training and improve its feature quality. During training, we gradually reduce depth feature retention.

4. Experiments

Implementation: We utilize two widely used UDA benchmarks. The synthetic source domain datasets consist of GTA [37], 24,966 images with a resolution of 1914×1052, and SYNTHIA [39], 9,400 images with a resolution of 1280×760. The target dataset Cityscapes [7] includes 2975 training and 500 validation images each with a resolution of 2048×1024. We employ self-supervised monocular depth estimation, MonoDepth2 [10], trained on image sequences from VIPER(GTA) [38] and SYNTHIA-SEQ, for the source domain. For the target domain, we obtain disparity estimations from UniMatch [59], trained on a synthetic dataset [34]. To demonstrate the *plugin* capability of MICDrop we also apply it to state-of-the-art methods [19–21], using a light-weight MiT-B3 [56] depth encoder. We initialize the RGB encoder and decode head with the publicly available pre-trained weights and the depth encoder with ImageNet weights. We use an AdamW [33] optimizer with a learning rate of 6×10^{-5} for the depth encoder and 6×10^{-4} for the remaining modules. Furthermore, we apply a linear warm-up for 1.5k iterations followed by polynomial decay with a factor of 0.9. Following prior works [19, 21, 45], we use an EMA [44] teacher with α=0.999, batch size of 2 and strong data augmentations [45]. We freeze the RGB encoder and train the rest of the network for 20k iterations.

Main Results: To validate the effectiveness of MICDrop and its *plugin* capabilities, we evaluate the mean Intersection over Union (mIoU) across the three state-of-the-art methods [19–21] as shown in Tab. 1 using the classes shared by source and target domain. Applying MICDrop on GTA, the results improve current methods DAFormer [19], $\text{MIC}_{DAFormer}$ [21] and HRDA [20] by 1.8 mIoU, 1.2 mIoU and 1.0 mIoU respectively. Using DAFormer with a ResNet-101 backbone, in Tab. 1, MICDrop achieves a significant gain of 4.1 mIoU and outperforms the previous SOTA depth-guided UDA method CorDA [53]. Building on top the best performing model MIC_{HRDA}, we can further boost results

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.)																				
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
DAFormer [19]	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
DAFormer [19]	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$_{DAFormer}$ [21]	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 [20]	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$_{HRDA}$ [21]	*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.)																				
DAFormer [19]	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 [20]	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$_{HRDA}$ [21]	*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 [13] and Segformer [56] backbones. On both, GTA and SYNTHIA, MICDrop achieves consistent improvements demonstrating the effectiveness of our masking strategy and fusion module.

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$_{HRDA}$ [21]	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 (dilation 0.005).

by 0.7 mIoU, a significant improvement considering the saturation of this benchmark (94% of the oracle performance). Thus, our light-weight modules and complementary dropout consistently show improvements, offering orthogonal contribution to input masking [21]. Diving into the details, we notice predominately improvements 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 aggregating local depth features, as these local depth discontinuities at boundary regions serve as a strong cue. Second, large low prevalence classes, *e.g.*, truck, bus, or train, benefit from both global and local depth features. Due to their size, global reasoning can improve the segmentation consistency, while locally smooth, continuous depth reduces the likelihood of changes in the semantics within a local window. In Tab. 1 MICDrop achieves consistent improvements on SYNTHIA, *i.e.* 1.1 mIoU for DAFormer, 1.0 mIoU for HRDA, and 0.6 mIoU for MIC$_{HRDA}$.

Boundary Analysis: Tab. 2 additionally studies the boundary IoU [6]. Compared to the standard IoU it improves by a significantly larger margin (1.6 vs 0.7). MICDrop particularly improves fine structures (*e.g.*, pole or sign) and classes that are prone to oversegmentation (*e.g.*, truck and building), quantitatively supporting our motivation in Fig 1.

Ablation Studies: For a fair comparison, we finetune the pretrained baseline model without any changes but did not observe any performance improvements (68.3 ±0.2 mIoU). We validate our design by exploring various masking and feature fusion strategies using DAFormer and the GTA benchmark.

Masking Strategy	Mask RGB	Mask 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
Local Self-Attn	69.7 ±0.1
Global Cross-Attn	68.1 ±0.8
Both (ours)	**70.1** ±0.1

(b) **Feature Fusion ablation.**

Table 3. **Ablation study.** We use DAFormer [19] 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 standard deviation are reported over 3 seeds.

Cross-Modal Complementary Dropout: Adding depth information without masking improves our baseline by 0.8 mIoU. Additional RGB features masking further improves by 0.2 mIoU. However, applying *independent* masking to both RGB and depth features exhibits no improvements, resulting in a high standard deviation. 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, and 1.8 mIoU over the DAFormer baseline. We demonstrate that complementary masking across levels is crucial, as having different masks per level leads to a decrease of 0.4 mIoU. These findings support that complementary masking is crucial for leveraging depth information for semantic segmentation in UDA, achieving a great balance between geometric and visual scene information.

Feature Fusion: In Tab. 3b, we first explore a naive addition of features and state-of-the-art RGB-D method CMX [61], both of which only attain suboptimal performance in our context. Turning our focus to the *individual* efficacy of our global and local feature fusion blocks, we observe that the local self-attention block outperformed our naive baseline, whereas the global cross-attention block exhibits training instability with no improvement. Analogous to the results observed with CMX, these findings underscore the significance of controlling the flow of local information and the importance of the *complement* design of the local and global attention mechanism. This combination led to a notable improvement of an additional 0.4 mIoU over local self-attention, achieving an overall gain of 1.8 mIoU over the baseline [19].

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 between 0.7 and 1.8 mIoU, achieving state-of-the-art results. Thus, MICDrop demonstrates the effectiveness of utilizing depth in UDA without retraining existing encoders. The plugin design of MICDrop facilitates easy integration into future domain-adaptive semantic segmentation methods. We hope that our simple and effective approach inspires further research into leveraging complementary cues in UDA.

References

[1] Nikita Araslanov and Stefan Roth. Self-supervised augmentation consistency for adapting semantic segmentation. In *CVPR*, 2021. 2

[2] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *ECCV*, 2022. 2

[3] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. In *ICLR*, 2021. 2

[4] Mu Chen, Zhedong Zheng, Yi Yang, and Tat-Seng Chua. Pipa: Pixel-and patch-wise self-supervised learning for domain adaptative semantic segmentation. *ACM Multimedia*, 2023. 1, 2

[5] Xiaokang Chen, Kwan-Yee Lin, Jingbo Wang, Wayne Wu, Chen Qian, Hongsheng Li, and Gang Zeng. Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation. In *ECCV*, 2020. 2

[6] Bowen Cheng, Ross Girshick, Piotr Dollár, Alexander C. Berg, and Alexander Kirillov. Boundary IoU: Improving object-centric image segmentation evaluation. In *CVPR*, 2021. 4

[7] 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 *CVPR*, 2016. 1, 3

[8] Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, Nenghai Yu, and Baining Guo. Peco: Perceptual codebook for bert pre-training of vision transformers. In *AAAI*, 2023. 2

[9] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. Dropblock: A regularization method for convolutional networks. *NeurIPS*, 2018. 3

[10] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *ICCV*, 2019. 1, 3

[11] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. *NeurIPS*, 17, 2004. 2

[12] Vitor Guizilini, Jie Li, Rareș Ambruș, and Adrien Gaidon. Geometric unsupervised domain adaptation for semantic segmentation. In *ICCV*, 2021. 2

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4

[14] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 2

[15] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 1997. 3

[16] 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. 2

[17] 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 *ICML*, 2018. 2

[18] Lukas Hoyer, Dengxin Dai, Yuhua Chen, Adrian Koring, Suman Saha, and Luc Van Gool. Three ways to improve semantic segmentation with self-supervised depth estimation. In *CVPR*, pages 11130–11140, 2021. 2

[19] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *CVPR*, 2022. 1, 2, 3, 4

[20] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Hrda: Context-aware high-resolution domain-adaptive semantic segmentation. In *ECCV*, 2022. 1, 2, 3, 4

[21] Lukas Hoyer, Dengxin Dai, Haoran Wang, and Luc Van Gool. Mic: Masked image consistency for context-enhanced domain adaptation. In *CVPR*, 2023. 1, 2, 3, 4

[22] 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. *IJCV*, 2023. 2

[23] Lukas Hoyer, David Joseph Tan, Muhammad Ferjad Naeem, Luc Van Gool, and Federico Tombari. SemiVL: Semi-supervised semantic segmentation with vision-language guidance. *arXiv preprint arXiv:2311.16241*, 2023. 2

[24] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Domain adaptive and generalizable network architectures and training strategies for semantic image segmentation. *IEEE TPAMI*, 46 (1):220–235, 2024. 1

[25] Xinxin Hu, Kailun Yang, Lei Fei, and Kaiwei Wang. Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation. In *ICIP*, 2019. 2

[26] Maximilian Jaritz, Tuan-Hung Vu, Raoul de Charette, Emilie Wirbel, and Patrick Pérez. xMUDA: Cross-modal unsupervised domain adaptation for 3D semantic segmentation. In *CVPR*, 2020. 2

[27] Yuru Jia, Lukas Hoyer, Shengyu Huang, Tianfu Wang, Luc Van Gool, Konrad Schindler, and Anton Obukhov. Dginstyle: Domain-generalizable semantic segmentation with image diffusion models and stylized semantic control. *arXiv preprint arXiv:2312.03048*, 2023. 1

[28] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, page 896, 2013. 2

[29] Kuan-Hui Lee, German Ros, Jie Li, and Adrien Gaidon. Spigan: Privileged adversarial learning from simulation. *arXiv preprint arXiv:1810.03756*, 2018. 2

[30] Bonan Li, Yinhan Hu, Xuecheng Nie, Congying Han, Xiangjian Jiang, Tiande Guo, and Luoqi Liu. Dropkey for vision transformer. In *CVPR*, 2023. 3

[31] Nian Liu, Ni Zhang, and Junwei Han. Learning selective self-mutual attention for rgb-d saliency detection. In *CVPR*, 2020. 2

[32] Ivan Lopes, Tuan-Hung Vu, and Raoul de Charette. Cross-task attention mechanism for dense multi-task learning. In *WACV*, 2023. 2

[33] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 3

[34] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *CVPR*, 2016. 3

[35] Ke Mei, Chuang Zhu, Jiaqi Zou, and Shanghang Zhang. Instance adaptive self-training for unsupervised domain adaptation. In *ECCV*, 2020. 2

[36] Yingwei Pan, Ting Yao, Yehao Li, Yu Wang, Chong-Wah Ngo, and Tao Mei. Transferrable prototypical networks for unsupervised domain adaptation. In *CVPR*, 2019. 2

[37] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *ECCV*, 2016. 3

[38] Stephan R Richter, Zeeshan Hayder, and Vladlen Koltun. Playing for benchmarks. In *ICCV*, 2017. 3

[39] 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 *CVPR*, 2016. 3

[40] 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 *CVPR*, 2021. 2

[41] Suman Saha, Lukas Hoyer, Anton Obukhov, Dengxin Dai, and Luc Van Gool. Edaps: Enhanced domain-adaptive panoptic segmentation. In *ICCV*, 2023. 2

[42] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *CVPR*, 2018. 2

[43] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *ICCV*, 2021. 1

[44] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 2017. 3

[45] Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. Dacs: Domain adaptation via cross-domain mixed sampling. In *WACV*, 2021. 2, 3

[46] Thanh-Dat Truong, Ngan Le, Bhiksha Raj, Jackson Cothren, and Khoa Luu. Fredom: Fairness domain adaptation approach to semantic scene understanding. In *CVPR*, 2023. 2

[47] 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 *CVPR*, 2018. 2

[48] Ozan Unal, Dengxin Dai, Lukas Hoyer, Yigit Baran Can, and Luc Van Gool. 2d feature distillation for weakly-and semi-supervised 3d semantic segmentation. In *WACV*, pages 7336–7345, 2024. 2

[49] Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Mti-net: Multi-scale task interaction networks for multi-task learning. In *ECCV*, 2020. 2

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 3

[51] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *CVPR*, 2019. 2

[52] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Dada: Depth-aware domain adaptation in semantic segmentation. In *CVPR*, 2019. 2

[53] Qin Wang, Dengxin Dai, Lukas Hoyer, Luc Van Gool, and Olga Fink. Domain adaptive semantic segmentation with self-supervised depth estimation. In *ICCV*, 2021. 2, 3, 4

[54] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *CVPR*, pages 14668–14678, 2022. 2

[55] Binhui Xie, Shuang Li, Mingjia Li, Chi Harold Liu, Gao Huang, and Guoren Wang. Sepico: Semantic-guided pixel contrast for domain adaptive semantic segmentation. *IEEE TPAMI*, 2023. 1, 2

[56] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *NeurIPS*, 2021. 3, 4

[57] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *CVPR*, 2022. 2

[58] Dan Xu, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In *CVPR*, 2018. 2, 3

[59] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *IEEE TPAMI*, 2023. 1, 3

[60] 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. 2

[61] Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *IEEE Transactions on Intelligent Transportation Systems*, 2023. 2, 4

[62] 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 *CVPR*, 2021. 2

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

[64] Weichen Zhang, Wanli Ouyang, Wen Li, and Dong Xu. Collaborative and adversarial network for unsupervised domain adaptation. In *CVPR*, 2018. 2

[65] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Yan Yan, Nicu Sebe, and Jian Yang. Pattern-affinitive propagation across depth, surface normal and semantic segmentation. In *CVPR*, 2019. 2

[66] 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

[67] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *ECCV*, 2018. 2

[68] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *ECCV*, 2018. 2