UniDAformer: Unified Domain Adaptive Panoptic Segmentation Transformer via Hierarchical Mask Calibration (Supplementary Materials)

A. Experiments

A.1. Datasets

We evaluate UniDAformer over four datasets that have been widely used in domain adaptive visual recognition tasks:
1) Cityscapes [8] aims for street scene understanding and autonomous driving. The images are collected under normal weather conditions from 50 European cities, including 2,975 images as training set and 500 images as validation set, with panoptic segmentation annotations of 30 categories. All the images have the same resolution of 1024 × 2048.
2) Foggy cityscapes [21] is a synthesized dataset that is derived on top of Cityscapes by including simulated fog. This dataset inherits the dense panoptic segmentation annotations of 30 categories from Cityscapes.
3) SYNTHIA [20] is a large-scale synthesized dataset with 9,400 images that stimulated with virtual environments. This dataset provides both instance-level annotations for instance segmentation and pixel-level annotations for semantic segmentation. We generate panoptic segmentation annotations by fusing instance-level annotations and pixel-level annotations. All the images have the same resolution of 760 × 1280.
4) VIPER [19] is a visual perception benchmark with more than 250K high-resolution video frames collected from the computer game Grand Theft Auto V (GTA5). This dataset provides various types of annotations, including instance-level annotations and pixel-level annotations. We generate panoptic segmentation annotations by fusing instance-level annotations and pixel-level annotations. All the images have the same resolution of 1080 × 1920.

A.2. Implementation Details

For the experiments with DETR [4], we employ ResNet-50 [13] as backbone (pre-trained on ImageNet [9]). We adopt SGD optimizer [3] with a momentum 0.9 and a weight decay $1e-4$. The initial learning rate is $2e-4$. Note we follow [10] to modify the DETR [4] architecture and additionally add a semantic segmentation [15] head on it.

For the experiments with PSN [16], we adopt Deeplab-V2 [5] as semantic branch network and Mask R-CNN [12] as instance segmentation branch. The experiments employ ResNet-101 [13] as backbone (pre-trained on ImageNet [9]). The model are trained with the SGD optimizer [3] with learning rate 2.5×10^{-4}, momentum 0.9, and weight decay 10^{-4}.

For hyper-parameters in UniDAformer, we fix the momentum coefficient γ at 0.999 as in [11] and the update coefficient γ' in Eq.6 at 0.999. In Superpixel-wise Calibration, we compute superpixels using SLIC algorithm [1] and the number of superpixels I is fixed at 500.

A.3. Discussions

A.3.1 Parameter Analysis

Momentum Model Coefficient. As described in Section Method in the main text, we adopt a momentum model for pseudo label generation as its slow and smooth parameter updates facilitates more stable and consistent pseudo labels generation along the training process. Here we study how momentum model coefficient γ affects the adaptation performance over task SYNTHIA → Cityscapes with DETR [4]. As Table 1 shows, UniDAformer is robust when γ is large enough (from 0.99 to 0.9999) while its performance starts to drop slightly when γ becomes too small, which further demonstrates the effectiveness of the slow-update momentum model for pseudo label generation.

Update Coefficient. The update coefficient γ' in Eq.6 in the main text decides the update speed of mask centroids: the smaller it is, the faster mask centroids change. Table 2 shows it affects segmentation over SYNTHIA → Cityscapes with

Method	Momentum Model Coefficient γ														
	0.5			0.9			0.99			0.999			0.9999		
	mSQ	mRQ	mPQ	mSQ	mRQ	mPQ	mSQ	mRQ	mPQ	mSQ	mRQ	mPQ	mSQ	mRQ	mPQ
UniDAformer	60.6	39.2	30.8	62.5	40.3	31.7	63.1	42.3	32.8	64.7	42.2	33.0	64.2	41.7	32.9

Table 1. The momentum model coefficient γ affects domain adaptation. The experiments are conducted over task SYNTHIA → Cityscapes.

DETR [4]. We can see that UniDAformer is robust when γ' is relatively large (from 0.99 to 0.999) otherwise its performance starts to drop slightly. This shows that smooth centroid update is helpful, without which the mask centroid and the calibrated pseudo masks could become unstable.

Method	Update Coefficient γ'														
	0.5			0.9			0.99			0.999			0.9999		
	mSQ	mRQ	mPQ	mSQ	mRQ	mPQ	mSQ	mRQ	mPQ	mSQ	mRQ	mPQ	mSQ	mRQ	mPQ
UniDAformer	62.2	40.8	31.5	62.4	41.7	32.5	64.5	42.0	32.8	64.7	42.2	33.0	64.3	42.1	32.9

Table 2. The mask centroid update coefficient γ' defined in Eq.6 affects domain adaptation. The experiments are conducted over task SYNTHIA → Cityscapes.

Number of Superpixels. Parameter I decides the number of the computed superpixels in Superpixel Calibration. We studied the sensitivity of I over task SYNTHIA → Cityscapes. As Table 3 shows, UniDAformer is quite tolerant to parameter I and the best performance is obtained when $I = 500$.

Method	The Number of Superpixels I														
	400			500			600			700			1000		
	mSQ	mRQ	mPQ	mSQ	mRQ	mPQ	mSQ	mRQ	mPQ	mSQ	mRQ	mPQ	mSQ	mRQ	mPQ
UniDAformer	63.7	42.2	32.7	64.7	42.2	33.0	63.8	41.9	32.8	62.4	41.3	32.6	63.7	42.5	32.9

Table 3. The number of superpixels I affects domain adaptation. The experiments are conducted over task SYNTHIA → Cityscapes.

A.3.2 Comparison with Existing Online Self-training Methods

We compare UniDAformer with several semantic segmentation methods [2,18] that perform online self-training. In particular, existing online self-training methods directly enforce pixel-wise prediction consistency across different image augmentations for learning robust representation. In contrast, UniDAformer focuses on the false prediction issue and introduces HMC to calibrate pseudo masks during self-training. Experiments in Table 4 show that UniDAformer outperforms [18] and [2] clearly, largely due to the proposed HMC that helps to generate more accurate pseudo masks under massive false predictions.

PixMatch [18]			SAC [2]			UniDAformer		
mSQ	mRQ	mPQ	mSQ	mRQ	mPQ	mSQ	mRQ	mPQ
60.7	34.2	25.9	60.7	34.2	25.9	64.7	42.2	33.0

Table 4. Comparison with existing online self-training UDA methods [2,18] using unified panoptic segmentation architecture [4] over task SYNTHIA → Cityscapes.

A.3.3 Comparison with Existing Superpixel-based UDA Methods

As described in Section Method in the main text, the core of HMC is a hierarchical design that corrects pseudo labels jointly and collaboratively based on information flowing across the three levels. Superpixel-wise calibration is the middle stage that plays a role in linking coarse region-wise calibration and fine pixel-wise calibration. To the best of our knowledge, superpixel was not used for domain adaptive panoptic segmentation before though it has been used for domain adaptive semantic segmentation [23, 24]. However, [23, 24] use superpixels to capture and transfer spatial relations which is very different from our work in terms of tasks, motivations, and objectives of using superpixels. Additionally, Table 5 shows that UniDAformer outperforms [23] and [24] clearly, largely because the proposed coarse-to-fine calibration ($i.e.$, from region to superpixel and pixel) helps to rectify the pseudo masks effectively.

CDA [23]			RPT [24]			UniDAformer		
mSQ	mRQ	mPQ	mSQ	mRQ	mPQ	mSQ	mRQ	mPQ
62.1	33.8	25.6	61.2	34.7	26.3	64.7	42.2	33.0

Table 5. Comparison with existing superpixel-based UDA methods [23, 24] using unified panoptic segmentation architecture [4] over task SYNTHIA → Cityscapes.

A.3.4 Synergetic Experiments

As described in Section Introduction in the main text, our Hierarchical Mask Calibration (HMC) introduces little extra computation overhead ($i.e.$, 0.01% parameters overhead) and could be used as a plug-in. Here we study the synergetic benefits of the proposed HMC by incorporating it into several domain adaptation methods under task SYNTHIA → Cityscapes. As table 6 shows, the incorporation of HMC improves domain adaptation methods consistently across all evaluation metrics ($i.e.$, mSQ, mRQ and mPQ). The experiment results indicate that the proposed HMC is complementary to existing domain adaptation methods and could be used as a plug-in with consistent performance improvement.

Method	Base			+ HMC			Gain		
	mSQ	mRQ	mPQ	mSQ	mRQ	mPQ	mSQ	mRQ	mPQ
DAF [6]	59.0	28.3	20.9	62.2	39.9	30.1	+3.2	+11.6	+9.2
ADVENT [22]	60.2	33.0	25.0	63.7	41.9	32.5	+3.5	+8.9	+7.5
CVRN [14]	61.4	35.9	27.9	63.3	43.4	33.6	+1.9	+7.5	+5.7

Table 6. Synergetic Experiments of HMC: the proposed HMC can be used as a plug-in and the incorporation of HMC brings performance improvement to existing domain adaptation methods consistently.

A.3.5 Generalization across Different Unified Architectures

We examine the proposed UniDAformer over other two unified architectures, $i.e.$, MaskFormer [7] and PanopticFCN [17] under task SYNTHIA → Cityscapes. The experimental results in Table 7 show that our UniDAformer achieves similar improvements as with the DETR [4] architecture, indicating that UniDAformer can work with different unified architectures with consistent improvements.

Method	Baseline			UniDAformer			Gain		
	mSQ	mRQ	mPQ	mSQ	mRQ	mPQ	mSQ	mRQ	mPQ
MaskF [7]	56.6	19.2	16.2	60.7	34.2	25.9	+4.1	+15.0	9.7
PanFCN [17]	47.5	19.7	15.8	60.5	34.8	26.5	+13.0	+15.1	10.7

Table 7. Generalization across different unified architectures [7, 17] over task SYNTHIA → Cityscapes.

A.3.6 Why Unified Networks Have Poor Adaptation Baseline?

Table 1 in the main text shows that most unified panoptic segmentation networks outperform traditional multi-branch panoptic segmentation network by large margins under the supervised setup while opposite results are observed under unsupervised domain adaptation setup. Here we provide possible insights behind it. In panoptic segmentation, things and stuff predictions rely on different types of features [17], $i.e.$, things prediction requires instance-aware features that vary among semantic categories as well as different instance identities, while stuff prediction requires semantic-consistency features that vary according to different semantic categories only. Several recent unified panoptic segmentation works [4, 7, 17] propose to tackle such conflict for supervised learning. However, these works can not handle such conflict under unsupervised domain adaptation setup as they were designed for supervised learning where the ground-truth annotations are available. As a result, the feature conflict between things and stuff still exists in unified domain adaptation panoptic segmentation problem and leads to large performance drops as shown in Table 1 of the main text. Specifically, such drops are reflected primarily in severe false prediction issue as illustrated in Fig.4 (a) of the main text.

B. Qualitative Results

B.1. Visualization of the Calibration Process

We present visual illustrations of Hierarchical Mask Calibration over task SYNTHIA → Cityscapes. Fig. 1 shows the corresponding calibrated mask over each level. It can be seen that the HMC-calibrated masks M' achieve higher IoU than the original pseudo masks $\hat{M}$, which indicates the superior ability of HMC in correcting pseudo masks.

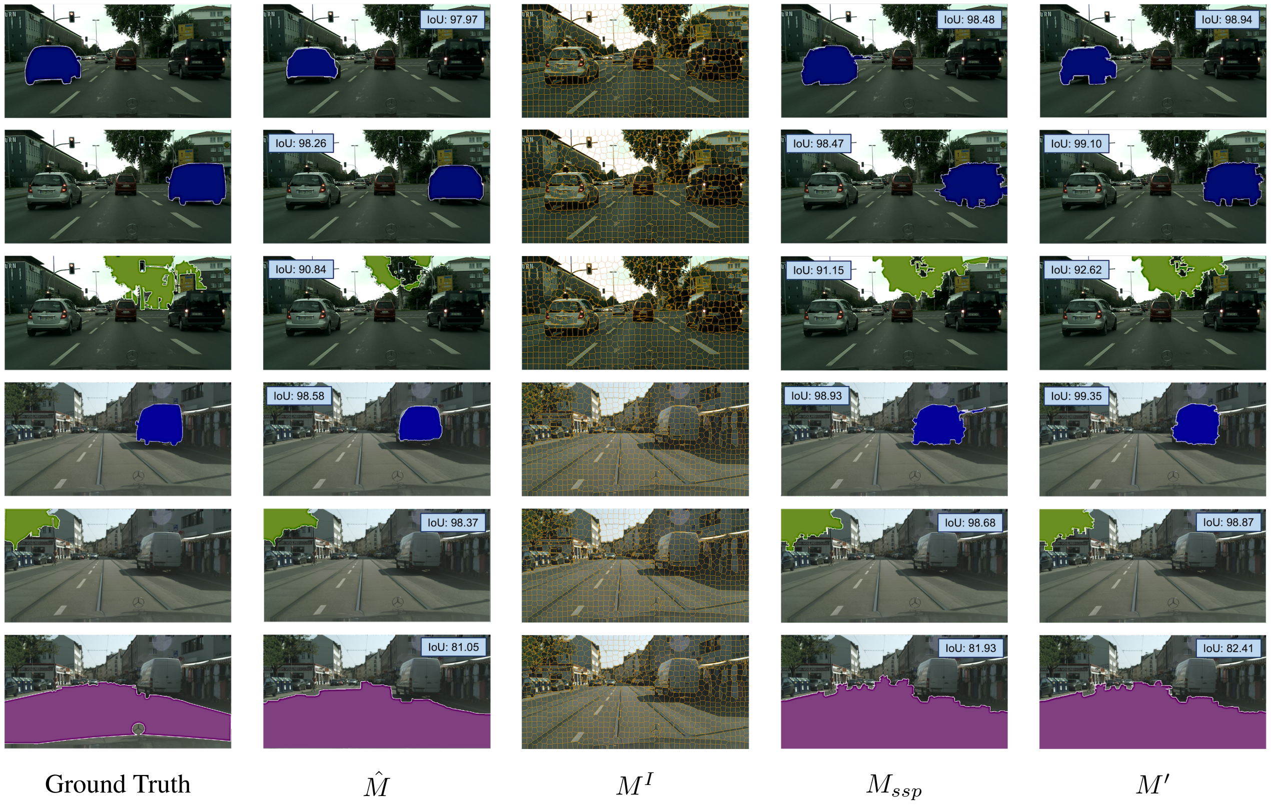

Figure 1. Visualization of HMC-calibrated masks over task SYNTHIA → Cityscapes. We sample 6 masks from 2 images as examples. The ground truth of each mask is shown in the first column, and original pseudo mask $\hat{M}$, computed superpixels M^I, superpixel-wise calibrated mask M_{sp} and pixel-wise calibrated mask M' are shown from the second to fifth columns. The insertion over union (IoU) is computed between each mask and its corresponding ground-truth mask.

B.2. Qualitative Comparisons

We present qualitative illustrations and comparisons over task SYNTHIA → Cityscapes, Cityscapes → Foggy cityscapes and VIPER → Cityscapes. As Fig. 2 show, UniDAformer yields the best segmentation with more true positives and less false predictions consistently which is aligned well with the quantitative results.

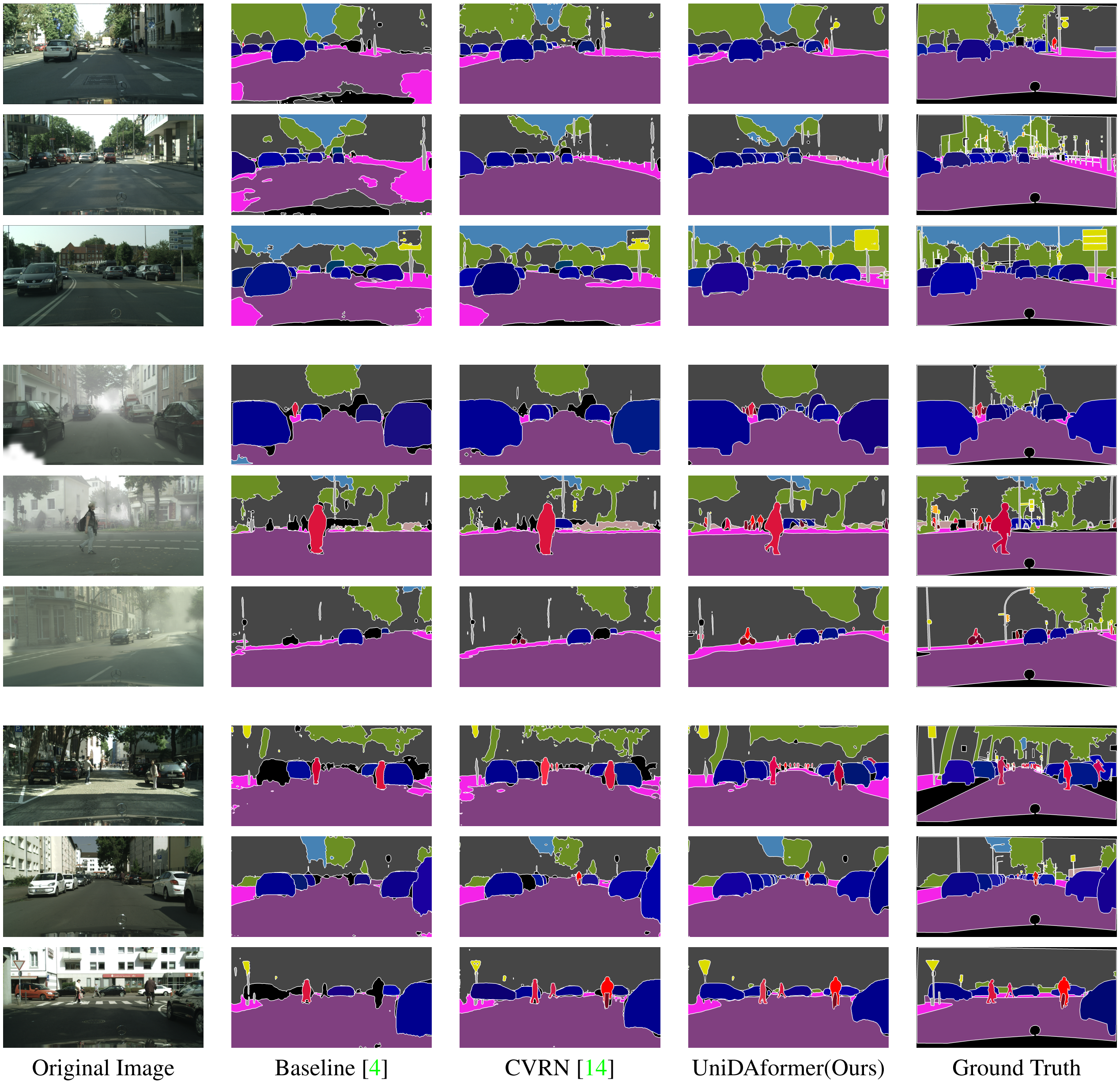

Figure 2. Qualitative comparison of UniDAformer with the baseline model [4] and CVRN [14] over three tasks including SYNTHIA → Cityscapes as shown in rows 1-3, Cityscapes → Foggy Cityscapes as shown in rows 4-6 and VIPER → Cityscapes as shown in rows 7-9, respectively. The proposed UniDAformer yields best segmentation with more true positives, less false predictions and finer boundary.

C. Social Impacts and Limitations

This work presents a new unified learning pipeline for domain adaptive panoptic segmentation, which has clear advantages in less parameters and simpler training and inference pipeline. In another word, unified domain adaptive panoptic segmentation benefits the computer vision community by providing a new solution for domain adaptive panoptic segmentation that involves much fewer parameters and simplifies the training and inference pipeline greatly. However, the explored techniques in this work are still at an early stage and thus our proposed method can serve as an auxiliary tool in applications instead of the hard control system that could lead to harmful consequences.

References

[1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012. 1

[2] 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] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010. 1

[4] 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. 1, 2, 3, 4, 5

[5] 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. 1

[6] 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 conference on computer vision and pattern recognition*, pages 3339–3348, 2018. 3

[7] 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, 2021. 3, 4

[8] 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 conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 1

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[10] Bin Dong, Fangao Zeng, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. Solq: Segmenting objects by learning queries. *Advances in Neural Information Processing Systems*, 34:21898–21909, 2021. 1

[11] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 1

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[14] 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. 3, 5

[15] 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

[16] 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

[17] Yanwei Li, Hengshuang Zhao, Xiaojuan Qi, Liwei Wang, Zeming Li, Jian Sun, and Jiaya Jia. Fully convolutional networks for panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 214–223, 2021. 3, 4

[18] 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

[19] Stephan R Richter, Zeeshan Hayder, and Vladlen Koltun. Playing for benchmarks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2213–2222, 2017. 1

[20] 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 conference on computer vision and pattern recognition*, pages 3234–3243, 2016. 1

[21] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126(9):973–992, 2018. 1

[22] 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. 3

[23] Yang Zhang, Philip David, and Boqing Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2020–2030, 2017. 3

[24] Yiheng Zhang, Zhaofan Qiu, Ting Yao, Chong-Wah Ngo, Dong Liu, and Tao Mei. Transferring and regularizing prediction for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9621–9630, 2020. 3