MC-PanDA: Mask Confidence for Panoptic Domain Adaptation

Ivan Martinović[1], Josip Šarić[1], and Siniša Šegvić[1]

University of Zagreb, Faculty of Electrical Engineering and Computing, Croatia
{ivan.martinovic,josip.saric,sinisa.segvic}@fer.hr

Abstract. Domain adaptive panoptic segmentation promises to resolve the long tail of corner cases in natural scene understanding. Previous state of the art addresses this problem with cross-task consistency, careful system-level optimization and heuristic improvement of teacher predictions. In contrast, we propose to build upon remarkable capability of mask transformers to estimate their own prediction uncertainty. Our method avoids noise amplification by leveraging fine-grained confidence of panoptic teacher predictions. In particular, we modulate the loss with mask-wide confidence and discourage back-propagation in pixels with uncertain teacher or confident student. Experimental evaluation on standard benchmarks reveals a substantial contribution of the proposed selection techniques. We report 47.4 PQ on Synthia→Cityscapes, which corresponds to an improvement of 6.2 percentage points over the state of the art. The source code is available at `github.com/helen1c/MC-PanDA`.

Keywords: Unsupervised Domain Adaptation · Panoptic Segmentation · Consistency Learning · Mean Teacher · Uncertainty Quantification

1 Introduction

Panoptic segmentation is a recently developed technique that aims to unify all computer vision tasks related to scene understanding [6, 12, 27]. Consistent performance improvement has brought exciting applications within reach of practitioners [11,31,56]. However, many important domains require application-specific datasets that are very expensive to annotate [58, 68]. Furthermore, even fairly standard domains suffer from the long tail of corner cases that are not easily collected [44, 51, 57, 59]. Synthetic datasets offer a promising avenue to address the shortage of training labels, provided that domain shift is somehow accounted for [25, 41, 42]. Reducing the sensitivity to domain shift appears as a reasonable stance, since there will always be some divergence between the train and the test data [4, 5, 29, 64]. This state of affairs makes unsupervised domain adaptation an extremely attractive goal [2, 16, 47].

Most of the previous work in dense domain adaptation addresses semantic segmentation [8, 20, 23, 62, 65]. However, semantic segmentation methods are not easily upgraded to panoptic segmentation since they can not distinguish instances of the same class. This incompatibilty contributes to the relative scarceness of domain adaptive panoptic approaches in the literature [24, 43, 61].

Many approaches to discriminative learning on unlabeled images encourage smooth decision boundaries by consistency training. In practice, this boils down to learning invariance over a family of perturbations [30] by requiring similar predictions in related training samples [7]. The most effective methods involve strong perturbations [15] and disallow loss propagation from the perturbed samples towards the clean ones [18, 36]. This leads to self-learning setups with two branches, each of which holds an instance of the desired production model. The student branch receives the perturbed image and propagates the gradients to model parameters. The teacher branch receives the corresponding clean image, and stops the gradients. Best results are often obtained in the Mean Teacher setup where the teacher corresponds to the exponentially averaged student [48].

Consistency learning powers the most successful domain adaptive approaches for semantic segmentation [8,20,23]. Previous work in domain adaptive panoptics also start from this setup. CVRN [24] relies on cross-task and cross-style terms. UniDAformer [61] improves the teacher predictions through hierarchical mask calibration. EDAPS [43] outperforms all previous approaches by a wide margin through powerful shared encoder, separate decoders for things and stuff, as well as modulating the self-supervised loss with image-wide confidence.

This paper empowers domain adaptive panoptics with fine-grained uncertainty quantification. We denote our method MC-PanDA since we implement it atop a mask transformer [11] that localizes instances and stuff segments with dense activation maps called masks. Figure 1 shows that the teacher recovers mask-wide confidences λ_i and panoptic predictions. Moreover, we recover

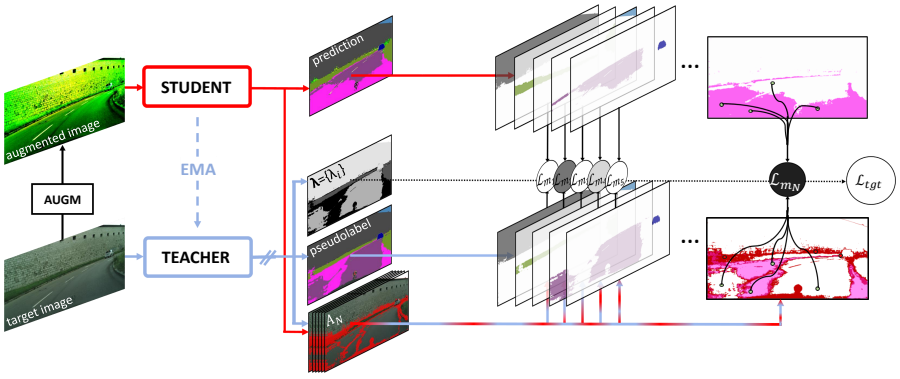

Fig. 1: MC-PanDA complements the Mean Teacher for domain-adaptive panoptics with fine-grained uncertainty quantification. We modulate the dense localization loss of the i-th mask $\mathcal{L}_{m_i}$ with the mask-wide confidence λ_i, and sample it with respect to pixel-level affinity A_i that blends teacher confidence Φ and student uncertainty (*cf*. Eq. 5). Our contributions encourage cautious self-learning from uncertain pseudo-labels. Note that we show detailed loss sampling and weighting only for the mask N. This procedure is carried out for each student mask that maps to a non-empty teacher mask. For simplicity, the figure omits source domain branch and SegMix augmentation.

the dense sampling affinity A_i for each mask i by blending fine-grained teacher confidence Φ with student uncertainty. Mask-wide confidences λ_i and sampling affinities A_i allow us to discourage self-learning in uncertain masks and at locations with inappropriate pixel-level uncertainty. In summary, we contribute two novel techniques for domain adaptive panoptics: i) mask-wide loss scaling (MLS) according to aggregated region-wide confidence λ_i, and ii) confidence-based point filtering (CBPF) that favours learning in points with confident teacher and uncertain student as indicated by A_i. Experiments on standard benchmarks for domain adaptive panoptics reveal substantial improvements of the generalization performance. Our method outperforms the state of the art by a large margin and brings us a step closer to real-world applications of synthetic training data.

2 Related Work

Panoptic Segmentation. Although many real-world applications require both semantic and instance segmentation, most of the early research considered these tasks in isolation. Recent definition of panoptic segmentation as the joint task brought much attention to this problem [27]. Early attempts extend Mask R-CNN [19] with a semantic segmentation branch and consider different fusion strategies for things and stuff predictions [26, 27, 32, 55]. Panoptic Deeplab extends the semantic segmentation pipeline with class-agnostic instance predictions based on center and offset regression [10, 46].

Unified Panoptics with Mask Transformers. Recent work proposes a unified framework that localizes instances and stuff segments with dense sigmoidal maps called *masks* [11,31,56]. Masks are recovered by scoring dense features with the associated embeddings [12]. Mask embeddings are recovered through direct set prediction with an appropriate transformer module [6]. Besides the simplified inference, these models currently achieve the state-of-the-art on panoptic segmentation benchmarks [33, 63]. Moreover, several recent works point out remarkable capability of mask transformers to estimate their own prediction uncertainty [17]. Methods based on mask transformers represent the current state of the art in dense anomaly detection [1, 14, 37, 40].

Unsupervised Domain Adaptation. This field involves a source domain and the target domain. The source domain contains labeled data that involves a domain shift with respect to the target domain. Thus, standard training on the source domain fails to bring satisfactory generalization on target data [44, 57]. Domain adaptation aims to improve the performance by joint training with unlabeled target data, which results in the following compound loss [2, 47]:

$$\mathcal{L}_{uda} = \mathcal{L}_{src} + \mathcal{L}_{tgt} \tag{1}$$

Domain adaptation focuses on the second loss term to find a way to exploit unlabeled target data [39,52,66,69]. The most performant approaches build upon consistency learning with Mean Teacher as described in the introduction [23,43]. Several approaches [62, 65] for domain-adaptive semantic segmentation leverage pixel-level uncertainty. Different from them, our method aggregates region-wide

uncertainties, exploits both the student uncertainty and the teacher uncertainty, and relies on sparse point sampling [28] instead of per-pixel loss scaling [62, 65].

Domain Adaptive Panoptics. The current state of the art is achieved by EDAPS [43], which extends consistency learning with image-wide scaling of the self-supervised loss [21, 22, 50]. This method encourages stable learning by reducing the gradient magnitude in images with uncertain teacher predictions. However, we find this suboptimal because it equally downgrades all image pixels, even when the prediction uncertainty varies across the image. This will certainly be the case when some kind of mixing data augmentation is used, where training images are crafted by pasting source scenery into target images [15, 50].

Our method is most closely related to UniDAformer [60], the only prior work that considers domain adaptation of a mask transformer. However, their work relies on handcrafted improvement of teacher predictions, while we present the first panoptic adaptation approach that relies upon region-wide and fine-grained uncertainty. Recent strong performance of mask transformers in dense outlier detection [17, 37, 40] suggests that our approach has a considerable chance of success. Contrary, UniDAformer underperforms with mask transformers and reports the best performance with a multi-branch architecture (*cf*. Table 1 and [60]).

3 Method

We first recap panoptic segmentation with direct set prediction in subsection 3.1. Then, we describe a baseline domain adaptation with a mask transformer in 3.2. Finally, subsections 3.3 and 3.4 present our contributions based on per-mask loss modulation and loss subsampling according to pixel-level confidence.

3.1 Panoptic Segmentation with Direct Set Prediction

Recent methods for scene understanding [6, 11, 31, 56] can handle semantic, instance or panoptic segmentation without changing the loss function or the model architecture. These models directly detect instances and stuff segments, and describe them with distinct segmentation masks. The dense feature extractor produces pixel embeddings $E_p \in \mathbb{R}^{H \times W \times d}$, where H and W stand for height and width. The transformer decoder observes the features and classifies each of the N mask queries across $(C+1)$ classes. The $(C+1)$-th class (no-object) indicates that the corresponding mask is unused in this particular image. The transformer decoder also produces mask embeddings $E_m \in \mathbb{R}^{N \times d}$ that identify the corresponding pixel embeddings E_p through dot-product similarity. Combining the two embeddings through generalized matmul and sigmoid activation produces pixel-to-mask assignments $\sigma \in \mathbb{R}^{N \times H \times W}$. Thus, each mask is defined with $\sigma_i \in \mathbb{R}^{H \times W}$ and class distribution $P_i = (p_i^{(1)}, p_i^{(2)}, ..., p_i^{(C+1)})$.

The training process minimizes the difference between the ground truth masks $\{(\sigma_i^{GT}, y_i^{GT})\}_i^{N^{GT}}$ and the predictions $\{(\sigma_i, P_i)\}_i^N$. The loss computation requires bipartite matching $\mathcal{M}$ which maps predictions onto their ground truth counterparts. Note that the set of ground truth masks has to be extended with

empty masks ($\sigma_i^{GT} = \mathbf{0}, y_i^{GT} = C+1$) in order to match the number of predictions. Given $\mathcal{M}$, we can decompose the loss into recognition and localization:

$$\mathcal{L}^{MT} = \sum_{i}^{N} \mathcal{L}_{cls}(P_i, y^{GT}_{\mathcal{M}(i)}) + \sum_{y^{GT}_{\mathcal{M}(i)} \neq C+1} \mathcal{L}_{mask}(\sigma_i, \sigma^{GT}_{\mathcal{M}(i)}) \tag{2}$$

The recognition terms $\mathcal{L}_{cls}$ require correct semantic classification, while the localization terms $\mathcal{L}_{mask}$ optimize per-pixel assignments. Note that $\mathcal{L}_{mask}$ is computed only for masks matched with non-empty ground truth, and that $\mathcal{L}^{MT}$ from (2) corresponds to $\mathcal{L}_{src}$ from (1). During inference, each pixel (row, column) is assigned the mask M that maximizes the pixel-level confidence ρ that is expressed as a product of recognition and localization scores:

$$M(r,c) = \text{argmax}_i(\rho_{i,r,c}); \quad \rho_{i,r,c} = \max_{y \neq C+1} P_i(y) \cdot \sigma_i(r,c) \tag{3}$$

3.2 Baseline Consistency Learning with Mean Teacher

This section presents our baseline Mean Teacher setup for panoptic self-training with mask transformers. We feed the teacher with clean images and block the gradients through the corresponding branch. We perturb the student images with color augmentation [9], random application of Gaussian smoothing, and SegMix - our panoptic adaptation of ClassMix [39,50]. Instead of classes, SegMix samples half of the panoptic segments from a random source image and pastes them atop the target image. We recover the teacher predictions and convert them to hard pseudo-labels consisting of the semantic class and dense assignment map. We obtain the student loss $\mathcal{L}_{tgt}$ from the mask transformer loss $\mathcal{L}^{MT}$ by replacing the ground truth with the pseudo-labels, *cf*. equations (1) and (2). We initialize the teacher with pretrained weights and keep them frozen in the early stage of domain adaptive training [3]. After that, we set the teacher to the temporal exponential moving average (EMA) of the student [49]. Nevertheless, our baseline students still tend to deteriorate due to noisy pseudo-labels. Moreover, we notice that our baseline teachers produce many false positive masks. Training on such pseudo-labels tends to amplify the noise due to Mean Teacher setup. The following subsections present our solution to this problem.

3.3 Mask-wide Loss Scaling

Our baseline underperforms due to positive feedback loop between the student and the teacher. We propose to alleviate this effect by modulating the per-mask localization term $\mathcal{L}_{mask}$ of the student loss $\mathcal{L}_{tgt}$ with mask-wide teacher confidence λ_i. This allows the student to learn from the confident teacher masks while postponing the learning from the uncertain ones. We scale our localization loss as follows:

$$\mathcal{L}^{\text{MC}}_{\text{loc}} = \sum_{y^{\text{teach}}_{\mathcal{M}(i)} \neq C+1} \lambda_i \cdot \mathcal{L}_{mask}(\sigma_i, \sigma^{\text{teach}}_{\mathcal{M}(i)}); \quad \lambda_i = \frac{\sum_{r,c \in M_i^F} [\![\rho_{i,r,c} > \tau_1]\!]}{|M_i^F|} \tag{4}$$

Note that M_i^F represents the foreground locations for mask i, $[\![.]\!]$ — Iverson indicator function, and τ_1 a threshold. The pixel-level confidence ρ is defined in (3). The mask-wide teacher confidence λ_i corresponds to the ratio of the foreground pixels where the pixel-level confidence is larger than the threshold.

3.4 Confidence-based Point Filtering

Training dense prediction models on high-resolution images can be extremely memory intensive. This problem can be alleviated by training on a carefully chosen sample of N_p dense predictions instead of on the whole prediction tensor [28]. The procedure starts by sampling a random oversized set of $3N_p$ floating-point locations. We subsample the initial oversized set by choosing $\beta \cdot N_p$ points with the largest uncertainty ($\beta \in [0, 1]$), and random $(1 - \beta) \cdot N_p$ points. However, the original procedure is inappropriate for consistency training. In fact, the teacher and the student will often be uncertain at the same locations since they are presented with the same image (up to a perturbation). Thus, blind favoring of uncertain student points would increase the chance of sampling incorrect pseudo-labels. On the other hand, favouring highly confident points would impair the learning process by providing uninformative gradients. Thus, we propose to favour points with low student and high teacher confidence by means of a dense sampling affinity A_i:

$$\mathrm{A}_i\,(r, c) = \begin{cases} -\infty & : \Phi_{r,c}^{\text{teach}} < \tau_2 \\ -|s_{i,r,c}| & : \text{otherwise} \end{cases}. \tag{5}$$

Note that $\Phi_{r,c}^{\text{teach}}$ denotes the teacher confidence that will be defined before the end of this subsection, while $s_{i,r,c}$ denotes the pre-activation of the student mask-assignment σ. Thus, if the teacher confidence in some point (r,c) is lower than the threshold τ_2, we prevent its sampling by setting the sampling affinity to $-\infty$. If the teacher is confident, we set the sampling affinity as the negative absolute value of the student pixel-to-mask pre-activation. We estimate the teacher confidence as follows:

$$\Phi_{r,c}^{\text{teach}} = \max_i \rho_{i,r,c}^{\text{teach}} \tag{6}$$

This score assigns the lowest confidence to the locations that are not claimed by any of the masks, and to the locations claimed by masks with low classification confidence. This formulation of point sampling is conservative as it prevents any mask from training on low confidence teacher predictions.

Figure 2 illustrates the teacher's computational graph given the pixel-to-mask assignment scores σ and classification probabilities P. We observe that the proposed additions to the baseline consistency represent small computational overhead over the standard panoptic inference.

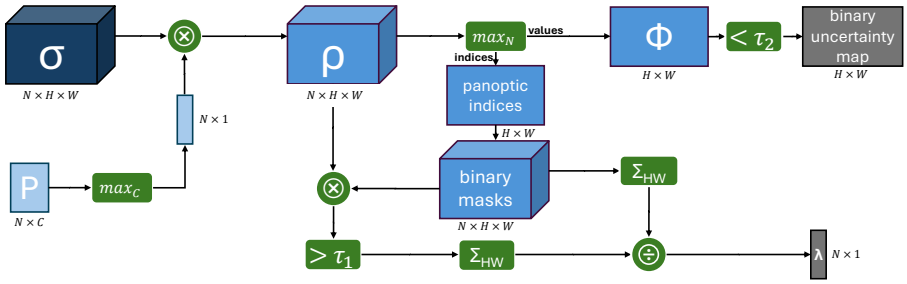

Fig. 2: The teacher multiplies the dense assignment masks σ with mask-wide max-softmax in order to recover dense per-mask confidence ρ. Taking the max along the mask axis of ρ and thresholding with respect to τ_2 reveals binary map of uncertain pixels (5). Furthermore, the mask axis arg-max of ρ delivers the map of panoptic indices, which we further convert to per-mask binary maps by converting indices to their one-hot encodings. Finally, we determine per-mask teacher confidence $\boldsymbol{\lambda} = \{\lambda_i\}$ (4) as the relative count of dense per-mask confidences that are greater than τ_1.

4 Experiments

4.1 Implementation Details

Datasets: We experiment on two real-world datasets: Cityscapes [13] and Mapillary Vistas [38]. Cityscapes comprises 2975 training and 500 validation images of European urban scenes in fair weather and 1024×2048 resolution. Vistas comprises 18,000 training and 2000 validation images of worldwide scenes in varying weather conditions and mean resolution of 8.4MPix. For synthetic data, we consider Synthia [42] with 9400 images at 1280×760 resolution, and Foggy Cityscapes [45] with 2975 training and 500 validation images of 1024×2048 pixels, and attenuation factor $\beta = 0.02$.
Architecture: Our experiments involve panoptic Mask2Former [11] with Swin-B [34] backbone, and multi-scale deformable attention [67] in the pixel decoder. We use the MiT-B5 backbone [54] in comparisons with EDAPS [43].
Training: We conduct the training in two stages [3]. The first stage initializes the backbone with ImageNet weights and pre-trains the teacher on the source domain. The second stage performs the domain adaptation training on unlabeled target domain and labeled source domain. We pre-train the teacher on Synthia/Cityscapes for 20k/40k iterations and freeze the teacher for initial 15k/20k iterations of the second stage [3]. Subsequently, we set the teacher weights to the exponential moving average [49] of the student with the decay factor $\alpha = 0.999$. We train for 90k iterations on batches of two source domain and two target domain images. We oversample the rare classes in the source domain [53], and apply random scaling, horizontal flipping, color augmentation, and random 512×1024 cropping in both domains. We apply additional strong augmentations to the student image as described in section 3.2. We recover target domain pseudo-labels through default panoptic inference [11] with the teacher, and train the student

with consistency and fine-grained confidence. We set the initial learning rate to 0.0001, weight decay to 0.05, and use AdamW [35]. We set $\tau_1 = 0.99$ and $\tau_2 = 0.8$ except on Synthia→Vistas where we report experiments for three values of (τ_1, τ_2). All experiments are conducted on a single A100-40GB.

Evaluation: We evaluate our models according to panoptic quality (PQ) [27] that can be factored into segmentation quality (SQ) and recognition quality (RQ). We report PQ for each category, as well as the mean PQ. Synthia [42] comprises 16 annotated classes that correspond to a subset of the standard Cityscapes taxonomy. Consequently, the experiments on Synthia report PQ_{16} as the mean over the 16 Synthia classes.

4.2 Comparison with the State-of-the-Art

Table 1 compares MC-PanDA with the domain-adaptive panoptic segmentation methods from the literature in two setups: Synthia→Cityscapes (left) and Synthia→Vistas (right). Our method performs well across all metrics in both setups. Our Cityscapes performance (47.4 PQ_{16}) outperforms the previous state of the art for 6.2 percentage points. On Synthia→Vistas, our model achieves 37.9 $\text{PQ}_{16}$with the default hyperparameters. We had noticed a deterioration of validation performance for a single training run. Hence, we have performed another experiment with a stricter loss subsampling threshold $\tau_2 = 0.9$, which resulted in 38.7 PQ_{16}. We have also tried to account for greater variety of stuff classes by setting $\tau_1^{\text{stuff}} = 0.9$. This experiment yielded 39.6 PQ_{16}. In the end, we have included the median of the three compound assays in the table. Note that all reported performances are averaged over three training runs.

Table 1: Performance evaluation on Synthia→Cityscapes and Synthia→Vistas. Two-branch UniDAFormer [61] (UniDAF-PSN) works better than the mask transformer counterpart (UniDAF-DETR). All our experiments are averaged over 3 random seeds.

Method	Synthia→City			Synthia→Vistas		
	SQ_{16}	RQ_{16}	PQ_{16}	SQ_{16}	RQ_{16}	PQ_{16}
CVRN [24]	66.6	40.9	32.1	65.3	28.1	21.3
UniDAF-DETR [61]	64.7	42.2	33.0	–	–	–
UniDAF-PSN [27,61]	66.9	44.3	34.2	–	–	–
EDAPS [43]	72.7	53.6	41.2	71.7	46.1	36.6
MC-PanDA (ours)	**76.7**	**59.3**	**47.4**	71.0	**49.8**	**38.7**

Figure 3 shows qualitative results on two Cityscapes scenes (rows 1-2) and a single Vistas scene (row 3). The columns show the input image, ground truth, predictions of the current SOTA [43], and our predictions. In the first scene, both models fail to recognize terrain due to the absence of this class in the source Synthia dataset. Besides that, we observe that our model delivers much better

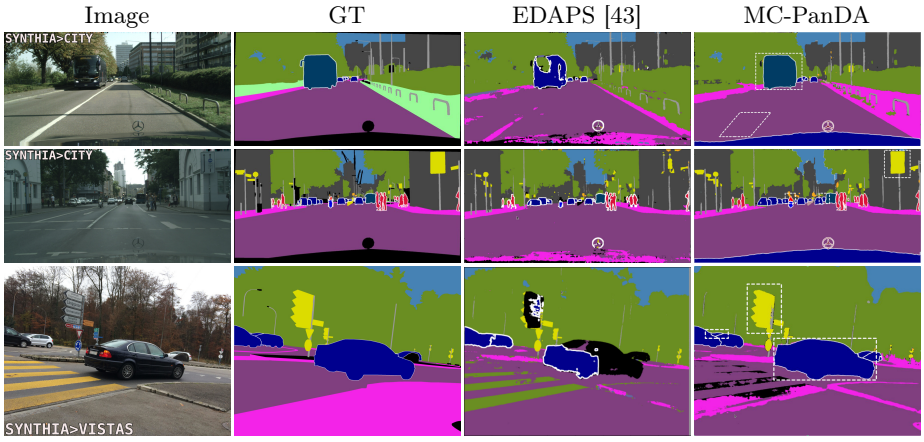

Fig. 3: Qualitative comparison with the state-of-the-art [43] on Synthia→Cityscapes and Synthia→Vistas. Dashed polygons indicate regions where our method prevails.

recognition of the road surface. Moreover, it correctly localizes and recognizes the bus while EDAPS fails in both tasks. In the second scene, our model prevails on the traffic sign and rider. In the third scene, our method delivers significantly better segmentation of the car and the traffic light. Both models fail on the crosswalk: EDAPS classifies it as vegetation, while our model rejects prediction in some pixels and predicts sidewalk in others.

Table 2 compares MC-PanDA with the methods from the literature when we use Cityscapes as the source dataset, while Foggy Cityscapes and Vistas serve as targets. Note that in the latter case, both source and target domain are based on real images. Following the previous work, we report PQ_{16} averaged over 16 Synthia classes, as well as PQ_{19} averaged over 19 Cityscapes classes. Our models outperform previous work by a large margin and set a new state of the art. We observe a substantial performance improvement on Vistas in comparison to the Synthia experiments. This reveals a significant space for improvement in bridging the domain shift between the synthetic and real domains.

Table 2: Performance evaluation on Cityscapes→Foggy Cityscapes and Cityscapes → Vistas. All our experiments are averaged over three random seeds.

	City→Foggy				City→Vistas			
Method	SQ_{16}	RQ_{16}	PQ_{16}	PQ_{19}	SQ_{16}	RQ_{16}	PQ_{16}	PQ_{19}
CVRN [24]	72.7	46.7	35.7	–	73.8	42.8	33.5	–
UniDAF [61]	72.9	49.5	37.6	–	–	–	–	–
EDAPS [43]	79.2	70.5	56.7	–	75.9	53.4	41.2	–
MC-PanDA	**82.5**	**76.3**	**63.7**	62.0	**79.3**	**66.6**	**53.8**	51.7

Table 3 complements the previous two tables with per-class evaluations of panoptic quality. In experiments with Synthia as the source domain, our method achieves the best performance on most classes and competitive peformance elsewhere. We observe that fence represents the hardest class in these setups, while sky and road are the easiest. When we use Cityscapes as the source dataset, our method prevails on all classes, except on rider when evaluating on Foggy Cityscapes. We observe significantly higher panoptic quality for class fence than in domain adaptation from Synthia. This suggests a large domain shift between fences in Synthia and the two real datasets.

Table 3: Per-class performance evaluation on four domain adaptation setups. All our experiments are averaged over three random seeds.

Method																	PQ_{16}
	Synthia→Cityscapes																
CVRN [24]	86.6	33.8	74.6	3.4	0.0	10.0	5.7	13.5	80.3	76.3	26.0	18.0	34.1	37.4	7.3	6.2	32.1
UniDAF [61]	73.7	26.5	71.9	1.0	0.0	7.6	9.9	12.4	81.4	77.4	27.4	23.1	47.0	40.9	12.6	15.4	33.0
UniDAF† [61]	**87.7**	34.0	73.2	1.3	0.0	8.1	9.9	6.7	78.2	74.0	37.6	25.3	40.7	37.4	15.0	18.8	34.2
EDAPS [43]	77.5	36.9	80.1	**17.2**	**1.8**	29.2	**33.5**	40.9	82.6	80.4	43.5	33.8	45.6	35.6	18.0	2.8	41.2
MC-PanDA	87.2	**51.8**	**82.5**	16.1	1.7	**36.3**	26.1	**54.3**	**86.3**	**86.4**	**48.3**	**37.7**	**46.9**	**45.8**	**27.4**	**23.9**	**47.4**
	Synthia→Vistas																
CVRN [24]	33.4	7.4	32.9	1.6	0.0	4.3	0.4	6.5	50.8	76.8	30.6	15.2	44.8	18.8	7.9	9.5	21.3
EDAPS [43]	77.5	25.3	59.9	**14.9**	0.0	27.5	33.1	37.1	**72.6**	**92.2**	32.9	**16.4**	47.5	**31.4**	13.9	3.7	36.6
MC-PanDA	**82.7**	**26.5**	**61.0**	5.5	0.0	**39.9**	**34.4**	**51.3**	62.1	85.6	**41.9**	11.4	**50.6**	25.1	**23.0**	**18.1**	**38.7**
	Cityscapes→Foggy Cityscapes																
CVRN [24]	93.6	52.3	65.3	7.5	15.9	5.2	7.4	22.3	57.8	48.7	32.9	30.9	49.6	38.9	18.0	25.2	35.7
UniDAF [61]	93.9	53.1	63.9	8.7	14.0	3.8	10.0	26.0	53.5	49.6	38.0	35.4	57.5	44.2	28.9	29.8	37.6
EDAPS [43]	91.0	68.5	80.9	24.1	29.0	50.1	47.2	67.0	85.3	71.8	50.9	**51.2**	64.7	47.7	36.9	41.5	56.7
MC-PanDA	**98.0**	**80.6**	**85.8**	**45.7**	**43.4**	**60.3**	**49.4**	**73.8**	**87.9**	**81.7**	**53.5**	47.8	**65.2**	**61.6**	**40.2**	**44.3**	**63.7**
	Cityscapes→Vistas																
CVRN [24]	77.3	21.0	47.8	10.5	13.4	7.5	14.1	25.1	62.1	86.4	37.7	20.4	55.0	21.7	14.3	21.4	33.5
EDAPS [43]	58.8	43.4	57.1	25.6	29.1	34.3	35.5	41.2	77.8	59.1	35.0	23.8	56.7	36.0	24.3	25.5	41.2
MC-PanDA	**88.4**	**49.1**	**75.2**	**35.2**	**39.7**	**50.3**	**45.2**	**54.0**	**81.1**	**96.2**	**46.1**	**30.1**	**57.2**	**42.0**	**33.9**	**37.3**	**53.8**

4.3 Ablations and Validations

This section validates the performance impact of our contributions on Synthia → Cityscapes. For all domain adaptation experiments we report mean and standard deviation over three runs. Please find additional ablations in the supplement, including a detailed comparison with the consistency baseline, the impact of τ_1 and τ_2 on the final performance, and ablations on Synthia→Vistas.

Table 4 quantifies the contributions of mask-wide loss scaling (MLS) and confidence-based point filtering (CBPF) to the panoptic performance of adapted models. We report results for MLS with two different thresholds: $\tau_1 = 0.968$ [43, 50] and $\tau_1 = 0.99$. The first row shows the domain generalization performance of a supervised model trained only on Synthia images. We notice that our consistency baseline from section 3.2 achieves the improvement of 8.8 points. Furthermore, self-training with MLS brings additional improvement of around 5 points,

which is almost 14 PQ points better than the domain generalization baseline. Confidence-based point filtering (CBPF) contributes a further improvement of around 3 PQ points. Finally, our complete method achieves 47.4 PQ points which represents an improvement of 16.6 PQ points w.r.t. the supervised baseline and 7.8 PQ points w.r.t. baseline consistency training.

Table 4: Ablation study on Synthia→Cityscapes: baseline consistency (BC), mask-level loss scaling (MLS), and confidence-based point filtering (CBPF). Top row represents supervised training on Synthia. We report $\text{mean}_{\pm\text{std}}$ over three random seeds.

BC	$\text{MLS}_{.968}$	$\text{MLS}_{.99}$	$\text{CBPF}_{0.8}$	SQ_{16}	↴	RQ_{16}	↴	PQ_{16}	↴
–	–	–	–	70.7	↴	40.4	↴	30.8	↴
✓	–	–	–	$73.4_{\pm 0.6}$	+2.7	$50.0_{\pm 1.7}$	+9.6	$39.6_{\pm 1.5}$	+8.8
✓	✓	–	–	$75.1_{\pm 0.5}$	+4.4	$56.1_{\pm 1.8}$	+15.7	$44.4_{\pm 1.5}$	+13.6
✓	–	✓	–	$75.3_{\pm 0.2}$	+4.6	$56.5_{\pm 0.9}$	+16.1	$44.6_{\pm 0.7}$	+13.8
✓	–	–	✓	$75.8_{\pm 0.4}$	+5.1	$55.7_{\pm 0.2}$	+15.3	$44.1_{\pm 0.2}$	+13.3
✓	✓	–	✓	$76.5_{\pm 0.0}$	+5.8	$59.2_{\pm 1.0}$	+18.8	$47.2_{\pm 0.5}$	+16.4
✓	–	✓	✓	$\mathbf{76.7}_{\pm 0.4}$	**+6.0**	$\mathbf{59.3}_{\pm 1.0}$	**+18.9**	$\mathbf{47.4}_{\pm 0.8}$	**+16.6**

Figure 4 illustrates the difference in pseudo-label quality of our consistency baseline (top) and MC-PanDA (bottom) at three training checkpoints. We observe that the pseudo-label quality improves as the training progresses. Moreover, we notice that the baseline starts hallucinating false positive segments, which significantly deteriorates panoptic quality. We believe this is due to the learning on noisy pseudo-labels in a positive feedback loop. On the other hand, our method reduces those hallucinations by downscaling the gradients of low-confidence masks and avoiding pixels with uncertain mask assignment. Note that both methods suppress very small masks during teacher inference.

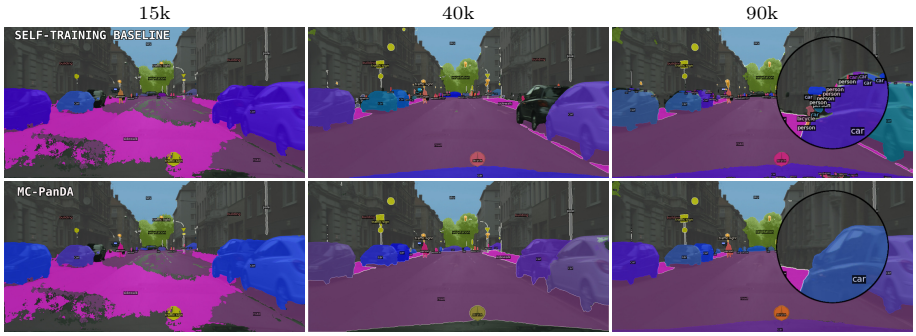

Fig. 4: Panoptic predictions in different training iterations. Top row presents our consistency baseline, while the bottom row presents our approach. Best viewed zoomed in.

We proceed by investigating alternative formulations of our proposed contributions. Table 5 compares our mask-wide loss scaling (MLS) with image-wide loss scaling (ILS) that assigns each mask with the same image-wide scaling factor [43]. We formalize the image-wide loss weight as follows:

$$\lambda_i = \lambda^{\text{ILS}} = \frac{\sum_{r,c} [\![\Phi^{\text{teach}}_{r,c} > \tau_{\text{ILS}}]\!]}{HW}. \tag{7}$$

We experiment with three different τ_{ILS} thresholds $\{0.95, 0.968, 0.99\}$ and report the performance of the best choice ($\tau_{\text{ILS}} = 0.968$) as an optimistic baseline. Top section shows results of ILS and MLS without confidence-based point filtering (CBPF), while the bottom section includes CBPF. We observe a clear advantage of mask-wide loss weighting in both cases. We believe this happens since ILS down-scales the loss gradients for many valid masks. Table 6 validates the proposed loss subsampling according to CBPF. The baseline point-sampling [28] favors points with high student uncertainty, which increases chances of sampling incorrect pseudo labels. CBPF reduces this effect by preventing training in points with low teacher confidence as presented in Figure 2. We also experiment with $\Phi^{\text{teach}}_{i,r,c} = \sigma(|s^{\text{teach}}_i|)_{r,c}$ as an alternative formulation of the teacher confidence, which considers only the mask that corresponds to the actual loss component. This experiment also considers completely random point samples (random sampling, $\beta = 0.0$). We observe that our formulation (all-masks) outperforms the alternative (per-mask) by 2.7 PQ points. We believe that per-mask underperformance occurs due to over-confident negative mask assignments where many pixels get incorrectly predicted as not belonging to the corresponding mask.

Table 5: Comparison of image-wide (ILS) and mask-wide (MLS) loss scaling on Synthia→Cityscapes. We report an optimistic ILS performance as a maximum across three different τ_{ILS} thresholds.

ILS	MLS	CBPF	SQ_{16}	RQ_{16}	PQ_{16}
✓	–	–	$73.7_{\pm 0.5}$	$50.5_{\pm 1.1}$	$39.7_{\pm 0.8}$
–	✓	–	$75.3_{\pm 0.2}$	$56.5_{\pm 1.0}$	$44.7_{\pm 0.7}$
✓	–	✓	$74.8_{\pm 0.4}$	$53.4_{\pm 0.2}$	$42.1_{\pm 0.5}$
–	✓	✓	$76.7_{\pm 0.4}$	$59.3_{\pm 1.0}$	$47.4_{\pm 0.8}$

Table 6: Validation of loss subsampling with confidence-based point filtering (CBPF) on Synthia→Cityscapes. The random sampling policy applies the loss across a random set of image points.

CBPF ($\tau_2 = 0.8$) filtering method	SQ_{16}	RQ_{16}	PQ_{16}
random sampling	$74.9_{\pm 0.1}$	$56.1_{\pm 1.4}$	$44.2_{\pm 1.1}$
per-mask	$75.2_{\pm 0.4}$	$56.5_{\pm 0.1}$	$44.7_{\pm 0.1}$
all-masks (eq. 6)	$76.7_{\pm 0.4}$	$59.3_{\pm 1.0}$	$47.4_{\pm 0.8}$

Table 7 compares our method with the current state-of-the-art [43] in experiments with matching backbones. We retrain EDAPS with our Swin-B backbone [34] and default hyperparameters, as well as MC-PanDA with the MiT-B5 backbone [54]. We observe 3.5 PQ points advantage of our method with MiT-B5 backbone, and 8.1 PQ points with Swin-B. Note that the number of parameters of both models are roughly the same.

Table 7: Backbone validation on Synthia→Cityscapes and comparison with EDAPS [43] using the same backbone. We report mean$_{\pm\text{std}}$ over three random seeds. † denotes our training with the publicly available source code.

Method	Backbone	#Params	SQ_{16}	RQ_{16}	PQ_{16}
EDAPS [43]	MiT-B5	104.9M	$72.7_{\pm 0.2}$	$53.6_{\pm 0.5}$	$41.2_{\pm 0.4}$
EDAPS†	Swin-B	110.8M	$73.1_{\pm 0.4}$	$51.4_{\pm 1.6}$	$39.3_{\pm 1.5}$
MC-PanDA	MiT-B5	103.2M	$75.1_{\pm 0.2}$	$56.5_{\pm 0.3}$	$44.7_{\pm 0.2}$
MC-PanDA	Swin-B	107.0M	$76.7_{\pm 0.4}$	$59.3_{\pm 1.0}$	$47.4_{\pm 0.8}$

4.4 Qualitative Analysis

Figure 5 illustrates advantages of our method on two qualitative examples. The figure presents two masks at three different training checkpoints and the corresponding mask-wide loss-scaling factors λ_i. We observe a strong positive correlation between the mask quality and the scaling factor, which suggests that our technique indeed suppresses gradients in low-quality masks. This behavior is especially important during early training when many road and traffic-sign pixels get incorrectly excluded from the corresponding masks.

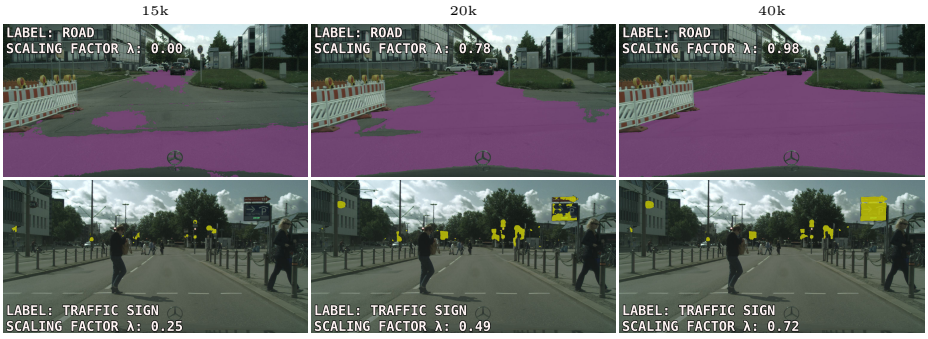

Fig. 5: We visualize the masks corresponding to road and traffic signs, and their MLS factors at three checkpoints. There is a strong correlation between the MLS factor λ and visual quality.

Figure 6 shows that mask-wide loss-scaling and confidence-based point filtering exhibit complementary contributions to the selection of confident points for domain-adaptive learning. We observe that both the person mask (red, left) and the motorcycle mask (blue, right) fail to attract all corresponding pixels. We also observe that these false negative regions coincide with the teacher's low confidence points (red, middle). The figure shows that our two techniques prevent the student model to learn from incorrect pseudo-labels.

Fig. 6: Mask-wide loss-scaling and confidence-based point filtering exhibit complementary contributions to the selection of pixels for domain-adaptive learning. The central image highlights the two regions where back-propagation is withheld due to low teacher confidence. The other two images show that these two regions correspond to false negative pixels in the person mask (left) and the motorcycle mask (right).

5 Limitations

Experiments on Synthia→Vistas show that our method may require dataset-specific hyper parameters in order to deliver the full extent of its generalization potential. Our experiments have been performed with batches of 2+2 images in order to fit on a single GPU with 40GB RAM. We consider that as a limitation since we imagine that many researchers do not have access to such equipment. We suspect that better performance would be obtained with larger batches, especially when training on datasets with high intra-class variation such as Vistas, since the recommended Mask2Former batch size is 16 [11].

6 Conclusion

All unsupervised domain adaptation methods strive to avoid random solutions due to noise amplification in the positive feedback. Our method addresses that goal in the frame of a siamese consistency with hard predictions by relying on fine-grained estimates of panoptic confidence. In particular, we have proposed to weight the dense self-supervised loss with mask-wide confidence, and subsample it with confidence-based point filtering. This improves the training convergence on target-domain images by discouraging self-learning at uncertain image locations. Extensive experimental evaluation reveals consistent advantage with respect to the published state of the art. In spite of encouraging experimental performance, important challenges still remain. We consider automatic and per-class selection of hyper-parameters as prominent avenues for future work.

Acknowledgments

This work has been supported by Croatian Recovery and Resilience Fund - NextGenerationEU (grant C1.4 R5-I2.01.0001), Croatian Science Foundation (grant IP-2020-02-5851 ADEPT), and the Advanced computing service provided by the University of Zagreb University Computing Centre - SRCE.

References

1. Ackermann, J., Sakaridis, C., Yu, F.: Maskomaly: Zero-shot mask anomaly segmentation. In: 34th British Machine Vision Conference 2022, BMVC 2022, Aberdeen, UK, November 20-24, 2023. p. 329. BMVA Press (2023)
2. Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.W.: A theory of learning from different domains. Mach. Learn. **79**(1-2), 151–175 (2010)
3. Berrada, T., Couprie, C., Alahari, K., Verbeek, J.: Guided distillation for semi-supervised instance segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 475–483 (2024)
4. Bevandic, P., Orsic, M., Grubisic, I., Saric, J., Segvic, S.: Weakly supervised training of universal visual concepts for multi-domain semantic segmentation. Int. J. Comput. Vis. (2024)
5. Blum, H., Sarlin, P., Nieto, J.I., Siegwart, R., Cadena, C.: The fishyscapes benchmark: Measuring blind spots in semantic segmentation. Int. J. Comput. Vis. **129**(11), 3119–3135 (2021)
6. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
7. Chapelle, O., Schölkopf, B., Zien, A. (eds.): Semi-Supervised Learning. The MIT Press (2006)
8. Chen, M., Zheng, Z., Yang, Y., Chua, T.: Pipa: Pixel- and patch-wise self-supervised learning for domain adaptative semantic segmentation. In: El-Saddik, A., Mei, T., Cucchiara, R., Bertini, M., Vallejo, D.P.T., Atrey, P.K., Hossain, M.S. (eds.) Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023. pp. 1905–1914. ACM (2023)
9. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
10. Cheng, B., Collins, M.D., Zhu, Y., Liu, T., Huang, T.S., Adam, H., Chen, L.C.: Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12475–12485 (2020)
11. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1290–1299 (2022)
12. Cheng, B., Schwing, A., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation. Advances in Neural Information Processing Systems **34**, 17864–17875 (2021)
13. 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: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
14. Delic, A., Grcic, M., Segvic, S.: Outlier detection by ensembling uncertainty with negative objectness. CoRR **abs/2402.15374** (2024)
15. French, G., Laine, S., Aila, T., Mackiewicz, M., Finlayson, G.D.: Semi-supervised semantic segmentation needs strong, varied perturbations. In: 31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020. BMVA Press (2020)

16. Ganin, Y., Lempitsky, V.S.: Unsupervised domain adaptation by backpropagation. In: Bach, F.R., Blei, D.M. (eds.) Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015. vol. 37, pp. 1180–1189 (2015)
17. Grcic, M., Saric, J., Segvic, S.: On advantages of mask-level recognition for outlier-aware segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023 - Workshops, Vancouver, BC, Canada, June 17-24, 2023. pp. 2937–2947. IEEE (2023)
18. Grubisic, I., Orsic, M., Segvic, S.: A baseline for semi-supervised learning of efficient semantic segmentation models (2021)
19. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
20. Hoyer, L., Dai, D., Gool, L.V.: Domain adaptive and generalizable network architectures and training strategies for semantic image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **46**(1), 220–235 (2024)
21. Hoyer, L., Dai, D., Van Gool, L.: Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9924–9935 (2022)
22. Hoyer, L., Dai, D., Van Gool, L.: Hrda: Context-aware high-resolution domain-adaptive semantic segmentation. In: European Conference on Computer Vision. pp. 372–391. Springer (2022)
23. Hoyer, L., Dai, D., Wang, H., Gool, L.V.: MIC: masked image consistency for context-enhanced domain adaptation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. pp. 11721–11732. IEEE (2023)
24. Huang, J., Guan, D., Xiao, A., Lu, S.: Cross-view regularization for domain adaptive panoptic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10133–10144 (2021)
25. Kim, D., Woo, S., Lee, J., Kweon, I.S.: Video panoptic segmentation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020. pp. 9856–9865. Computer Vision Foundation / IEEE (2020)
26. Kirillov, A., Girshick, R., He, K., Dollár, P.: Panoptic feature pyramid networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6399–6408 (2019)
27. Kirillov, A., He, K., Girshick, R., Rother, C., Dollár, P.: Panoptic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9404–9413 (2019)
28. Kirillov, A., Wu, Y., He, K., Girshick, R.: Pointrend: Image segmentation as rendering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9799–9808 (2020)
29. Lambert, J., Liu, Z., Sener, O., Hays, J., Koltun, V.: Mseg: A composite dataset for multi-domain semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **45**(1), 796–810 (2023)
30. Lenc, K., Vedaldi, A.: Understanding image representations by measuring their equivariance and equivalence. Int. J. Comput. Vis. **127**(5), 456–476 (2019)
31. Li, F., Zhang, H., Xu, H., Liu, S., Zhang, L., Ni, L.M., Shum, H.Y.: Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3041–3050 (2023)

32. Li, J., Raventos, A., Bhargava, A., Tagawa, T., Gaidon, A.: Learning to fuse things and stuff. arXiv preprint arXiv:1812.01192 (2018)
33. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
34. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
35. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2018)
36. Miyato, T., Maeda, S., Koyama, M., Ishii, S.: Virtual adversarial training: A regularization method for supervised and semi-supervised learning. IEEE Trans. Pattern Anal. Mach. Intell. **41**(8), 1979–1993 (2019)
37. Nayal, N., Yavuz, M., Henriques, J.F., Güney, F.: Rba: Segmenting unknown regions rejected by all. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023. pp. 711–722. IEEE (2023), https://doi.org/10.1109/ICCV51070.2023.00072
38. Neuhold, G., Ollmann, T., Rota Bulo, S., Kontschieder, P.: The mapillary vistas dataset for semantic understanding of street scenes. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
39. Olsson, V., Tranheden, W., Pinto, J., Svensson, L.: Classmix: Segmentation-based data augmentation for semi-supervised learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1369–1378 (2021)
40. Rai, S.N., Cermelli, F., Fontanel, D., Masone, C., Caputo, B.: Unmasking anomalies in road-scene segmentation. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023. pp. 4014–4023. IEEE (2023)
41. Richter, S.R., Vineet, V., Roth, S., Koltun, V.: Playing for data: Ground truth from computer games. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II. Lecture Notes in Computer Science, vol. 9906, pp. 102–118. Springer (2016)
42. 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: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
43. Saha, S., Hoyer, L., Obukhov, A., Dai, D., Van Gool, L.: Edaps: Enhanced domain-adaptive panoptic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 19234–19245 (October 2023)
44. Sakaridis, C., Dai, D., Gool, L.V.: ACDC: the adverse conditions dataset with correspondences for semantic driving scene understanding. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. pp. 10745–10755. IEEE (2021)
45. Sakaridis, C., Dai, D., Van Gool, L.: Semantic foggy scene understanding with synthetic data. International Journal of Computer Vision **126**, 973–992 (2018)
46. Saric, J., Orsic, M., Segvic, S.: Panoptic swiftnet: Pyramidal fusion for real-time panoptic segmentation. Remote. Sens. **15**(8), 1968 (2023)
47. Sun, B., Saenko, K.: Deep CORAL: correlation alignment for deep domain adaptation. In: Hua, G., Jégou, H. (eds.) Computer Vision - ECCV 2016 Workshops

- Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III. Lecture Notes in Computer Science, vol. 9915, pp. 443–450 (2016)
48. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. pp. 1195–1204 (2017)
49. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems **30** (2017)
50. Tranheden, W., Olsson, V., Pinto, J., Svensson, L.: Dacs: Domain adaptation via cross-domain mixed sampling. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1379–1389 (2021)
51. Uijlings, J.R.R., Mensink, T., Ferrari, V.: The missing link: Finding label relations across datasets. In: Avidan, S., Brostow, G.J., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VIII. Lecture Notes in Computer Science, vol. 13668, pp. 540–556. Springer (2022)
52. Vu, T.H., Jain, H., Bucher, M., Cord, M., Pérez, P.: Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2517–2526 (2019)
53. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2. `https://github.com/facebookresearch/detectron2` (2019)
54. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems. pp. 12077–12090 (2021)
55. Xiong, Y., Liao, R., Zhao, H., Hu, R., Bai, M., Yumer, E., Urtasun, R.: Upsnet: A unified panoptic segmentation network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8818–8826 (2019)
56. Yu, Q., Wang, H., Qiao, S., Collins, M., Zhu, Y., Adam, H., Yuille, A., Chen, L.C.: k-means mask transformer. In: European Conference on Computer Vision. pp. 288–307. Springer (2022)
57. Zendel, O., Honauer, K., Murschitz, M., Steininger, D., Domínguez, G.F.: Wilddash - creating hazard-aware benchmarks. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VI. Lecture Notes in Computer Science, vol. 11210, pp. 407–421. Springer (2018)
58. Zendel, O., Murschitz, M., Zeilinger, M., Steininger, D., Abbasi, S., Beleznai, C.: Railsem19: A dataset for semantic rail scene understanding. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 32–40. Computer Vision Foundation / IEEE (2019)
59. Zendel, O., Schörghuber, M., Rainer, B., Murschitz, M., Beleznai, C.: Unifying panoptic segmentation for autonomous driving. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. pp. 21319–21328. IEEE (2022)
60. Zhang, J., Huang, J., Lu, S.: Hierarchical mask calibration for unified domain adaptive panoptic segmentation. arXiv preprint arXiv:2206.15083 (2022)

61. Zhang, J., Huang, J., Zhang, X., Lu, S.: Unidaformer: Unified domain adaptive panoptic segmentation transformer via hierarchical mask calibration. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. pp. 11227–11237. IEEE (2023)
62. Zheng, Z., Yang, Y.: Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. IJCV (2021)
63. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 633–641 (2017)
64. Zhou, K., Liu, Z., Qiao, Y., Xiang, T., Loy, C.C.: Domain generalization: A survey. IEEE Trans. Pattern Anal. Mach. Intell. **45**(4), 4396–4415 (2023)
65. Zhou, Q., Feng, Z., Gu, Q., Cheng, G., Lu, X., Shi, J., Ma, L.: Uncertainty-aware consistency regularization for cross-domain semantic segmentation. CVIU (2022)
66. 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 **33**(2), 804–817 (2022)
67. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. In: International Conference on Learning Representations (2020)
68. Zlateski, A., Jaroensri, R., Sharma, P., Durand, F.: On the importance of label quality for semantic segmentation. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. pp. 1479–1487. Computer Vision Foundation / IEEE Computer Society (2018)
69. Zou, Y., Yu, Z., Kumar, B., Wang, J.: Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In: Proceedings of the European conference on computer vision (ECCV). pp. 289–305 (2018)