Hierarchical Mask Calibration for Unified Domain Adaptive Panoptic Segmentation

Jingyi Zhang, Jiaxing Huang, Shijian Lu∗
{Jingyi.Zhang, Jiaxing.Huang, Shijian.Lu}@ntu.edu.sg
Nanyang Technological University

Abstract. Domain adaptive panoptic segmentation aims to mitigate data annotation challenge by leveraging off-the-shelf annotated data in one or multiple related source domains. However, existing studies employ two networks for instance segmentation and semantic segmentation separately which lead to a large amount of network parameters with complicated and computationally intensive training and inference processes. We design UniDAPS, a Unified Domain Adaptive Panoptic Segmentation network that is simple but capable of achieving domain adaptive instance segmentation and semantic segmentation simultaneously within a single network. UniDAPS introduces Hierarchical Mask Calibration (HMC) that rectifies the predicted pseudo masks, pseudo superpixels and pseudo pixels and performs network re-training via an online self-training process on the fly. It has three unique features: 1) it enables unified domain adaptive panoptic adaptation; 2) it mitigates false predictions and improves domain adaptive panoptic segmentation effectively; 3) it is end-to-end trainable with much less parameters and simpler training and inference pipeline. Extensive experiments over multiple public benchmarks show that UniDAPS achieves superior domain adaptive panoptic segmentation as compared with the state-of-the-art.

Keywords: Panoptic Segmentation, Unsupervised Domain Adaptation

1 Introduction

Panoptic segmentation [28] performs instance segmentation for things and semantic segmentation for stuff, which assigns each image pixel with a semantic category and a unique identity simultaneously. With the advance of deep neural networks [29,18,5,19,41,17], panoptic segmentation [28,31,27,53,4,7,32,52,9,8] has achieved very impressive performance under the supervision of large quantities of densely-annotated training data. However, collecting densely-annotated panoptic data is prohibitively laborious and time-consuming [11,35,10] which has become one major constraint along this line of research. One possible alternative is to leverage off-the-shelf labeled data from one or multiple related source

* denotes corresponding author.
Authors Suppressed Due to Excessive Length

Fig. 1. Existing domain adaptive panoptic segmentation [20] adapts things and stuff separately by using two independent networks and fuses the two network outputs to produce the final panoptic segmentation, leading to more network parameters as well as complicated and computationally intensive training and inference processes. Differently, UniDAPS employs a single unified network to jointly adapt things and stuff which involves much less parameters and simplifies the training and inference pipeline greatly.

domains. Nevertheless, the source-trained models often experience clear performance drop while applied to various target domains that usually have different data distributions as compared with the source domains [20].

Table 1. Panoptic segmentation with traditional multi-branch architecture [28] and recent unified architectures [32, 9, 4]: The Supervised Setup trains with the Cityscapes [10] and tests on the same dataset. The UDA Setup trains with the SYNTHIA [43] and tests on Cityscapes. It can be seen that the Performance Drops between the two learning setups come more from mRQ than from mSQ consistently across different architectures. In addition, such a phenomenon is more severe for unified architectures. This demonstrates a clear false prediction issue in unified domain adaptive panoptic segmentation as mRQ is computed with false positives and false negatives.

|                      | Multi-branch Architecture | Unified Architecture |
|----------------------|---------------------------|----------------------|
|                      | PSN [28]                  | Panoptic FCN [32]    |
|                      | mSQ mRQ mPQ               | MaskFormer [9]       |
|                      | mSQ mRQ mPQ               | DETR [4]             |
| Supervised Setup     | 75.5 60.2 47.7            | 79.7 73.1 59.6       |
|                      | 79.1 62.6 51.1            | 79.1 64.1 51.9       |
| Adaptation Setup     | 59.0 27.8 20.1            | 47.5 19.7 15.8       |
|                      | 56.6 19.2 16.2            | 56.4 21.8 18.3       |
| Performance Drop     | -16.5 -32.4 -27.6         | -32.2 -53.4 -43.8    |
|                      | -22.5 -43.4 -34.9         | -22.7 -42.3 -33.6    |

Domain adaptive panoptic segmentation can mitigate the inter-domain discrepancy by aligning one or multiple labeled source domains and an unlabeled target domain [20]. To the best of our knowledge, CVRN [20] is the only such work, which tackles domain adaptive panoptic segmentation challenges by exploiting the distinct natures of instance segmentation and semantic segmentation. Specifically, CVRN introduces cross-view regularization to guide the two segmentation tasks to complement and regularize each other and achieves very impressive performance. However, CVRN relies on a multi-branch segmentation architecture that implements two separate networks and a two-phase pipeline as illustrated in Fig. 1 (a). This sophisticated design directly doubles network
parameters, slows down the network, and hinders it from being end-to-end trainable. A unified panoptic adaptation network that can effectively handle the two segmentation tasks with a single network is more desired.

We design a unified domain adaptive panoptic segmentation (UniDAPS) network as shown in Fig. 1 (b). Our design is based on the observation that one essential challenge in unified panoptic adaptation comes from a severe false prediction issue. As shown in Table 1, most recent unified panoptic segmentation architectures [4,32,9] outperform traditional multi-branch ones [28] by large margins under the supervised setup while opposite results are observed under unsupervised domain adaptation setup. Such contradictory results are more severe for the recognition quality in mRQ. This shows that the panoptic quality drop mainly comes from False Positives (FP) and False Negatives (FN) as mRQ is attributed by all predictions (True Positives, False Negatives and False Positives) while the segmentation quality in mSQ is attributed by True Positives (TP) only.

The proposed UniDAPS addresses the false prediction issue by introducing a Hierarchical Mask Calibration (HMC) technique, which calibrates the predicted pseudo masks, pseudo superpixels as well as pseudo pixels. With the corrected masks, UniDAPS re-trains the network via an online self-training process on the fly. Specifically, HMC treats both things and stuff predictions as masks uniformly and corrects each predicted pseudo mask hierarchically in a coarse-to-fine manner, i.e., from mask level that calibrates the overall category of each mask to superpixel and pixel levels that calibrate the superpixel and pixels around the boundary of each mask (which are more susceptible to prediction errors).

UniDAPS has three unique features. First, it achieves unified panoptic adaptation by treating things and stuff as masks and adapting them uniformly. Second, it mitigates the severe false prediction issue effectively by calibrating the predicted pseudo masks iteratively and progressively. Third, it is end-to-end trainable with much less parameters and simpler training and inference pipeline. Besides, the proposed HMC introduces little extra computation overhead and could be used as a plug-in.

We summarize our contributions in three aspects. First, we propose UniDAPS, the first end-to-end unified domain adaptive panoptic segmentation network to the best our knowledge. Second, we design Hierarchical Mask Calibration with online self-training, which allows to calibrate the predicted pseudo masks on the fly during self-training. Third, extensive experiments over multiple public benchmarks show that UniDAPS achieves superior segmentation accuracy and efficiency as compared with the state-of-the-art.

2 Related Works

Panoptic Segmentation is a challenging task that assigns each image pixel with a semantic category and a unique identity. The pioneer work [28] employs two networks for instance segmentation and semantic segmentation separately, and then combines the outputs of the two segmentation networks to acquire
panoptic segmentation. The later studies [53, 31, 47, 32, 52, 9, 8] simplify the complex pipeline by unifying the segmentation of things and stuff within single network. For example, DETR [4] predicts boxes around both things and stuff classes, and makes a final panoptic prediction by adding an FPN-style segmentation head. Panoptic segmentation has achieved very impressive accuracy but requires a large amount of densely-annotated training data that are often laborious and time-consuming to collect. Domain adaptive panoptic segmentation (DAPS), which leverages off-the-shelf annotated data for mitigating the data annotation constraint, is instead largely neglected.

Unsupervised Domain Adaptation (UDA) aims to exploit labeled source-domain data to learn a well-performing model on unlabeled target-domain data. In recent years, it has been studied extensively for various computer vision tasks, including image classification [14, 47, 40, 45, 62, 49, 40, 37, 12], instance segmentation/detection [6, 23, 46, 54, 3, 30, 60, 16, 44] and semantic segmentation [58, 62, 55, 21, 22, 57, 33]. On the other hand, domain adaptive panoptic segmentation is largely neglected despite its great values in various visual tasks and practical applications. To the best of our knowledge, CVRN [20] is the only work, which exploits the distinct natures of instance segmentation and semantic segmentation and introduces cross-view regularization to guide the two tasks to complement and regularize each other for panoptic adaptation. However, CVRN achieves panoptic adaptation by using two separate adaptation networks for things and stuff respectively, which directly doubles network parameters, slows down the network, and hinders it from being end-to-end trainable. In contrast, our proposed UniDAPS greatly simplifies training and inference pipeline by unifying the adaptation of things and stuff in a single panoptic adaptation network.

Self-training is a mainstream unsupervised domain adaptation technique that re trains networks with pseudo-labeled target-domain data. Most existing self-training methods [61, 23, 24, 56, 26, 62, 51, 34, 57] involve an iterative retraining process for effective learning from pseudo-labeled data. In each training iteration, an offline pseudo label generation process is involved which predicts and selects pseudo labels according to their confidence. For example, [61] proposes class-balanced self-training (CBST) that globally selects the same proportion of predictions as pseudo labels for each category for overcoming class-imbalance issues. To sidestep the cumbersome multi-round and offline training process, several studies [38, 2] explore ‘online’ self-training for semantic segmentation by directly enforcing pixel-wise consistency of predictions from different data augmentations. Differently, the proposed UniDAPS focuses on the false prediction issue in unified domain adaptive panoptic segmentation. It achieves effective ‘online’ self-training with a Hierarchical Mask Calibration technique which allows pseudo label calibration and correction on the fly.
Fig. 2. Overview of proposed Unified Domain Adaptive Panoptic Segmentation network (UniDAPS): it involves two flows, i.e., a pseudo mask calibration flow that calibrates pseudo masks with momentum model $G^m$, and an unsupervised training flow that optimizes model $G$ with the generated pseudo masks. For pseudo mask calibration, we feed a given unlabeled target image $x_t$ into the momentum model $G^m$ to produce $\hat{y}_t$, which is further used to calibrate pseudo masks by HMC via a coarse-to-fine manner (i.e., from mask level to superpixel and pixel levels). For network optimization, we conduct simple augmentations (i.e., resize, crop and flip) for $x_t$ and its calibrated pseudo masks $\hat{y}_t'$, and then optimize model $G$ with self-training loss $L_{self}$.

3 Method

3.1 Task Definition

This work focuses on domain adaptive panoptic segmentation. The training data involves a labeled source domain $D_s = \{(x^i_s, y^i_s)\}_{i=1}^{N_s}$ ($y^i_s$ is the panoptic annotation of sample $x^i_s$) and an unlabeled target domain $D_t = \{x^i_t\}_{i=1}^{N_t}$. The goal is to learn a model $G$ from $D_s$ and $D_t$ that well performs in $D_t$. The baseline model is trained with the source domain data $D_s$ only:

$$L_{sup} = l(G(x_s)), y_s),$$

where $l(\cdot)$ denotes the panoptic segmentation loss that consists of a matching cost and a Hungarian loss [4].

3.2 UniDAPS Overview

This subsection presents the overall framework of proposed UniDAPS, which consists of a supervised training process over the labeled source domain and an unsupervised training process over the unlabeled target domain. For the supervised training, the source samples $(x_s, y_s)$ are fed to a panoptic segmentation model $G$ that is optimized via the supervised loss $L_{sup}$ as defined in Eq. 1.

The unsupervised training involves two flows as illustrated in Fig. 2. The first flow calibrates pseudo masks with the momentum model $G^m$ (the moving
averaged of $G$, i.e., $\theta_G^{m} \leftarrow \gamma \theta_G^{m} + (1 - \gamma)\theta_G$, and $\gamma$ is a momentum coefficient) via Hierarchical Mask Calibration (HMC) module. The second flow optimizes $G$ with the calibrated pseudo masks. Specifically, we first apply simple data augmentations (i.e., resize, crop and flip) to $x_t$ and $\hat{y}_t'$ to obtain $x_{\text{aug}}$ and $\hat{y}'_{\text{aug}}$. The network model $G$ is then optimized with the augmented data and the self-training loss $L_{\text{self}}$ as defined in Eq. 9.

### 3.3 Hierarchical Mask Calibration

One key component in the proposed UniDAPS is HMC that calibrates the predicted pseudo masks and enables effective pseudo-label retraining on the fly. HMC treats both things and stuff predictions as masks uniformly and corrects each predicted mask hierarchically in a coarse-to-fine manner. The correction involves three consecutive stages of calibration including a Mask-wise Calibration, a Superpixel-wise Calibration and a Pixel-wise Calibration as illustrated in Fig. 2. First, Mask-wise Calibration corrects the overall category of each mask by adaptively re-weighting its category-wise probabilities. Leveraging the feature that superpixels adhere well to the boundaries of things and stuff, Superpixel-wise Calibration then adjusts the shape of each mask by considering the boundary of the computed superpixels. Finally, Pixel-wise Calibration introduces pixel-level categorization information and further refines the boundary of each mask with a simple pixel-wise voting mechanism.

As the proposed calibration technique works for all predicted pseudo masks (i.e., things and stuff) uniformly, we take one pseudo mask $\hat{y}_t^k$ from $\hat{y}_t = \{\hat{y}_t^k\}_{k=1}^K$ as an example for illustration. Each pseudo mask $\hat{y}_t^k$ includes a predicted category $\hat{c} = \arg\max_c p^c$ ($p^c \in \{p^c\}_{c=1}^C$ is the probability of belonging to the $c$-th category) and a predicted binary mask $\hat{M}$ of size $H \times W$.

**Mask-wise Calibration** corrects the predicted category $\hat{c}$ by re-weighting its category-wise probability $p^c$ as following:

$$
c' = \arg\max_{c}(w^c \otimes p^c),
$$

where $\otimes$ denotes the element-wise multiplication and $w^c \in \{w^c\}_{c=1}^C$ is the calibration weight for the corresponding $c$-th category probability.

The calibration weight $w$ is calculated according to the distance between $\hat{y}_t^k$ and the mask centroids $\delta^c$ over feature space. Given the feature $f_t \in \mathbb{R}^{E \times H \times W}$ as generated by the momentum backbone, we pool the feature within the mask $\hat{M}$ into a mask-wise vector $v_m^k \in \mathbb{R}^E$ as follows:

$$
v_m^k = \text{GAP}(\hat{M} \otimes f_t),
$$

where GAP($\cdot$) denotes the global average pooling operation.

Generally, if the mask-wise vector $v_m^k$ is far from the $c$-th mask centroid $\delta^c$, the pseudo mask $\hat{y}_t^k$ should be assigned with a lower probability of belonging to the $c$-th category, and vice versa. Therefore, the calibration weight in Eq. 2 is defined as follows:

$$
w^c = \text{Softmax}(-||v_m^k - \delta^c||_1),
$$
where the distance is measured using L1 distance and softmax operation is performed along the category dimension.

Here we demonstrate how we compute and update the mask centroids along the training process. The mask centroids are first initialized by all target predictions from the baseline model. For each category, the mask centroid \( \delta^c \) is defined as follows:

\[
\delta^c = \frac{\sum_{x_t \in D_t} \sum_{k \in K} v^k_m \cdot 1(c^k = c)}{\sum_{x_t \in D_t} \sum_{k \in K} 1(c^k = c)},
\]

where \( 1 \) is an indicator function that returns ‘1’ if the vector \( v^k_m \) belongs to \( c \)-th category, and ‘0’ otherwise.

Along training process, we update the mask centroids with the current batch of data:

\[
\delta^c \leftarrow \gamma' \delta^c + (1 - \gamma') \delta^c_*,
\]

where \( \delta^c_* \) is the mask centroid calculated with the current data and model, and \( \gamma' \) is a update coefficient for smooth centroid update.

**Superpixel-wise Calibration:** Following mask-wise calibration, we first correct the shape of the pseudo mask by exploiting superpixels that adhere well to the boundaries of things and stuff \([1]\). To this end, we first compute a superpixel map \( M^I \) which includes total \( I \) superpixels \( M^{(i)} \) for target image \( x_t \). Then, we select the superpixels that overlap with the original mask \( \hat{M} \) to form an adjusted binary mask \( M_{sp} \) as follows:

\[
M_{sp} = \bigcap_{i \in I} M^{(i)} \cdot 1(A^{(i)} > 0),
\]
where 1 is an indicator function and we denote the overlapping area between $i$-th superpixel and the mask $\hat{M}$ as $A^{(i)}$.

The superpixel-based mask $M_{sp}$ adjusts the original mask $\hat{M}$ with the computed superpixels which adheres better to the edge of things or stuff, as illustrated in Figs. 3(b) and (c).

**Pixel-wise Calibration:** Following superpixel-wise calibration, we further introduce pixel-level categorization information and refine the boundary of the superpixel-based mask $M_{sp}$ in a more precise manner.

We design a simple pixel-wise voting mechanism to achieve the fine-grained calibration: the superpixels are discarded from the $M_{sp}$ if a majority of pixel-wise feature vectors within it are inconsistent with the overall category of pseudo mask $\hat{y}^k_t$ as illustrated in Fig. 3(d). Let $v^k_p$ denote a pixel-wise feature vector within superpixel $M^{(j)} \in M_{sp}$, and we determine that it is inconsistent with the pseudo mask if it does not satisfy the following condition:

$$\arg\min_c (||v^k_p - \delta^c||) = c', \quad (8)$$

where $c'$ is the corrected category of the pseudo mask $\hat{y}^k_t$. Such pixel-wise voting effectively suppresses the side effect of outlier pixels by enforcing that pixels within the same superpixel should share the same semantic category [1].

The final calibrated pseudo masks $\hat{y}'_t$ for target image $x_t$ can be acquired by repeating the above-mentioned processes for all the calibrated pseudo mask (i.e., $\hat{y}'_t = \{\hat{y}^k'_t\}^K_{k=1}$, where $\hat{y}^k'_t = \{c', M'\}$).

### 3.4 Network Optimization

With the calibrated pseudo masks $\hat{y}'_t$, the self-training loss $L_{self}$ can be formulated as follows:

$$L_{self} = l(G(x_{aug}), \hat{y}'_{aug}), \quad (9)$$

where $l(\cdot)$ denotes the panoptic segmentation loss that consists of a matching cost and a Hungarian loss [4]. $\hat{y}'_{aug}$ and $x_{aug}$ are the simple augmentations (i.e., resize, crop and flip) of $\hat{y}'_t$ and $x_t$, respectively.

The overall training objective is defined by minimizing the supervised loss $L_{sup}$ and the unsupervised loss $L_{self}$:

$$\arg\min_G L_{sup} + L_{self}. \quad (10)$$

### 4 Experiment

This section presents experiments including datasets, evaluation metric, ablation studies, comparisons with the state-of-the-art and discussions. Due to the space limit, the implementation details are provided in the appendix.
4.1 Datasets

We evaluate UniDAPS over three widely used domain adaptation tasks with four datasets:

1) SYNTHIA [43] → Cityscapes [10] which aims for domain adaptation from synthetic images to real-world images. The training set in SYNTHIA are adopted as source domain and the training set in Cityscapes are considered as target domain. The evaluation is performed on the validation set of Cityscapes.

2) Cityscapes [10] → Foggy Cityscapes [48] which aims for domain adaptation across different weather conditions, where Cityscapes is used as source domain and Foggy Cityscapes is considered as target domain. The adaptation performance is evaluated over the validation set of Foggy Cityscapes.

3) VIPER [42] → Cityscapes [10] which aims for domain adaptation from synthetic images to real-world images. We adopt the training set of VIPER as source domain and the training set in Cityscapes as target domain. The evaluation is performed on the validation set of Cityscapes.

4.2 Evaluation Metric

In evaluations, we adopt three panoptic segmentation metrics [28] including segmentation quality (SQ), recognition quality (RQ) and panoptic quality (PQ) as in [28,27,32,20]. For each category, PQ can be computed as the multiplication of the corresponding SQ term and RQ term as follows:

\[
PQ = \sum_{(p,g) \in TP} \frac{\text{IoU}(p,g)}{|TP|} \times \frac{|TP|}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|},
\]

where \(g\) is the ground truth segment and \(p\) is the matched prediction. TP, FP and FN denote true positives, false positives and false negatives, respectively. IoU is the insertion over union metric [13] which is widely used in semantic segmentation evaluations. With the above definitions, RQ captures the proportion of TP in all predictions, SQ captures the segmentation quality within TP while PQ integrates PQ and SQ and captures the overall panoptic segmentation quality.

4.3 Ablation Studies

The core of the proposed UniDAPS is Hierarchical Mask Calibration that consists of a Mask-wise Calibration, a Superpixel-wise Calibration and a Pixel-wise Calibration. We first study the three calibration modules to examine how they contribute to the overall domain adaptive panoptic segmentation.

Table 2 shows experimental results over task SYNTHIA → Cityscapes. It can be seen that the baseline in the first row (trained with the labeled source data only) does not perform well due to domain shifts. Including self-training over unlabeled target data in the 2nd row improves the baseline from 18.3 to 22.6 in
Table 2. Ablation study of the proposed Hierarchical Mask Calibration technique over task SYNTTHIA $\rightarrow$ Cityscapes, where ‘Cal.’ stands for Calibration

| Self-training | Mask-wise Cal. | Superpixel-wise Cal. | Pixel-wise Cal. | mSQ   | mRQ   | mPQ   |
|---------------|----------------|---------------------|----------------|-------|-------|-------|
|               | ✓              | ✓                   | ✓              | 56.4  | 21.8  | 18.3  |
| ✓             |               | ✓                   | ✓              | 59.5  | 29.9  | 22.6  |
| ✓             | ✓             |                     | ✓              | 61.2  | 36.9  | 28.7  |
| ✓             | ✓             | ✓                   |               | 63.0  | 32.4  | 26.2  |
| ✓             | ✓             | ✓                   |               | 62.6  | 31.8  | 24.2  |
| ✓             | ✓             | ✓                   |               | 63.4  | 39.9  | 30.9  |
| ✓             | ✓             | ✓                   |               | 63.2  | 38.9  | 30.1  |
| ✓             | ✓             | ✓                   |               | 64.3  | 32.7  | 26.9  |
| ✓             | ✓             | ✓                   |               | 64.7  | 42.2  | 33.0  |

Table 3. Experiments with unified panoptic segmentation architecture [4] over task SYNTTHIA $\rightarrow$ Cityscapes. PQ is computed for each category. Mean SQ (mSQ), mean RQ (mSQ), mean PQ (mPQ) are computed over all categories

| SYNTTHIA $\rightarrow$ Cityscapes Panoptic Segmentation |
|--------------------------------------------------------|
| Methods       | road side, build, wall, fence, pole, light, sign, vege, sky, pers, rider, car, bus, mot, bike | mSQ   | mRQ   | mPQ   |
| Baseline [2]  | 33.9 7.3 45.6 0.0 0.0 2.9 5.2 7.6 65.0 57.1 19.2 5.4 22.3 14.9 1.3 4.8 56.4 21.8 18.3 |
| DAF [6]       | 34.6 7.5 53.8 0.0 0.0 2.1 3.1 2.6 72.8 67.5 21.3 9.4 28.2 22.7 4.4 4.4 59.0 28.3 20.9 |
| FDA [55]      | 29.2 5.8 63.6 0.1 0.0 4.9 4.0 4.8 73.9 62.5 24.5 11.4 32.1 27.1 5.2 8.0 59.1 30.3 22.3 |
| CRST [62]     | 46.8 11.8 56.8 0.9 0.0 4.0 2.6 3.5 70.9 64.3 20.7 11.8 32.9 32.2 6.2 7.9 62.5 31.9 23.3 |
| SVMin [15]    | 48.0 11.8 56.9 0.5 0.0 3.9 3.8 4.4 72.5 68.9 26.4 15.0 35.3 25.8 6.3 7.6 63.3 32.6 24.2 |
| AdvEnt [51]   | 55.9 14.4 64.0 0.0 0.0 4.6 3.3 2.7 75.5 72.3 24.9 9.7 33.8 26.7 5.2 7.1 60.2 33.0 25.0 |
| CVRN [20]     | 66.2 19.4 72.5 2.1 0.0 3.8 6.5 4.4 79.7 75.1 26.5 11.5 36.6 34.1 7.1 8.2 61.4 35.9 27.9 |
| UniDAPS       | 73.7 26.5 71.9 1.0 0.0 7.6 9.9 12.4 81.4 77.4 27.4 23.1 47.0 40.9 12.6 15.4 64.7 42.2 33.0 |

mPQ. On top of the self-training, including any of the three calibration modules improves the segmentation consistently as shown in rows 3-5. Specifically, mask-wise calibration improves mRQ more (15.1 above the baseline) than the other two calibration modules (10.6 and 10.0), showing that mask-wise calibration suppresses false predictions effectively by calibrating the overall category of each mask. On the other hand, superpixel-wise and pixel-wise calibrations improve mSQ more than mask-wise calibration (6.6 and 6.2 vs 4.8), showing that superpixel-wise and pixel-wise calibrations focus on refining the boundary of each mask.

The three calibration modules correct pseudo masks from different levels which complement each in domain adaptive panoptic segmentation. We can observe that combining any two modules further improves mSQ, mRQ and mPQ consistently as shown in rows 6-8, and combining all three achieves the best mSQ, mRQ and mPQ. Such experimental results are well aligned with the motivation and design of the proposed hierarchical mask calibration.
Table 4. Experiments with unified panoptic segmentation architecture [4] over task Cityscapes → Foggy cityscapes. PQ is computed for each category. Mean SQ (mSQ), mean RQ (mRQ), mean PQ (mPQ) are computed over all categories

| Methods     | road side. | build. | wall | fence | pole | light | sign | vege. | sky | pers. | rider | car | bus | mot. | bike | mSQ   | mRQ   | mPQ   |
|-------------|------------|--------|------|-------|------|-------|------|-------|-----|-------|-------|-----|-----|------|------|-------|-------|-------|
| Baseline    | 92.5       | 48.9   | 60.6 | 6.0   | 10.7 | 5.3   | 9.9  | 23.6  | 55.6| 22.3  | 15.4  | 38.5| 23.7| 1.6  | 2.8  | 70.0  | 38.6  | 29.2  |
| DAF [6]     | 94.0       | 54.5   | 57.7 | 6.7   | 10.0 | 7.0   | 6.6  | 25.5  | 44.6| 26.7  | 16.7  | 42.2| 36.6| 4.5  | 16.9 | 70.6  | 41.7  | 31.8  |
| FDA [5]     | 93.8       | 53.1   | 62.2 | 8.2   | 13.4 | 7.3   | 7.6  | 28.9  | 50.8| 49.7  | 25.0  | 22.6| 42.9| 36.3 | 10.3 | 15.2  | 71.4  | 43.5  | 33.0  |
| AdvEnt [51] | 93.8       | 52.7   | 56.3 | 5.7   | 13.5 | 10.0  | 10.9 | 27.7  | 40.7| 57.9  | 27.8  | 29.4| 44.7| 28.6 | 11.6 | 20.8  | 72.3  | 43.7  | 33.3  |
| CRST [62]   | 91.8       | 49.7   | 66.1 | 6.4   | 14.5 | 5.2   | 8.6  | 21.5  | 56.3| 50.7  | 30.5  | 30.7| 46.3| 34.2 | 11.7 | 22.1  | 72.2  | 44.9  | 34.1  |
| SVMin [15]  | 93.4       | 53.4   | 62.2 | 12.3  | 15.9 | 7.0   | 8.5  | 18.0  | 54.3| 57.1  | 31.2  | 29.6| 45.2| 35.6 | 11.5 | 22.7  | 72.4  | 45.5  | 34.8  |
| UniDAPS     | 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  | 72.9  | 49.5  | 37.6  |

Table 5. Experiments with unified panoptic segmentation architecture [4] over task VIPER → Cityscapes. PQ is computed for each category. Mean SQ (mSQ), mean RQ (mRQ), mean PQ (mPQ) are computed over all categories

| Methods     | road side. | build. | fence | light | sign | vege. | sky | pers. | car | bus | mot. | bike | mSQ   | mRQ   | mPQ   |
|-------------|------------|--------|-------|-------|------|-------|-----|-------|-----|-----|------|------|-------|-------|-------|
| Baseline    | 25.2       | 5.7    | 35.1  | 0.0   | 5.9  | 3.9   | 75.3| 21.2  | 39.7| 21.4 | 11.4 | 0.0  | 59.7  | 32.0  | 24.1  |
| DAF [6]     | 56.6       | 7.3    | 41.0  | 0.0   | 3.5  | 2.7   | 76.4| 70.2  | 19.0 | 34.3| 14.2 | 6.3  | 0.0   | 61.1  | 33.3  | 25.5  |
| FDA [5]     | 50.0       | 7.6    | 59.4  | 0.0   | 6.2  | 6.1   | 73.3| 65.9  | 19.4 | 38.2| 15.5 | 8.1  | 0.0   | 61.0  | 35.2  | 26.9  |
| AdvEnt [51] | 52.6       | 10.8   | 51.0  | 0.0   | 2.0  | 4.8   | 73.9| 70.1  | 15.9 | 38.2| 19.9 | 12.4 | 0.0   | 61.2  | 35.4  | 27.0  |
| CRST [62]   | 68.7       | 9.1    | 54.4  | 0.0   | 2.4  | 2.7   | 76.3| 69.9  | 21.2 | 34.0| 21.9 | 7.7  | 0.0   | 61.0  | 36.5  | 28.3  |
| SVMin [15]  | 87.6       | 14.2   | 70.7  | 0.0   | 4.1  | 6.3   | 74.4| 70.0  | 16.9 | 32.5| 2.4  | 11.0 | 1.2   | 61.3  | 37.5  | 29.9  |
| CVRN [20]   | 75.1       | 18.8   | 59.9  | 0.0   | 9.1  | 6.5   | 76.8| 71.1  | 22.3 | 37.0| 15.5 | 8.6  | 3.8   | 66.4  | 40.2  | 31.1  |
| UniDAPS     | 87.1       | 22.1   | 71.1  | 0.0   | 8.2  | 8.6   | 78.3| 71.8  | 25.4| 46.8| 13.7 | 12.8 | 2.8   | 68.9  | 43.0  | 34.5  |

4.4 Comparisons with the State-of-the-art

Due to the lack of prior studies on unified domain adaptive panoptic segmentation, we conduct two sets of experiments to benchmark UniDAPS with the state-of-the-art.

In the first set of experiments, we benchmark UniDAPS over the unified panoptic segmentation architecture (i.e., DETR [4]) by reproducing the state-of-the-art [20] with DETR. Specifically, we re-implement the cross-style regularization (one of two cross-view designs) in DETR to reproduce CVRN (cross-task regularization relies on multi-branch architecture and cannot work in the unified architecture). Following [20], we also reproduce several domain adaptive methods by directly implementing their adaptation module in DETR. We perform comparisons over three domain adaptive panoptic segmentation tasks as shown in Tables 3 and 5. It can be seen that UniDAPS improves the baseline [4] by large margins (8.3, 20.4 and 14.7 in mSQ, mRQ and mPQ) and it also outperforms the state-of-the-art clearly for SYNTHIA → Cityscapes. In particular, UniDAPS improves more in mRQ as compared with the state-of-the-art, indicating that it corrects more false predictions effectively. Similar experimental results are observed on the other two tasks as shown in Tables 4 and 5.
Table 6. Experiments with multi-branch panoptic segmentation architecture \[28\] over task SYNTHIA → Cityscapes. Mean SQ (mSQ), mean RQ (mSQ), mean PQ (mPQ) are computed over all categories.

| Methods   | road | side. | build. | wall | fence | pole | light | sign | vege. | sky | pers. | rider | car | bus | mot. | bike | mSQ | mRQ | mPQ |
|-----------|------|-------|--------|------|-------|------|-------|------|-------|-----|-------|-------|----|-----|-----|------|-----|-----|-----|
| PSN \[28\] | 32.3 | 5.1  | 85.8  | 0.9  | 0.0  | 4.6  | 61.7  | 61.3 | 27.6  | 9.5 | 32.8  | 22.6  | 1.0 | 2.7 | 59.0| 27.8 | 20.1|
| FDA \[55\]  | 79.0 | 22.0 | 61.8  | 1.1  | 0.0  | 5.6  | 5.5   | 9.5  | 51.6  | 34.1| 31.0  | 5.2   | 8.8 | 65.0| 35.6| 26.6 |
| CRST \[62\]  | 75.4 | 19.0 | 70.8  | 1.4  | 0.0  | 7.3  | 0.7   | 3.3  | 71.7  | 69.2 | 23.7  | 16.3  | 34.1| 31.0| 6.9 | 26.6 |
| AdvEnt \[51\] | 87.1 | 32.4 | 69.7  | 1.1  | 0.0  | 3.8  | 0.7   | 2.3  | 71.7  | 72.0 | 28.2  | 21.1  | 6.3 | 4.9 | 65.6| 28.1 |
| CVRN \[20\]  | 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| 66.9| 44.3| 34.2 |
| UniDAPS    | 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| 66.9 |

Fig. 4. Prediction quality analysis over task SYNTHIA → Cityscapes.

In the second set of experiments, we benchmark UniDAPS over the multi-branch panoptic segmentation architecture (i.e., PSN \[28\]). Since HMC introduces little extra computation overhead and can be incorporated as a plug-in, we directly apply HMC (with the online self-training loss) on the multi-branch architecture for benchmarking. Table 6 shows experimental results on SYNTHIA → Cityscapes. We can see that UniDAPS outperforms CVRN in mSQ, mRQ and mPQ consistently. In addition, it similarly improves mRQ by large margins, which further verifies the motivation and design of the proposed HMC.

4.5 Discussions

Prediction Quality Analysis. UniDAPS suppresses false predictions effectively via HMC. We examine it over task SYNTHIA → Cityscapes with DETR \[4\]. As discussed in Section 4.2, the predictions in panoptic segmentation consists of three parts including TP, FP and FN. We compute the proportion of each part over all predictions and Fig. 4 shows experimental results. We can observe that UniDAPS produces clearly more TP and less FN and FP as compared with both baseline \[4\] and the state-of-the-art \[15,20\]. This demonstrates the superiority of UniDAPS in suppressing false predictions in domain adaptive panoptic segmentation.

Efficiency Comparison with CVRN \[20\]. Beyond segmentation accuracy, we also benchmark UniDAPS with multi-branch panoptic adaptation network CVRN \[20\] in parameter number, training speed and inference speed. As Table 7 shows, UniDAPS has clearly less parameters and its training and inference time...
Table 7. Efficiency comparison with multi-branch panoptic adaptation network CVRN [20] in terms of parameter number, training speed and inference speed

| Methods          | Architecture       | Parameter | Training Speed | Inference Speed |
|------------------|--------------------|-----------|----------------|-----------------|
| CVRN [20]        | Multi-branch Network | 185.58 M  | 0.27 fps       | 0.36 fps        |
| UniDAPS          | Unified Network    | 58.75 M   | 2.16 fps       | 7.24 fps        |

Fig. 5. Qualitative comparison of UniDAPS with the baseline model [4] and CVRN [20] over task SYNTHIA → Cityscapes. The proposed UniDAPS yields best segmentation with more true positives, less false predictions and finer boundary.

is much shorter than CVRN as well, demonstrating its great simplicity and efficiency.

Qualitative Results. Fig. 5 shows qualitative comparisons on SYNTHIA → Cityscapes. It can be observed that qualitative segmentation is well aligned with the quantitative results. Specifically, the baseline DETR [4] produces the worst segmentation and the state-of-the-art improves but tends to miss small things and produce false predictions. UniDAPS further improves and yields better segmentation with more TP and less false predictions.

5 Conclusion

This paper presents UniDAPS, a unified domain adaptive panoptic segmentation network. UniDAPS introduces a Hierarchical Mask Calibration (HMC) technique to calibrate the predicted pseudo masks on the fly during re-training. UniDAPS has three unique features: 1) it achieves unified panoptic adaptation by treating things and stuff as masks and adapting them uniformly; 2) it mitigates the severe false prediction issue effectively by calibrating the predicted pseudo masks iteratively and progressively; 3) it is end-to-end trainable with much less parameters and simpler training and inference pipeline. Extensive experiments over multiple public benchmarks show that UniDAPS achieves superior segmentation accuracy and efficiency as compared with the state-of-the-art. Moving forwards, we will continue to investigate innovative techniques for simple and effective domain adaptive panoptic segmentation.
References

1. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: Slic superpixels compared to state-of-the-art superpixel methods. IEEE transactions on pattern analysis and machine intelligence 34(11), 2274–2282 (2012)

2. Araslanov, N., Roth, S.: Self-supervised augmentation consistency for adapting semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15384–15394 (2021)

3. Cai, Q., Pan, Y., Ngo, C.W., Tian, X., Duan, L., Yao, T.: Exploring object relation in mean teacher for cross-domain detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11457–11466 (2019)

4. 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)

5. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: 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)

6. Chen, Y., Li, W., Sakaridis, C., Dai, D., Van Gool, L.: Domain adaptive faster r-cnn for object detection in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3339–3348 (2018)

7. 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)

8. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. arXiv preprint arXiv:2112.01527 (2021)

9. Cheng, B., Schwing, A., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation. Advances in Neural Information Processing Systems 34 (2021)

10. 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. pp. 3213–3223 (2016)

11. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)

12. Du, Z., Li, J., Su, H., Zhu, L., Lu, K.: Cross-domain gradient discrepancy minimization for unsupervised domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3937–3946 (2021)

13. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International journal of computer vision 111(1), 98–136 (2015)

14. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: International conference on machine learning. pp. 1180–1189. PMLR (2015)

15. Guan, D., Huang, J., Lu, S., Xiao, A.: Scale variance minimization for unsupervised domain adaptation in image segmentation. Pattern Recognition 112, 107764 (2021)

16. Guan, D., Huang, J., Xiao, A., Lu, S., Cao, Y.: Uncertainty-aware unsupervised domain adaptation in object detection. IEEE Transactions on Multimedia (2021)
17. 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)
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
19. Hoffman, J., Wang, D., Yu, F., Darrell, T.: Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. arXiv preprint arXiv:1612.02649 (2016)
20. 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)
21. Huang, J., Guan, D., Xiao, A., Lu, S.: Rda: Robust domain adaptation via fourier adversarial attacking. arXiv preprint arXiv:2106.02874 (2021)
22. Huang, J., Lu, S., Guan, D., Zhang, X.: Contextual-relation consistent domain adaptation for semantic segmentation. In: European Conference on Computer Vision. pp. 705–722. Springer (2020)
23. Inoue, N., Furuta, R., Yamasaki, T., Aizawa, K.: Cross-domain weakly-supervised object detection through progressive domain adaptation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5001–5009 (2018)
24. Iqbal, J., Ali, M.: MsIsI: Multi-level self-supervised learning for domain adaptation with spatially independent and semantically consistent labeling. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1864–1873 (2020)
25. Kim, M., Byun, H.: Learning texture invariant representation for domain adaptation of semantic segmentation. arXiv preprint arXiv:2003.00867 (2020)
26. Kim, S., Choi, J., Kim, T., Kim, C.: Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 6092–6101 (2019)
27. 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)
28. 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)
29. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems 25, 1097–1105 (2012)
30. Li, C., Du, D., Zhang, L., Wen, L., Luo, T., Wu, Y., Zhu, P.: Spatial attention pyramid network for unsupervised domain adaptation. In: European Conference on Computer Vision. pp. 481–497. Springer (2020)
31. Li, Y., Chen, X., Zhu, Z., Xie, L., Huang, G., Du, D., Wang, X.: Attention-guided unified network for panoptic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7026–7035 (2019)
32. Li, Y., Zhao, H., Qi, X., Wang, L., Li, Z., Sun, J., Jia, J.: Fully convolutional networks for panoptic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 214–223 (2021)
33. Li, Y., Yuan, L., Vasconcelos, N.: Bidirectional learning for domain adaptation of semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6936–6945 (2019)
34. Lian, Q., Lv, F., Duan, L., Gong, B.: Constructing self-motivated pyramid curriculums for cross-domain semantic segmentation: A non-adversarial approach. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6758–6767 (2019)
35. Lin, T.Y., Maire, M., Belongie, S., HaysPietro, J., Ramanan, P., DollárC, P., Zitnick, L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755 (2014)
36. Long, M., Zhu, H., Wang, J., Jordan, M.I.: Deep transfer learning with joint adaptation networks. In: International conference on machine learning. pp. 2208–2217. PMLR (2017)
37. Lu, Z., Yang, Y., Zhu, X., Liu, C., Song, Y.Z., Xiang, T.: Stochastic classifiers for unsupervised domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9111–9120 (2020)
38. Melas-Kyriazi, L., Manrai, A.K.: Pixmatch: Unsupervised domain adaptation via pixelwise consistency training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12435–12445 (2021)
39. Pan, F., Shin, I., Rameau, F., Lee, S., Kweon, I.S.: Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. arXiv preprint arXiv:2004.07703 (2020)
40. Pinheiro, P.O.: Unsupervised domain adaptation with similarity learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8004–8013 (2018)
41. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. arXiv preprint arXiv:1506.01497 (2015)
42. Richter, S.R., Hayder, Z., Koltun, V.: Playing for benchmarks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2213–2222 (2017)
43. 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. pp. 3234–3243 (2016)
44. Saha, S., Obukhov, A., Paudel, D.P., Kanakis, M., Chen, Y., Georgoulis, S., Van Gool, L.: Learning to relate depth and semantics for unsupervised domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8197–8207 (2021)
45. Saito, K., Kim, D., Sclaroff, S., Darrell, T., Saenko, K.: Semi-supervised domain adaptation via minimax entropy. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 8050–8058 (2019)
46. Saito, K., Ushiku, Y., Harada, T., Saenko, K.: Strong-weak distribution alignment for adaptive object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6956–6965 (2019)
47. Saito, K., Watanabe, K., Ushiku, Y., Harada, T.: Maximum classifier discrepancy for unsupervised domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3723–3732 (2018)
48. Sakaridis, C., Dai, D., Van Gool, L.: Semantic foggy scene understanding with synthetic data. International Journal of Computer Vision 126(9), 973–992 (2018)
49. Sankaranarayanan, S., Balaji, Y., Castillo, C.D., Chellappa, R.: Generate to adapt: Aligning domains using generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8503–8512 (2018)
50. Subhani, M.N., Ali, M.: Learning from scale-invariant examples for domain adaptation in semantic segmentation. arXiv preprint arXiv:2007.14449 (2020)
51. 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)

52. Wang, H., Zhu, Y., Adam, H., Yuille, A., Chen, L.C.: Max-deeplab: End-to-end panoptic segmentation with mask transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5463–5474 (2021)

53. 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)

54. Xu, C.D., Zhao, X.R., Jin, X., Wei, X.S.: Exploring categorical regularization for domain adaptive object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11724–11733 (2020)

55. Yang, Y., Soatto, S.: Fda: Fourier domain adaptation for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4085–4095 (2020)

56. Yu, F., Wang, D., Chen, Y., Karianakis, N., Yu, P., Lymberopoulos, D., Chen, X.: Unsupervised domain adaptation for object detection via cross-domain semi-supervised learning. arXiv preprint arXiv:1911.07158 (2019)

57. Zhang, P., Zhang, B., Zhang, T., Chen, D., Wang, Y., Wen, F.: Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12414–12424 (2021)

58. Zhang, Q., Zhang, J., Liu, W., Tao, D.: Category anchor-guided unsupervised domain adaptation for semantic segmentation. In: Advances in Neural Information Processing Systems. pp. 433–443 (2019)

59. Zhang, Y., David, P., Gong, B.: Curriculum domain adaptation for semantic segmentation of urban scenes. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2020–2030 (2017)

60. Zheng, Y., Huang, D., Liu, S., Wang, Y.: Cross-domain object detection through coarse-to-fine feature adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13766–13775 (2020)

61. 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)

62. Zou, Y., Yu, Z., Liu, X., Kumar, B., Wang, J.: Confidence regularized self-training. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5982–5991 (2019)