

Universal Domain Adaptation for Semantic Segmentation

Seun-An Choe¹

Keon-Hee Park¹

Jinwoo Choi¹

Gyeong-Moon Park^{2†}

¹Kyung Hee University, Yongin, Republic of Korea

²Korea University, Seoul, Republic of Korea

¹{dragoon0905, pgh2874, jinwoochoi}@khu.ac.kr ²gm-park@korea.ac.kr

Abstract

Unsupervised domain adaptation for semantic segmentation (UDA-SS) aims to transfer knowledge from labeled source data to unlabeled target data. However, traditional UDA-SS methods assume that category settings between source and target domains are known, which is unrealistic in real-world scenarios. This leads to performance degradation if private classes exist. To address this limitation, we propose Universal Domain Adaptation for Semantic Segmentation (UniDA-SS), achieving robust adaptation even without prior knowledge of category settings. We define the problem in the UniDA-SS scenario as low confidence scores of common classes in the target domain, which leads to confusion with private classes. To solve this problem, we propose UniMAP: UniDA-SS with Image Matching and Prototype-based Distinction, a novel framework composed of two key components. First, Domain-Specific Prototype-based Distinction (DSPD) divides each class into two domain-specific prototypes, enabling finer separation of domain-specific features and enhancing the identification of common classes across domains. Second, Target-based Image Matching (TIM) selects a source image containing the most common-class pixels based on the target pseudo-label and pairs it in a batch to promote effective learning of common classes. We also introduce a new UniDA-SS benchmark and demonstrate through various experiments that UniMAP significantly outperforms baselines. The code is available at <https://github.com/KU-VGI/UniMAP>.

1. Introduction

Semantic segmentation is a fundamental computer vision task that predicts the class of each pixel in an image

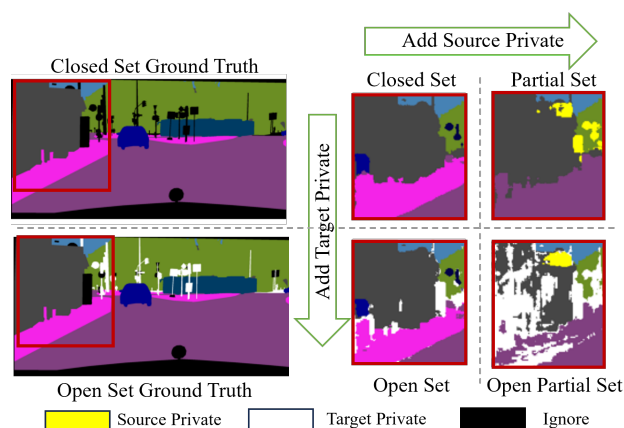


Figure 1. Visualization results of the UDA-SS models across different scenarios. We select MIC and BUS, which achieve the best performance in CDA-SS and ODA-SS, respectively, and visualize their results in PDA-SS and OPDA-SS. The images illustrate the performance degradation caused by the introduction of source-private classes.

and is essential in fields like autonomous driving, medical imaging, and human-machine interaction. However, training segmentation models requires pixel-level annotations, which are costly and time-consuming. To address this, researchers have explored Unsupervised Domain Adaptation for Semantic Segmentation (UDA-SS) methods, which aim to learn domain-invariant representations from labeled synthetic data (source) to unlabeled real-world data (target).

While UDA-SS has shown effectiveness in addressing domain shift, existing UDA-SS methods rely on the assumption that source and target categories are known in advance. This assumption is often impractical in real-world scenarios, as target labels are typically unavailable. As a result, the target domain frequently includes unseen classes that are absent in the source domain (target-private classes), or conversely, the source domain may contain classes not found in the target domain (source-private classes). This limitation can lead to negative transfer, where models incor-

[†]Corresponding author.

rectly align source-private classes with the target domain, resulting in significant performance degradation. To address these challenges, we introduce a new Universal Domain Adaptation for Semantic Segmentation (UniDA-SS) scenario, enabling adaptation without prior knowledge of category configurations and classifying target samples as “unknown” if they contain target-private classes.

To understand the challenges posed by the UniDA-SS scenario, we first evaluate the performance of existing UDA-SS methods under various domain adaptation settings. Figure 1 presents qualitative results of UDA-SS methods across various scenarios. Specifically, we select MIC [15] and BUS [6] as representative models for Closed Set Domain Adaptation (CDA-SS) and Open Set Domain Adaptation (ODA-SS), respectively, and analyze their performance in Partial Domain Adaptation (PDA-SS) and Open Partial Domain Adaptation (OPDA-SS) settings. CDA-SS assumes that the source and target domains share the same set of classes, while ODA-SS contains target-private classes that do not exist in the source domain. PDA-SS, on the other hand, assumes that the target domain contains only a subset of the source classes. OPDA-SS extends PDA-SS by adopting the open-set characteristic of ODA-SS, where both source-private and target-private classes exist simultaneously.

These evaluations reveal that adding source-private classes in transitions from CDA to PDA and from ODA to OPDA degrades performance. In PDA, common classes like “buildings” are often misclassified as source-private, while “sidewalk” regions are mistakenly predicted as “road”. Similarly, in OPDA, target-private regions are frequently confused with source-private or common classes. Most state-of-the-art UDA-SS methods depend on self-training with target pseudo-labels, heavily relying on pseudo-label confidence scores. Particularly, in ODA-SS scenarios such as BUS, confidence scores are also used to assign unknown pseudo-labels. When source-private classes are present, their feature similarity to some common classes increases, leading to a reduction in pseudo-label confidence. As a result, common classes may not be effectively learned, and if the confidence score drops below a certain threshold (τ_p), common classes are often misassigned as target-private classes. This misassignment hinders the effective learning of both common and target-private classes, further degrading adaptation performance.

To mitigate this issue, we propose a novel framework UniMAP, **Uni**DA-SS with **I**mage **M**atching and **P**rototype-based **D**istinction, aim to increase the confidence scores of common classes under unknown category settings. First, we introduce a Domain-Specific Prototype-based Distinction (DSPD) to distinguish between common classes and source-private classes while considering variations of common classes between the source and target domains. Un-

like conventional UDA-SS methods, which treat common classes as identical across domains, DSPD assigns two prototypes per class—one for source and one for target—to learn with one class while capturing domain-specific features. This approach enables independent learning of source and target-specific features, enhancing confidence scores for target predictions. Additionally, since common class pixel embeddings will have similar relative distances to the source and target prototypes, and the private class will be relatively close to any one prototype, we can use this to distinguish between common and private classes and assign higher weights to pixel embeddings that are more likely to belong to the common classes.

Second, to increase the confidence scores of the common classes, it is crucial to increase their pixel presence during training for robust domain-invariant representation. However, source-private classes often reduce the focus on common classes, hindering effective adaptation. To address this issue, we propose Target-based Image Matching (TIM), which prioritizes source images with the highest number of common class pixels based on target pseudo-labels. TIM compares target pseudo-labels and source ground truth at the pixel level, selecting the source images that overlap the most in common classes to pair with the target image in a single batch. This matching strategy facilitates domain-invariant learning of common classes, improving performance in a variety of scenarios. We also utilize a class-wise weighting strategy based on the target class distribution to assign higher weights to rare classes to address the class imbalance problem.

We summarize our main contributions as follows:

- We introduce a new task the Universal Domain Adaptation for Semantic Segmentation (UniDA-SS) task for the first time. To address this, we propose a novel framework called UniMAP, short for **Uni**DA-SS with **I**mage **M**atching and **P**rototype-based **D**istinction.
- To enhance pseudo-label confidence in the target domain, we propose Domain-Specific Prototype-based Distinction (DSPD), a pixel-level clustering approach that utilizes domain-specific prototypes to distinguish between common and private classes.
- We propose Target-based Image Matching (TIM), which enhances domain-invariant learning by prioritizing source images rich in common class pixels.
- We demonstrate the superiority of our approach by achieving state-of-the-art performance compared to existing UDA-SS methods through extensive experiments.

2. Related Work

2.1. Semantic Segmentation.

Semantic segmentation aims to classify each pixel in an image into a specific semantic. A foundational approach, Fully

Convolutional Networks (FCNs) [21], has demonstrated impressive performance in this task. To enhance contextual understanding, subsequent works have introduced methods such as dilated convolutions [4], global pooling [20], pyramid pooling [41], and attention mechanisms [42, 45]. More recently, transformer-based methods have achieved significant performance gains [37, 43]. Despite various studies, semantic segmentation models are still vulnerable to domain shifts or category shifts. To address this issue, we propose a universal domain adaptation for semantic segmentation that handles domain shifts and category shifts.

2.2. Unsupervised Domain Adaptation for Semantic Segmentation.

Unsupervised Domain Adaptation (UDA) aims to leverage labeled source data to achieve high performance on unlabeled target data. Existing UDA methods for semantic segmentation can be categorized into two approaches: adversarial learning-based and self-training. Adversarial learning-based methods [3, 9, 12, 16, 25, 32, 33] use an adversarial domain classifier to learn domain-invariant features. Self-training methods [5, 13, 14, 18, 19, 23, 31, 35, 36, 40, 46, 47] assign pseudo-labels to each pixel in the target domain using confidence thresholding, and several self-training approaches further enhance target domain performance by re-training the model with these pseudo-labels. Although UDA allows the model to be trained on the target domain without annotations, it requires prior knowledge of class overlap between the source and target domains, which limits the model’s applicability and generalizability. To overcome this limitation, we propose a universal domain adaptation approach for semantic segmentation, where the model can adapt to the target domain without requiring prior knowledge of class overlap.

2.3. Universal Domain Adaptation in Classification

Universal Domain Adaptation (UniDA) [38] was introduced to address various domain adaptation settings, such as closed-set, open-set, and partial domain adaptation. UniDA is a more challenging scenario because it operates without prior knowledge of the category configuration of the source and target domains. To tackle UniDA in classification tasks, prior works have focused on computing confidence scores for known classes and treating samples with lower scores as unknowns. CMU [10] proposed a thresholding function, while ROS [1] used the mean confidence score as a threshold, which results in neglecting about half of the target data as unknowns. DANCE [29] set a threshold based on the number of classes in the source domain. OVA Net [28] introduced training a threshold using source samples and adapting it to the target domain. While UniDA has been extensively studied in the context of classification tasks, it remains underexplored in semantic segmentation, which re-

quires a higher level of visual understanding due to the need for pixel-wise classification. In this work, we aim to investigate UniDA for semantic segmentation.

3. Method

3.1. Problem Formulation

In the UniDA-SS scenario, the goal is to transfer knowledge from a labeled source domain $D_s = \{X_s, Y_s\}$ to an unlabeled target domain $D_t = \{X_t\}$. The model is trained on the source images $X_s = \{x_s^1, x_s^2, \dots, x_s^{i_s}\}$ with the corresponding labels $Y_s = \{y_s^1, y_s^2, \dots, y_s^{i_s}\}$ and the target images $X_t = \{x_t^1, x_t^2, \dots, x_t^{i_t}\}$, where ground-truth labels are unavailable. Each image $x_s^{i_s} \in \mathbb{R}^{3 \times H \times W}$ and $y_s^{i_s} \in \mathbb{R}^{C_s \times H \times W}$ represent an i_s -th RGB image and its pixel-wise label. H and W are the height and width of the image, and C_s and C_t denote the sets of classes in the source and target domains, respectively. We aim to adapt the model to perform well on D_t , even though there is no prior knowledge of class overlap between C_s and C_t given. We define $C_c = C_s \cap C_t$ as the set of common classes, while $C_s \setminus C_c$ and $C_t \setminus C_c$ represent the classes private to each domain, respectively. To handle target-private samples in $C_t \setminus C_s$, we classify them as “unknown” without prior knowledge of their identities. Under this formulation, UniDA-SS requires addressing two challenges: (1) to classify common classes in C_c correctly and (2) to detect target-private classes in $C_t \setminus C_s$.

3.2. Baseline

We construct our UniDA-SS baseline by adopting a standard open-set self-training approach, partially following the ODA-SS formulation introduced in BUS [6]. BUS handles unknown target classes by appending an additional classification head node to predict unknown class. In our baseline, we adopt the same structural design as BUS but remove refinement components and the use of attached private class masks, resulting in a setup suitable for UniDA-SS.

In this baseline, the number of classifier heads is set to $(C_s + 1)$, where the $(C_s + 1)$ -th head corresponds to the unknown class. The segmentation network f_θ is trained with the labeled source data using the following categorical cross-entropy loss \mathcal{L}_{seg}^s :

$$\mathcal{L}_{seg}^s = - \sum_{j=1}^{H \cdot W} \sum_{c=1}^{C_s+1} y_s^{(j,c)} \log f_\theta(x_s)^{(j,c)}, \quad (1)$$

where $j \in \{1, 2, \dots, H \cdot W\}$ denotes the pixel index and $c \in \{1, 2, \dots, C_s + 1\}$ denotes the class index. The baseline utilizes a teacher network g_ϕ to generate the target pseudo-labels. g_ϕ is updated from f_θ via exponential moving average (EMA) [30] with a smoothing factor α . The pseudo-label $\hat{y}_{tp}^{(j)}$ for the j -th pixel considering unknown assigned

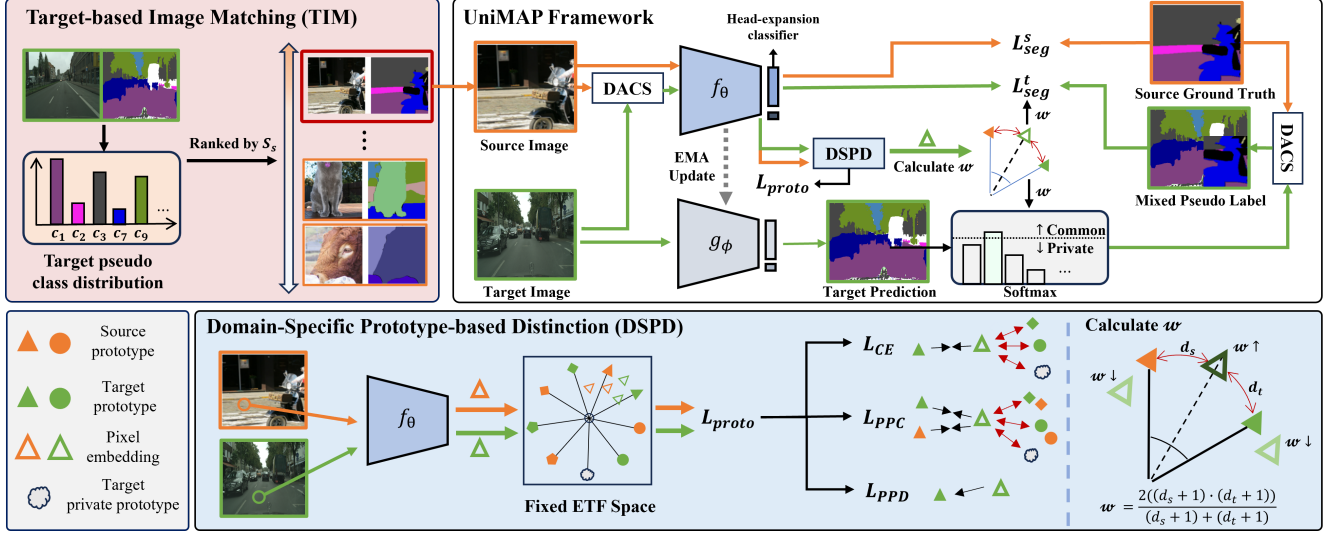


Figure 2. Overview of our proposed method, UniMAP. The top right illustrates the main training framework. The model is optimized with three main losses: the supervised segmentation loss on the source domain L_{seg}^s , the pseudo-label guided loss on the target domain L_{seg}^t using DACS [31], which is a domain mixing technique, and L_{proto} , the prototype-based loss L_{proto} computed in a fixed ETF space [26]. L_{proto} consists of three losses, which allows the prototype to have domain-specific information. Pixel-wise weight scaling factor w , is derived based on the relative distance between source and target prototypes, assigning higher weights to common classes that align well with both prototypes. These weights are used in generating target pseudo-labels and the target loss L_{seg}^t . On the top left is the framework of TIM. It computes the class distribution of the target pseudo-label and ranks source images based on class overlap using the similarity score S_s . The top-ranked source image is selected and paired with the target image in each training batch.

as follows:

$$\hat{y}_{tp}^{(j)} = \begin{cases} c', & \text{if } (\max_{c'} g_\phi(x_t)^{(j,c')} \geq \tau_p) \\ C_s + 1, & \text{otherwise} \end{cases}, \quad (2)$$

where $c' \in \{1, 2, \dots, C_s\}$ denotes a known classes and τ_p is a fixed threshold for assign unknown pseudo-labels. Then, we calculate the image-level reliability of the pseudo-label q_t as follows [31]:

$$q_t = \frac{1}{H \cdot W} \sum_{j=1}^{H \cdot W} \left[\max_{c'} g_\phi(x_t)^{(j,c')} \geq \tau_t \right], \quad (3)$$

where τ_t is a hyperparameter. The network f_θ is trained using the pseudo-labels and the corresponding confidence estimates with the using the weighted cross-entropy loss \mathcal{L}_{seg}^t :

$$\mathcal{L}_{seg}^t = - \sum_{j=1}^{H \cdot W} \sum_{c=1}^{C_s+1} q_t \cdot \hat{y}_{tp}^{(j,c)} \log f_\theta(x_t)^{(j,c)}. \quad (4)$$

Based on this baseline, we propose a novel framework called UniMAP, short for **Uni**DA-SS with **I**mage **M**atching and **P**rototype-based **D**istinction.

3.3. Domain-Specific Prototype-based Distinction

Prototype-based Learning. In conventional self-training-based UDA-SS methods, common classes from both the source and target domains are typically treated

as a unified class, assuming identical feature representations. However, in practice, common classes often exhibit domain-specific features (e.g., road appearance and texture differences between Europe and India). To address this issue, we leverage the concept from ProtoSeg [44]. ProtoSeg uses multiple non-learnable prototypes per class to represent diverse features within the pixel embedding space, adequately capturing inter-class variance. Building on this idea, we assign two prototypes per class, one for the source and one for the target. This allows the model to capture domain-specific features for each class while still learning them as a unified class, effectively enhancing the confidence scores for common classes in the target domain. To ensure that the source and target prototypes maintain a stable distance, we use a fixed Simplex Equiangular Tight Frame (ETF) [26], which guarantees equal cosine similarity and L2-norm across all prototype pairs. This structure enables consistent separation between the source and target prototypes, facilitating the learning of domain-specific features. The prototypes are defined as follows:

$$\{p_k\}_{k=1}^{2C+1} = \sqrt{\frac{2C+1}{2C}} U \left(I_{2C+1} - \frac{1}{2C+1} \mathbf{1}_{[2C+1]} \mathbf{1}_{[2C+1]}^T \right), \quad (5)$$

Each class has a pair of prototypes $p^c \in \{p_s^c, p_t^c\}$, with an additional prototype is defined for unknown classes $p^{C+1} \in \{p_t^{C+1}\}$. We employ three prototype-based loss functions adapted from ProtoSeg for each domain $D \in \{s, t\}$. First,

the cross entropy loss \mathcal{L}_{CE} that moves the target closer to the corresponding prototype and further away from the rest of the prototype as follows:

$$\mathcal{L}_{CE}^D = -\log \frac{\exp(i^\top p_D^c)}{\exp(i^\top p_D^c) + \sum_{c' \neq c} \exp(i^\top p_D^{c'})}, \quad (6)$$

where i represents the L2-normalized pixel embedding, using the label for source pixels and the pseudo-label for target pixels to determine the corresponding class c . Second, pixel-prototype contrastive learning strategy \mathcal{L}_{PPC} , which makes it closer to the corresponding prototype in the entire space and farther away from the rest as follows:

$$\mathcal{L}_{PPC}^D = -\log \frac{\sum_{p \in p^c} \exp(i^\top p^c / \tau)}{\sum_{p \in p^c} \exp(i^\top p^c / \tau) + \sum_{p^- \in P^-} \exp(i^\top p^- / \tau)}, \quad (7)$$

where P^- denotes set of prototypes excluding p^c . Finally, Pixel-Prototype Distance Optimization \mathcal{L}_{PPD} makes the distance of pixel embedding and prototype closer as:

$$\mathcal{L}_{PPD}^D = (1 - i^\top p_D^c)^2. \quad (8)$$

Therefore, we can organize \mathcal{L}_{proto} as follows:

$$\mathcal{L}_{proto} = \mathcal{L}_{CE} + \lambda_1 \mathcal{L}_{PPC} + \lambda_2 \mathcal{L}_{PPD}, \quad (9)$$

where λ_1 and λ_2 denote hyperparameters. Through the \mathcal{L}_{proto} , the model can capture domain-specific features while learning each class as a unified representation.

Prototype-based Weight Scaling. We further utilize prototypes to distinguish between common class and source-private. As training progresses, common-class pixel embeddings tend to align with both source and target prototypes, whereas private-class embeddings align with only one. Thus, when an embedding is similarly close to both prototypes, it is likely to be from a common class. Based on this, we assign a pixel-wise weight scaling factor w to reflect the likelihood of a pixel belonging to a common class:

$$w = \frac{2(d_s + 1)(d_t + 1)}{(d_s + 1) + (d_t + 1)}, \quad (10)$$

where d_s, d_t denote cosine similarity between pixel embedding i and the source and target prototypes p_s^c, p_t^c , respectively. The scaling factor w is then applied to Equation (11) during pseudo-label generation as follows:

$$\mathcal{L}_{seg}^t = - \sum_{j=1}^{H \cdot W} \sum_{c=1}^{C+1} w \cdot q_t \hat{y}_{tp}^{(j,c)} \log f_\theta(x_t)^{(j,c)}, \quad (11)$$

$$\hat{y}_{tp}^{(j)} = \begin{cases} c', & \text{if } (\max_{c'} g_\phi(x_t)^{(j,c')} \cdot w \geq \tau_p) \\ C+1, & \text{otherwise} \end{cases}. \quad (12)$$

The above method mitigates the assignment of a common class to target-private in the target pseudo-label and enhances the learning of pixels with a high probability of a common class.

3.4. Target-based Image Matching

To increase the confidence score of common classes, it is important to include as many common classes as possible in the training to learn domain-invariant representation. However, when source-private classes are added, the proportion of learning common classes decreases, making it difficult to learn domain-invariant representation. To solve this problem, we propose the Target-based Image Matching (TIM) method, which selects images containing as many common classes as possible from source images based on the classes appearing in the target pseudo-label. First, we calculate the proportion of each class present in the target pseudo-label \hat{y}_{tp} as follows:

$$f_c = \frac{n_c}{\sum_k n_k}, \quad (13)$$

where n_c denotes the number of pixels of class c in \hat{y}_{tp} . Utilizing f_c we calculate \hat{f}_c , which has a higher value for rare classes, as follows:

$$\hat{f}_c = \text{softmax}\left(\frac{1 - f_c}{T}\right), \quad (14)$$

where T denotes temperature. For each source image through \hat{f}_c , we measure S_s as follows:

$$S_s = \sum_{c \in c^*} n_c^s \hat{f}_c, \quad (15)$$

where n_c^s denotes the number of pixels of class c in y_s and c^* denotes set of overlapping classes between y_s and \hat{y}_{tp} . So, we select the source image with the highest S_s and pair it with the corresponding target image in a training batch. This approach allows us to effectively learn domain-invariant representations for common classes, which can improve performance in a variety of scenarios. It also mitigates class imbalance by prioritizing source images that contain more pixels from rare common classes, guided by class weighting based on the target class distribution.

4. Experiments

4.1. Experimental Setup

Datasets. We evaluated our method on two newly defined OPDA-SS benchmarks: Pascal-Context [24] \rightarrow

Pascal-Context → Cityscapes															
Method	Road	S.walk	Build.	Wall	Fence	Veget.	Sky	Car	Truck	Bus	M.bike	Bike	Common	Private	H-score
UAN [38]	61.78	13.14	78.14	0.03	5.60	20.01	81.50	33.2	36.24	4.90	15.48	13.01	31.93	4.30	7.47
UniOT [2]	62.34	15.64	75.69	0.05	4.61	21.50	78.10	34.3	35.04	5.94	12.98	15.85	32.84	6.85	10.76
MLNet [22]	71.28	12.94	68.63	0.00	6.15	19.73	81.7	22.8	27.04	4.45	11.68	10.72	30.81	6.43	10.61
DAFormer [13]	25.29	0.00	83.44	0.09	7.69	86.94	91.68	91.59	81.80	66.18	55.66	60.49	54.24	4.43	8.20
HRDA [14]	62.33	0.00	77.75	0.64	30.87	80.49	83.24	88.79	70.11	58.66	9.11	21.75	51.89	8.55	14.68
MIC [15]	40.49	0.21	79.40	0.00	8.35	85.74	89.58	84.78	46.87	47.23	47.78	53.59	48.67	7.85	13.51
BUS [6]	77.90	0.01	85.26	0.00	31.16	87.12	88.43	89.94	64.51	53.71	50.22	63.40	57.64	20.38	30.11
UniMAP (Ours)	84.15	16.77	86.38	0.00	35.12	88.26	89.45	90.75	64.54	59.25	49.98	66.63	60.94	31.27	41.33

Table 1. Semantic segmentation performance on Pascal-Context → Cityscapes OPDA-SS benchmarks. Our method outperformed baselines in common, private, and overall performance. White columns show individual common class scores, while “Common” in gray columns represents the average performance of common classes. The best results are highlighted in bold.

GTA5 → IDD																				
Method	Road	S.walk	Build.	Wall	Fence	Pole	Light	Sign	Veget.	Sky	Person	Rider	Car	Truck	Bus	M.bike	Bike	Common	Private	H-score
UAN [38]	97.38	61.33	62.24	36.27	16.41	24.11	8.96	58.29	78.82	94.15	57.06	30.09	68.98	72.92	42.66	64.93	7.85	49.20	3.14	5.92
UniOT [2]	96.99	41.19	63.61	34.63	18.96	28.35	3.96	54.07	72.9	92.89	53.9	32.36	81.82	72.85	63.84	63.28	5.18	51.82	7.44	13.01
MLNet [22]	95.59	9.87	55.53	17.26	12.14	12.69	5.81	64.13	72.69	91.57	0.00	17.92	69.59	65.65	50.35	60.76	5.39	41.58	4.23	7.68
DAFormer [13]	97.89	54.84	70.28	43.71	25.56	37.74	14.57	66.80	79.14	91.92	58.31	52.31	83.36	80.14	77.16	64.70	21.54	52.05	21.07	29.99
HRDA [14]	97.90	52.22	69.80	42.73	25.15	38.79	21.43	66.80	80.06	91.38	57.60	50.83	83.27	80.05	76.35	64.05	20.07	57.83	22.47	32.69
MIC [15]	95.18	39.64	67.66	43.19	23.08	36.32	17.06	65.09	85.39	94.48	53.37	57.35	79.67	81.47	65.86	65.40	20.27	56.42	24.68	34.82
BUS [6]	98.31	74.34	73.65	48.05	34.62	46.21	30.15	74.17	87.06	95.77	64.38	66.91	89.31	87.84	89.77	71.89	16.25	65.47	29.70	41.26
UniMAP (Ours)	98.13	62.50	76.12	85.74	27.48	46.56	26.07	59.63	90.44	96.31	65.87	66.85	82.83	87.08	68.33	70.27	35.45	64.08	34.78	45.51

Table 2. Semantic segmentation performance on GTA5 → IDD OPDA-SS benchmarks. Our method outperformed baselines in common, private, and overall performance. White columns show individual common class scores, while “Common” in gray columns represents the average performance of common classes. The best results are highlighted in bold.

Cityscapes [7], and GTA5 [27] → IDD [34], which we introduce to assess universal domain adaptation in more realistic settings involving both source-private and target-private classes. Pascal-Context → Cityscapes is a real-to-real scenario, and Pascal-Context contains both in-door and out-door, while Cityscapes only has a driving scene, so it is a scenario with a considerable amount of source-private classes. We selected 12 classes as common classes and the remaining 7 classes (“pole”, “light”, “sign”, “terrain”, “person”, “rider”, and “train”) are treated as target-private classes. GTA5 → IDD is a synthetic-to-real scenario and GTA5 features highly detailed synthetic driving scenes set in urban cityscapes, while IDD captures real-world driving scenarios on diverse roads in India. We used 17 classes as common classes, 2 source-private class (“terrain”, “train”), and 1 target-private class (“auto-rickshaw”).

Evaluation Protocols. In the OPDA-SS setting, both common class and target-private performance are important, so we evaluate methods using H-Score, which can fully reflect them. The H-score is calculated as the harmonic mean of the common mIoU (mean Intersection-over-Union) and the target-private IoU.

Implementation Details. This method is based on BUS. We used the multi-resolution self-training strategy and training parameter used in MIC [15]. The network used a MiT-

B5 [37] encoder and was initialized with ImageNet-1k [8] pretrained. The learning rate was 6e-5 for the backbone and 6e-4 for the decoder head, with a weight decay of 0.01 and linear learning rate warm-up over 1.5k steps. EMA factor α was 0.999 and the optimizer was AdamW [17]. ImageNet feature Distance [13], DACS [31] data augmentation, Masked Image Consistency module [15], and Dilation-Erosion-based Contrastive Loss [6] were used. We also modified some of the BUS methods to suit the OPDA setting. In OpenReMix [6], we applied only Resizing Object except Attaching Private and did not use refinement through MobileSAM [39]. For rare class sampling [13], we switched from calculating a distribution based on the existing source and applying it to source sampling to applying it to target sampling based on the target pseudo-label distribution. We trained on a batch of two 512×512 random crops for 40k iterations. The hyperparameter are set to: $\tau_p = 0.5$, $\tau_t = 0.968$, $\lambda_1 = 0.01$, $\lambda_2 = 0.01$, $\tau = 0.1$, and $T = 0.01$.

Baselines. Since there is no existing research on OPDA-SS, we performed experiments by changing the methods in different settings to suit the OPDA-SS. First, for UniDA for classification methods [2, 22, 38], we experimented by changing the backbone to a semantic segmentation model. In this case, we used the DeepLabv2 [4] segmentation network and ResNet-101 [11] as the backbone. For the CDA-SS methods [13–15], we added 1 dimension to the head dimension of the classifier to predict the target-private and

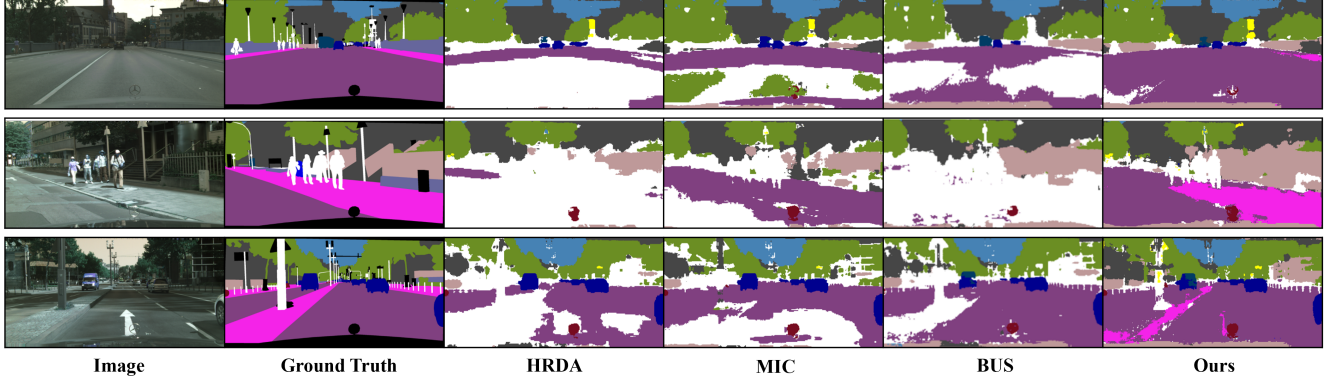


Figure 3. Qualitative results of OPDA-SS setting. We visualize the segmentation predictions from different methods on the Cityscapes dataset. White and yellow represent target-private and source-private classes, respectively, while other colors indicate common classes (e.g., purple for “road” and pink for “sidewalk”). Compared to HRDA, MIC, and BUS, our method more accurately segments both common and target-private classes.

assigned an unknown based on the confidence score [6]. Lastly, the ODA-SS method, BUS [6], was used as it is.

4.2. Comparisons with the State-of-the-Art

We compared performance on two benchmarks for OPDA-SS settings. Table 1 presents the semantic segmentation performance from Pascal-Context \rightarrow Cityscapes, while Table 2 presents the performance from GTA5 \rightarrow IDD. As shown in Table 1, UniMAP achieved outstanding performance in the Pascal-Context \rightarrow Cityscapes benchmark. Specifically, it outperformed previous approaches by a significant margin, with improvements of approximately 3.3 for Common, 10.89 for Private, and 11.22 in H-score. These results indicate that UniMAP effectively enables the model to learn both common and private classes. Notably, UniMAP surpassed BUS, the state-of-the-art in ODA-SS, in terms of private class performance. Although our method primarily focuses on capturing knowledge of common classes, it also enhances the identification of private classes due to improved representation learning. In addition, Table 2 shows the performance comparison for the GTA5 \rightarrow IDD benchmark. Our method demonstrated notable improvements in both Private and H-Score. In particular, while prior methods in CDA-SS showed inferior performance for Private and H-score, our approach led to significant gains of approximately 6.25 for Common, 10.3 for Private, and 9.69 for H-score. Although our method had a relatively lower performance than BUS in Common, it surpassed BUS in Private performance with a margin of about 5.08, ultimately leading to superior H-score results. Overall, the experimental findings demonstrate that our method delivers promising performance in OPDA-SS settings, which is critical for achieving effective UniDA-SS.

4.3. Qualitative Evaluation

We conducted qualitative experiments under the OPDA-SS setting. Figure 3 compared prediction maps from

UniMAP		Pascal-Context \rightarrow Cityscapes		
DSPD	TIM	Common	Private	H-Score
		53.79	26.54	36.03
✓		59.46	27.97	38.04
	✓	56.22	29.14	38.39
✓	✓	60.94	31.27	41.33

Table 3. Ablation study of our method on Pascal-Context \rightarrow Cityscapes. We evaluate the contributions of DSPD and TIM, where the baseline is BUS without private attaching and pseudo-label refinement. The best results are highlighted in bold.

DSPD		Pascal-Context \rightarrow Cityscapes		
w	\mathcal{L}_{proto}	Common	Private	H-Score
		53.79	26.54	36.03
✓		54.38	21.75	31.08
	✓	59.71	26.76	36.96
✓	✓	59.46	27.97	38.04

Table 4. Further ablation study of DSPD components on Pascal-Context \rightarrow Cityscapes. w represents pixel-wise weight scaling factor, and \mathcal{L}_{proto} represents the prototype loss function. The best results are highlighted in bold.

Cityscapes against baselines, where white and yellow represent target-private and source-private classes, respectively, while other colors denote common classes. Baseline methods such as HRDA, MIC, and BUS tend to either misclassify common classes as target-private or sacrifice common class accuracy to detect target-private regions. In contrast, UniMAP successfully predicted both common and target-private classes. Notably, it accurately identified the “sidewalk” class (pink) in rows 2 and 3, unlike other baselines. These results indicate that UniMAP effectively balances the identification of common and target-private classes.

4.4. Ablation Study

Ablation Study about UniMAP. Table 3 shows the experimental results of the ablation study of the performance

Method	Pascal-Context → Cityscapes								Common Average	H-Score Average
	Open Partial Set DA			Open Set DA			Partial Set DA	Closed Set DA		
	Common	Private	H-Score	Common	Private	H-Score	Common	Common		
DAF	54.24	4.43	8.19	44.27	12.07	18.97	35.18	46.48	44.51	12.46
HRDA	51.89	8.55	14.68	52.76	14.76	23.07	51.99	63.17	54.76	18.40
MIC	48.67	7.85	13.52	60.88	23.79	34.21	58.04	65.68	57.97	21.51
BUS	57.64	20.38	30.11	60.67	27.05	37.42	58.54	60.24	59.26	33.57
UniMAP (Ours)	60.94	31.27	41.33	58.50	24.73	34.76	59.44	64.74	60.86	37.90

Table 5. Experimental results on Pascal-Context → Cityscapes for various domain adaptation scenarios. For a fair comparison, all methods used a head-expansion model. The best results are highlighted in bold.

contribution of each component. As described in the Implementation Details section, the baseline model, derived by removing the Attaching Private and refinement pseudo-label module from the BUS, achieved an H-Score of 36.03. First, applying DSPD alone to the baseline, the H-Score improves to 38.04, increasing both Common and Private performance. This enhancement indicates that DSPD effectively captures domain-specific features, improving performance for both the common and target-private classes compared to the Baseline. Next, when only applying TIM alone to the baseline, also improves performance, achieving an H-Score of 38.39, with better Private. This result suggests that TIM successfully learns domain-invariant representations between source and target by leveraging target pseudo-labels, thus enhancing overall performance. Finally, when both DSPD and TIM are applied to the baseline, the model achieves the best performance, with an H-Score of 41.33. This demonstrates that DSPD and TIM work synergistically, enabling the model to achieve superior performance across both common and target-private classes.

Ablation Study about DSPD. Table 4 shows the impact of the individual components of DSPD, namely L_{proto} and w on performance in the Pascal-Context → Cityscapes scenario. The L_{proto} represents the pixel embedding loss in the ETF space, designed to guide pixel embeddings within a class to be closer to their respective prototypes. When only L_{proto} is applied, the model achieves a Common of 59.71, a Private of 26.76, and an H-Score of 36.96. This result suggests that L_{proto} alone can enhance the clustering of pixel embeddings around domain-specific prototypes, thereby improving overall performance compared to the baseline. The w , on the other hand, means a weighting mechanism based on the ETF prototype structure that estimates the common class more effectively and applies weights scaling accordingly. When only w is used, the Common drops to 54.38, and the Private score falls to 21.75, resulting in a lower H-Score of 31.08. This indicates that while w is utilized in distinguishing common classes, it is less effective without the guidance provided by L_{proto} . When both L_{proto} and w are combined, the model

achieves the best performance, with a Common of 59.46, a Private of 27.97, and an H-Score of 38.04. This demonstrates that the two components are complementary: L_{proto} enhances pixel embedding alignment with domain-specific prototypes, while w further boosts the ability to focus on common class pixels with appropriate weighting. Together, they yield a notable improvement in the overall H-Score.

Comparisons in Various Category Settings. We further compared the performance generalization ability of UniMAP across various domain adaptation settings. As shown in 5, while some existing methods achieve slightly better results in Closed Set and Open Set settings due to their specialized assumptions, UniMAP demonstrates clear advantages in Partial Set and Open Partial Set, where prior methods have not been actively explored. Notably, UniMAP achieves the highest scores, with a Common Average of 60.86 and an H-Score Average of 37.90, validating its robustness and effectiveness across varying category shift configurations. These results highlight the practicality of our framework for the real-world scenario, where category settings are often unknown.

5. Conclusion

In this paper, we proposed a new framework for UniDA-SS, called UniMAP. Since UniDA-SS must handle different domain configurations without prior knowledge of category settings, it is very important to identify and learn common classes across domains. To this end, UniMAP incorporates two key components: Domain-Specific Prototype-based Distinction (DSPD) and Target-based Image Matching (TIM). DSPD is used to estimate common classes from the unlabeled target domain, while TIM samples labeled source images to transfer knowledge to the target domain effectively. Experimental results show that our method improved average performance across different domain adaptation scenarios. We hope our approach sheds light on the necessity of universal domain adaptation for the semantic segmentation task.

Acknowledgment

This research was conducted with the support of the HANCOM InSpace Co., Ltd. (Hancom-Kyung Hee Artificial Intelligence Research Institute), and was supported by Korea Planning & Evaluation Institute of Industrial Technology (KEIT) grant funded by the Korea government (MOTIE) (RS-2024-00444344), and in part by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2019-II190079, Artificial Intelligence Graduate School Program (Korea University)).

References

- [1] Silvia Bucci, Mohammad Reza Loghmani, and Tatiana Tomasi. On the effectiveness of image rotation for open set domain adaptation. In *European conference on computer vision*, pages 422–438. Springer, 2020. 3
- [2] Wanxing Chang, Ye Shi, Hoang Tuan, and Jingya Wang. Unified optimal transport framework for universal domain adaptation. *Advances in Neural Information Processing Systems*, 35:29512–29524, 2022. 6
- [3] Cheng Chen, Qi Dou, Hao Chen, Jing Qin, and Pheng-Ann Heng. Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 865–872, 2019. 3
- [4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 3, 6
- [5] Minghao Chen, Hongyang Xue, and Deng Cai. Domain adaptation for semantic segmentation with maximum squares loss. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2090–2099, 2019. 3
- [6] Seun-An Choe, Ah-Hyung Shin, Keon-Hee Park, Jinwoo Choi, and Gyeong-Moon Park. Open-set domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23943–23953, 2024. 2, 3, 6, 7
- [7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 6
- [8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6
- [9] Liang Du, Jingang Tan, Hongye Yang, Jianfeng Feng, Xi-angyang Xue, Qibao Zheng, Xiaoqing Ye, and Xiaolin Zhang. Ssf-dan: Separated semantic feature based domain adaptation network for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 982–991, 2019. 3
- [10] Bo Fu, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. Learning to detect open classes for universal domain adaptation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 567–583. Springer, 2020. 3
- [11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6
- [12] Weixiang Hong, Zhenzhen Wang, Ming Yang, and Junsong Yuan. Conditional generative adversarial network for structured domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1335–1344, 2018. 3
- [13] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. 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*, pages 9924–9935, 2022. 3, 6
- [14] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Hrda: Context-aware high-resolution domain-adaptive semantic segmentation. In *European Conference on Computer Vision*, pages 372–391. Springer, 2022. 3, 6
- [15] Lukas Hoyer, Dengxin Dai, Haoran Wang, and Luc Van Gool. Mic: Masked image consistency for context-enhanced domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11721–11732, 2023. 2, 6
- [16] Myeongjin Kim and Hyeran Byun. Learning texture invariant representation for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12975–12984, 2020. 3
- [17] Xiangtai Li, Xia Li, Li Zhang, Guangliang Cheng, Jianping Shi, Zhouchen Lin, Shaohua Tan, and Yunhai Tong. Improving semantic segmentation via decoupled body and edge supervision. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 435–452. Springer, 2020. 6
- [18] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6936–6945, 2019. 3
- [19] Qing Lian, Fengmao Lv, Lixin Duan, and Boqing Gong. 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*, pages 6758–6767, 2019. 3
- [20] Wei Liu, Andrew Rabinovich, and Alexander C Berg. Parsenet: Looking wider to see better. *arXiv preprint arXiv:1506.04579*, 2015. 3
- [21] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 3

- [22] Yanzuo Lu, Meng Shen, Andy J Ma, Xiaohua Xie, and Jian-Huang Lai. Mlnet: Mutual learning network with neighborhood invariance for universal domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3900–3908, 2024. 6
- [23] Luke Melas-Kyriazi and Arjun K Manrai. Pixmatch: Unsupervised domain adaptation via pixelwise consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12435–12445, 2021. 3
- [24] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 891–898, 2014. 5
- [25] Fei Pan, Inkyu Shin, Francois Rameau, Seokju Lee, and In So Kweon. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3764–3773, 2020. 3
- [26] Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020. 4
- [27] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II 14*, pages 102–118. Springer, 2016. 6
- [28] Kuniaki Saito and Kate Saenko. Ovanet: One-vs-all network for universal domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9000–9009, 2021. 3
- [29] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, and Kate Saenko. Universal domain adaptation through self supervision. *Advances in neural information processing systems*, 33:16282–16292, 2020. 3
- [30] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 3
- [31] Wilhelm Tranehen, Viktor Olsson, Juliano Pinto, and Lennart Svensson. Dacs: Domain adaptation via cross-domain mixed sampling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1379–1389, 2021. 3, 4, 6
- [32] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schuster, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7472–7481, 2018. 3
- [33] Yi-Hsuan Tsai, Kihyuk Sohn, Samuel Schuster, and Manmohan Chandraker. Domain adaptation for structured output via discriminative patch representations. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1456–1465, 2019. 3
- [34] Girish Varma, Anbumani Subramanian, Anoop Namboodiri, Manmohan Chandraker, and CV Jawahar. Idd: A dataset for exploring problems of autonomous navigation in unconstrained environments. In *2019 IEEE winter conference on applications of computer vision (WACV)*, pages 1743–1751. IEEE, 2019. 6
- [35] Qin Wang, Dengxin Dai, Lukas Hoyer, Luc Van Gool, and Olga Fink. Domain adaptive semantic segmentation with self-supervised depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8515–8525, 2021. 3
- [36] Yuxi Wang, Junran Peng, and ZhaoXiang Zhang. Uncertainty-aware pseudo label refinery for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9092–9101, 2021. 3
- [37] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 3, 6
- [38] Kaichao You, Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Universal domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2720–2729, 2019. 3, 6
- [39] Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289*, 2023. 6
- [40] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12414–12424, 2021. 3
- [41] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 3
- [42] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. Pscanet: Pointwise spatial attention network for scene parsing. In *Proceedings of the European conference on computer vision (ECCV)*, pages 267–283, 2018. 3
- [43] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021. 3
- [44] Tianfei Zhou, Wenguan Wang, Ender Konukoglu, and Luc Van Gool. Rethinking semantic segmentation: A prototype view. In *Proceedings of the IEEE/CVF Conference*

- on *Computer Vision and Pattern Recognition*, pages 2582–2593, 2022. [4](#)
- [45] Zhen Zhu, Mengde Xu, Song Bai, Tengting Huang, and Xiang Bai. Asymmetric non-local neural networks for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 593–602, 2019. [3](#)
- [46] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pages 289–305, 2018. [3](#)
- [47] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5982–5991, 2019. [3](#)