MarCon: Max-Margin Contrastive Learning for Imbalanced Domain Adaptation Semantic Segmentation

Yibo Wang[1,2,3], **Ruikang Xu**[3], **Guangcheng Zhu**[3], **Cheng Peng**[1,3*], **Haobo Wang**[1,2,3*], **Runze Wu**[4], **Minmin Lin**[4], **Changjie Fan**[4]
[1]School of Software Technology, Zhejiang University
[2]Hangzhou High-Tech Zone (Binjiang) Institute of Blockchain and Data Security
[3]State Key Laboratory of Blockchain and Data Security, Zhejiang University
[4]NetEase Fuxi AI Lab, Hangzhou
{sonnytheone, 3220105566, zhuguangcheng, chengchng, wanghaobo}@zju.edu.cn, {wurunze1, linminmin01, fanchangjie}@corp.netease.com

Abstract

Unsupervised Domain Adaptation for Semantic Segmentation (UDA-SS) has seen significant progress in recent years. Existing UDA-SS approaches mostly adopt a pseudo-labeling schema to adapt model in the target domain, but they often overlook the inherent long-tailed data distribution in segmentation. We find that such scarce tail samples can lead to representation collapse for tail classes and further hinder the quality of pseudo-labels. To address these challenges, we propose MarCon, a framework that mitigates the long-tailed problem from both feature representation and pseudo-labeling perspectives. Specifically, to explicitly learn a Maximum-Margin Distribution, we derive a reformulated pixel-level contrastive learning objective by modeling feature distributions with the von Mises-Fisher (vMF) distribution. It enforces strict margins to enhance intra-class compactness and inter-class separability, preventing tail classes from being overwhelmed by head classes. Furthermore, to mitigate label noise, we introduce a Reliability-aware Filter (RaF) based on the vMF-derived metrics, which performs adaptive class-wise pixel reliability assessment to identify unreliable pixels and attenuate their contribution to model training, thereby mitigating confirmation bias. Extensive experiments on GTA $\rightarrow$ Cityscapes and SYNTHIA $\rightarrow$ Cityscapes demonstrate that MarCon consistently outperforms current leading methods across various transformer-based architectures.

1 Introduction

Semantic Segmentation (SS) is a fundamental task in computer vision that aims to assign a specific semantic label to every pixel in an image. Due to its capability for dense prediction, this technique has been extensively applied across various critical domains, including medical analysis [Ronneberger *et al.*, 2015] and autonomous driving [Zhang *et al.*, 2018]. However, most of the current advanced methods are built upon deep neural networks [Chen *et al.*, 2018; Zheng *et al.*, 2021], which are largely hungry for huge fully-supervised datasets [Long *et al.*, 2015]. Therefore, to circumvent such labor-intensive annotation work, Unsupervised Domain Adaptation Semantic Segmentation (UDA-SS) [Toldo *et al.*, 2020] provides a feasible cross-domain learning paradigm to transfer knowledge from the label-rich source domain to the vast unlabeled target domain.

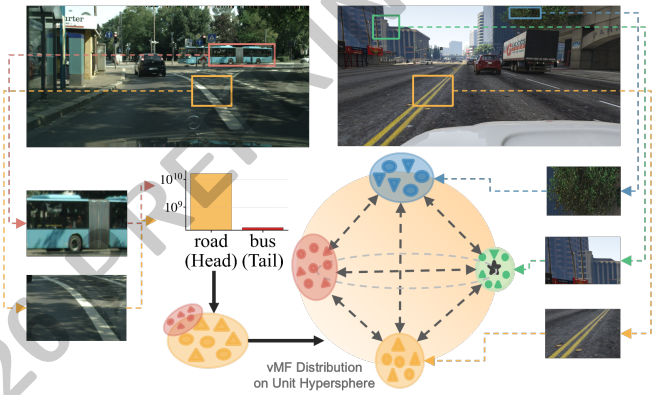


Figure 1: Illustration of the tail-class feature collapse challenge in UDA-SS and our proposed solution. **Left:** The inherent long-tailed class distribution (e.g., vast road pixels vs. scarce bus pixels) leads to feature collapse, where tail-class features (red region) are overwhelmed by head-class ones (orange region). **Right:** Our Max-Margin Contrastive Learning (MCL) models the feature space using vMF distributions. By enforcing strict margins, it ensures highly discriminative feature representations.

Typically, UDA-SS explores data distribution alignment within a domain-adaptation mechanism, which leverages the source information to facilitate precise pixel-wise classification on the target domain. Early works mostly followed the conventions of adversarial learning to enforce domain alignment from diverse aspects, e.g., input appearance, intermediate feature representations, and output spaces [Vu *et*

*Corresponding authors.

al., 2019; Hoffman *et al.*, 2018; Luo *et al.*, 2022b]. Subsequently, self-training strategies became mainstream due to the ability to mitigate domain-shift noise via filtering or rectifying pseudo-labels [Zhang *et al.*, 2021; Zhao *et al.*, 2024]. Nowadays, thanks to advanced Transformer-based architectures [Hoyer *et al.*, 2022a; Hoyer *et al.*, 2022b] and contrastive learning techniques, UDA-SS has flourished from the perspective of representation learning [Hoyer *et al.*, 2023].

Despite the promise, the recent UDA-SS research is still exposed to an overly idealized scenario of balanced datasets. As shown in Fig 1, SS data often reveals a severe long-tailed distribution, i.e., sparse tail-class pixels are usually hidden among a large number of head-class ones, which leads the training process to be dominated by head classes, triggering the feature collapse on tail classes. Moreover, such degradation is further exacerbated in UDA-SS, thereby restricting the effective knowledge transfer from the source to the target domain. Specifically, the target tail-class pixels are overwhelmed by the noisy pseudo-labels of head classes, which in turn introduces erroneous supervision for representation learning in a vicious cycle. Therefore, UDA-SS urgently needs a tail-effective learning paradigm that can enforce strict feature margins across classes and ensure the quality of pseudo-labels simultaneously.

To address the above challenges, we propose MarCon, a method that mitigates the long-tailed problem from two synergistic aspects — representation modeling and pseudo-labeling — in one coherent framework. Specifically, we first introduce Max-Margin Contrastive Learning (MCL) to rectify feature collapse. By explicitly modeling the feature space with the von Mises-Fisher (vMF) distribution, MCL reformulates pixel-level contrastive learning to enforce strict maximum margins. This constraint explicitly encourages the formation of a feature space where pixels of the same class form compact clusters, distinctly separated from other classes. Second, to guarantee the quality of the target domain pseudo-labels, we propose the Reliability-aware Filter (RaF). Leveraging the MCL-derived loss and spherical distribution affinity, RaF performs an adaptive and class-wise reliability assessment, which prevents the unreliable predictions from corrupting the model optimization. By coordinating the two modules above, MarCon explores a solution towards imbalanced UDA-SS for robust pixel segmentation. In summary, our main contributions are as follows:

- To our knowledge, this paper pioneers the exploration of a max-margin-view contrastive representation learning for imbalanced UDA-SS and proposes a tail-robust framework. As an integral part of MarCon, **Max-Margin Contrastive Learning** reformulates pixel-level contrastive learning, which enhances feature discriminability and mitigates the tail-class collapse.
- To realize cleaner target supervision, MarCon further integrates the **Reliability-aware Filter** module, which assesses target pixels' reliability and attenuates the impact of unreliable pixels to form a robust feedback loop.
- We verify the utility of our method through extensive testing on diverse UDA-SS benchmarks. MarCon demonstrates superior adaptability, consistently improving performance across all experimental configurations.

2 Related Work

2.1 Unsupervised Domain Adaptation Semantic Segmentation

Unsupervised Domain Adaptation Semantic Segmentation (UDA-SS) transfers knowledge from labeled source domains to unlabeled target domains, bypassing the need for laborious pixel-wise annotations. Recent UDA-SS works typically follow a self-supervised learning paradigm [Chen *et al.*, 2024b; Ren *et al.*, 2024; Li *et al.*, 2025b], especially some state-of-the-art methods adopt the teacher-student network for the target pixels pseudo-label prediction. Among them, DAFormer [Hoyer *et al.*, 2022a] first demonstrated the potential of Transformer architectures in UDA, introducing tailored training strategies to stabilize the learning process. However, due to the high computational cost of Transformers, input images are often downsampled. HRDA [Hoyer *et al.*, 2022b] proposed a multi-scale fusion training scheme to balance resolution with efficiency. To further enhance feature robustness, MIC [Hoyer *et al.*, 2023] leveraged the concept of masking, forcing the model to predict masked regions to strengthen its understanding of spatial context.

2.2 Long-Tail Learning in UDA-SS

Semantic segmentation inherently suffers from long-tailed distributions, where the training process is dominated by head classes. While strategies like loss re-weighting [Wang *et al.*, 2021] and over-sampling [Gupta *et al.*, 2019] have proven effective in supervised settings, the domain shifts in UDA-SS further compound this inherent class imbalance. To mitigate this challenge in UDA-SS, various strategies have been proposed. From a data augmentation perspective, methods like [Ghiasi *et al.*, 2021; Lee *et al.*, 2021] alleviate tail-class scarcity by copying source-domain tail regions onto target images. Regarding pseudo-labeling, several works [Zou *et al.*, 2018; Li *et al.*, 2022] have been proposed to mitigate the class imbalance during the pseudo-label generation. Additionally, some approaches [Li *et al.*, 2025a; Li and Li, 2021] introduce logit-level calibration to re-balance class predictions. Distinct from these methods, we propose a solution focusing on both the feature space and pseudo-label.

2.3 Contrastive Learning in UDA-SS

Contrastive Learning (CL) is a self-supervised learning framework that learns discriminative representations via instance similarity [Chen *et al.*, 2020]. By minimizing the distance between positive pairs while maximizing that of negative pairs, CL learns highly discriminative feature representations. Consequently, recent advances have incorporated CL into UDA-SS to align cross-domain representations.

Some early approaches directly adapted the standard Supervised Contrastive Loss (SCL) to perform a pixel-level paradigm [Vayyat *et al.*, 2022; Fan *et al.*, 2024]. Subsequent works acknowledged the importance of contextual information and applied contrastive constraints at both the pixel- and the patch-level [Chen *et al.*, 2023; Li *et al.*,

2023]. Besides, prototype-based methods [Jiang *et al.*, 2022; Xie *et al.*, 2023] contrast pixel embeddings against global semantic prototypes. Despite their efficiency, these methods overlook the inherent long-tailed distribution. In contrast, we reformulate the contrastive objective by explicitly modeling feature distributions via the vMF distribution, enforcing strict margins to secure discriminative space for tail classes.

3 Methods

In this section, we first define the UDA-SS task and then provide a detailed description of our MarCon framework.

Task Definition. In UDA-SS, the source-domain dataset $\mathcal{D}_s = \{(X_n^S, Y_n^S)\}_{n=1}^{N_s}$ consists of N_s images X_n^S with pixel-level labels Y_n^S in a class space $\mathcal{C}^S = \{1, \dots, C^S\}$, and the target-domain dataset $\mathcal{D}_t = \{X_n^T\}_{n=1}^{N_t}$ is provided without annotations and has a much broader label space $\mathcal{C}^T$, i.e., $\mathcal{C}^S \subseteq \mathcal{C}^T$. Then, the UDA-SS task aims to train a model F on training sets $\mathcal{D}_s$ and $\mathcal{D}_t^{train}$ to achieve good generalization performance on the test set $\mathcal{D}_t^{test}$. Generally, the segmentation model F is composed of a feature extractor g with a decoder h, and is ultimately trained by the cross-entropy loss.

As shown in Fig 2, our MarCon framework seamlessly integrates a well-designed contrastive learning branch into the self-training backbone. Specifically, we leverage Max-margin Contrastive Learning (MCL) to enforce a discriminative feature space and alleviate tail-class feature collapse. Moreover, to mitigate the noisy supervision of pseudo-labels, we design a Reliability-aware Filter (RaF) to dynamically distinguish reliable pixels via MCL reliability metrics.

3.1 Self-Training Frequency-aware Framework

To alleviate domain shift and filter high-frequency noise, we integrate the following two techniques into a unified pipeline.

Teacher-Student-Based Domain Adaption. Typically, the model F prioritizes the accurate supervision of the source training set $\mathcal{D}_s$ to ensure the basic semantic segmentation capability via the cross-entropy (CE) loss as:

$$\mathcal{L}^s = \frac{1}{N_s} \sum\nolimits_{n=1}^{N_s} \mathcal{L}_{\text{CE}}(F(X_n^S), Y_n^S). \tag{1}$$

However, relying solely on $\mathcal{D}_s$ to train the model F often leads to poor generalization performance on $\mathcal{D}_t$ due to the inherent domain shift. Consequently, we adopt the teacher-student-based self-training paradigm, following established protocols [Hoyer *et al.*, 2022a; Hoyer *et al.*, 2022b]. Specifically, we employ a teacher model F_Θ to generate the target pseudo-labels $\hat{Y}^T = \arg\max F_\Theta(X^T)$ for training the student model F_θ on $\mathcal{D}_t^{train}$ via the target domain CE loss $\mathcal{L}^t$:

$$\mathcal{L}^t = \frac{1}{N_t} \sum\nolimits_{n=1}^{N_t} \mathcal{L}_{\text{CE}}(F_\theta(X_n'^T), \hat{Y}_n^T). \tag{2}$$

Therefore, while using $\mathcal{L}^s$ to ensure strong segmentation ability, the student model F_θ also achieves effective adaptation on the target domain data via $\mathcal{L}^t$. Moreover, to further alleviate the domain shift, after the teacher model F_Θ generates pseudo-label $\hat{Y}^T$ for the original target-domain image X^T, we apply data augmentations, e.g., color jittering, Gaussian blur, and ClassMix, to generate an augmented variant X'^T following DACS [Tranheden *et al.*, 2021] for the training of the student model F_θ. Then, sharing the identical network architecture with the student model F_θ, the teacher model F_Θ updates the weight Θ via the Exponential Moving Average (EMA) of the weight θ:

$$\Theta \leftarrow \alpha\Theta + (1-\alpha)\theta, \tag{3}$$

where $\alpha \in [0, 1)$ is the smoothing coefficient.

Frequency-Aware Feature Fusion. To capture multi-scale semantic contexts and preserve spatial details, we adopt the Transformer-based model, like SegFormer, to achieve a hierarchical encoder g, which has been proven to be applicable for semantic segmentation [Xie *et al.*, 2021]. Considering that the drastic intra-class feature fluctuations inherent in semantic segmentation impede the process of multi-scale feature fusion, inspired by recent advances [Chen *et al.*, 2024a; Luo *et al.*, 2022a], we introduce a frequency-domain perspective to optimize the multi-scale feature fusion. Attributing the fluctuations to high-frequency noise, we use two complementary filters to attenuate noise-induced feature perturbations.

Specifically, we refer to this mechanism as the Frequency-Adaptive Fusion (FAF) module. First, FAF constructs a *Semantic Smoothing Filter (SSF)*. After the encoder g extracts the feature map $\mathbf{F}$, a lightweight convolutional predictor $\phi(\cdot)$ with the kernel-wise softmax normalization function $\sigma(\cdot)$ derives the smoothing kernel as $\mathbf{k}^{sem} = \sigma\left(\phi_{sem}(\mathbf{F})\right)$. Besides, to compensate for the high-frequency structural information degradation caused by indiscriminate smoothing, FAF introduces a complementary *Boundary Sharpening Filter (BSF)* via spectral inversion. By defining δ as the discrete unit impulse kernel (i.e. an all-pass spectrum), the sharpening kernel $\mathbf{k}^{bnd}$ is synthesized by subtracting the generated low-pass component from the identity: $\mathbf{k}^{\text{bnd}} = \delta - \sigma\left(\phi_{\text{bnd}}(\mathbf{F})\right)$, where δ is the Kronecker delta centered at the kernel origin. Operationally, FAF applies the low-pass SSF to low-resolution features and uses the high-pass BSF to reinforce high-resolution features, thereby completing the multi-scale fusion process.

Interestingly, as shown in Fig 5a, we observe that the effectiveness of the FAF module is compromised by domain shifts, yielding only marginal improvements. To this end, by alleviating tail-class feature collapse and constructing a discriminative feature space via strict margins, our proposed max-margin contrastive learning strategy enables the FAF module to regain its power in UDA-SS. This effect is empirically confirmed in our experimental analysis.

3.2 Max-Margin Targeted Contrastive Learning

Our primary goal is to alleviate the feature representation collapse of tail classes caused by the inherent long-tailed data distribution. To achieve this, we first theoretically quantify the discriminability of learned features.

Formally, for pixel $\mathbf{x}_i$ with label y_i, the margin γ is the cosine-similarity gap between its embedding $\mathbf{z}_i$ and true-class centroid $\boldsymbol{\mu}_{y_i}$ versus the highest confidence false-class centroid $\boldsymbol{\mu}_{c\neq y_i}$ as:

$$\gamma(\mathbf{x}_i, y_i) = \boldsymbol{\mu}_{y_i}^\top \mathbf{z}_i - \max_{c\neq y_i} \boldsymbol{\mu}_c^\top \mathbf{z}_i. \tag{4}$$

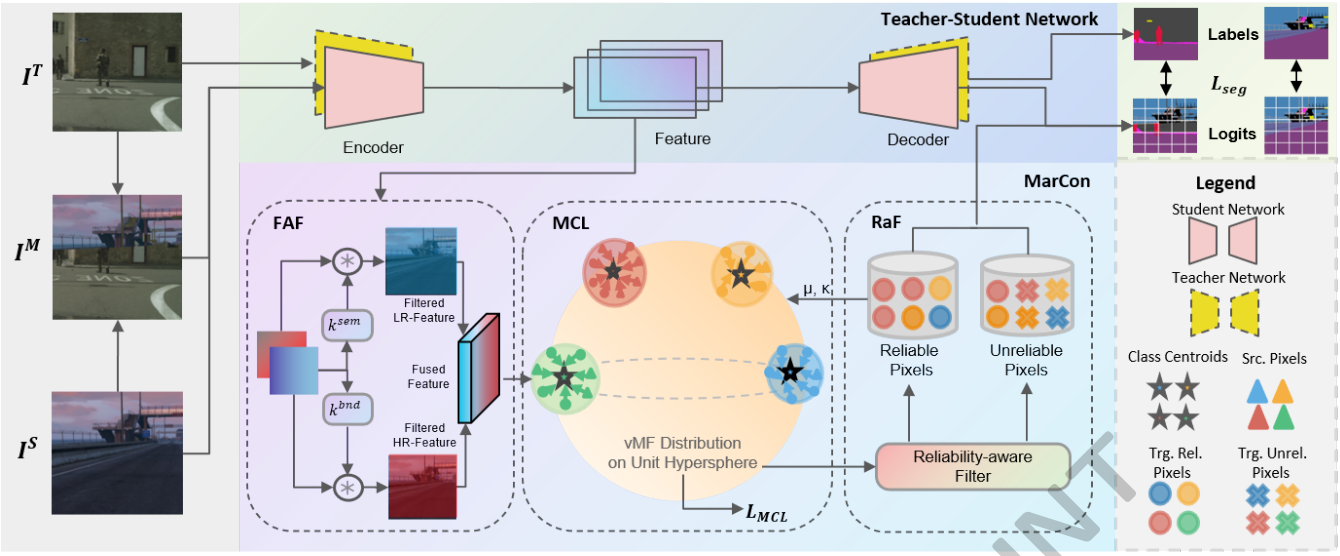


Figure 2: Overview of the proposed MarCon framework. We integrate our method into the Teacher-Student Network, which optimizes supervised source and pseudo-labeled target losses.Specifically, after the Frequency-Adaptive Fusion (FAF), we leverage Max-margin Contrastive Learning (MCL) to enforce a discriminative feature space and alleviate tail-class feature collapse. Moreover, to mitigate the noisy supervision of pseudo-labels, we design a Reliability-aware Filter (RaF) to dynamically distinguish reliable pixels via MCL-derived metrics.

Let $\mathcal{S}_c$ be the pixel set of class c, and class c's margin is $\gamma_c = \min_{\mathbf{x}\in\mathcal{S}_c} \gamma(\mathbf{x}, c)$. We find that the feature collapse of tail class pixels under long-tailed distributions is equivalent to their compressed class margins γ, which renders them highly susceptible to misclassification. Consequently, to facilitate robust feature learning across all classes, we aim to maximize the overall margin $\gamma_{\mathcal{C}} = \min_c \gamma_c$. To this end, we conduct a theoretical analysis from the optimization perspective to identify the conditions that guarantee margin maximization within a contrastive learning framework.

Proposition 1. *Let $\mathbf{z}_i \in \mathcal{Z}$ denote the ℓ_2-normalized contrastive feature of pixel $\mathbf{x}_i$ with label $y_i \in \mathcal{C}$. Under Gaussian kernel density estimation on the unit hypersphere, the optimization objective that maximizes the overall margin is:*

$$\mathcal{L}_{MM} = -\mathbb{E}_{i,\,\mathbf{z}^c\in\mathcal{Z}^c}\left[\log \frac{\pi(y_i)\exp(\mathbf{z}_i^\top \mathbf{z}^{y_i}/\tau)}{\sum_{c=1}^{C}\pi(c)\exp(\mathbf{z}_i^\top \mathbf{z}^c/\tau)}\right], \quad (5)$$

where $\pi(c)$ is the prior of class c, $\mathcal{Z}^c = \{\mathbf{z}_j | y_j = c\}$, and τ is a temperature factor. A detailed proof is showed in Appendix.

To facilitate the optimization of this target, it is essential to estimate the expectation of the exponential interaction term in Eq (5). Therefore, we commence by formulating an explicit model for the feature distribution.

Distribution Modeling. Given that contrastive features are constrained to the unit hypersphere, we assume they follow a mixture of vMF distributions, which is widely regarded as the directional equivalent of the Gaussian distribution on the hypersphere. Specifically, the probability density function of a unit vector $\mathbf{z}$ is defined as:

$$f_{\text{vMF}}(\mathbf{z}; \boldsymbol{\mu}, \kappa) = C_d(\kappa)\exp(\kappa \boldsymbol{\mu}^\top \mathbf{z}), \quad (6)$$

where $\boldsymbol{\mu}$ is the mean direction ($\|\boldsymbol{\mu}\| = 1$), $\kappa \geq 0$ is the concentration parameter and d is the feature dimension. $C_d(\kappa)$ is the normalization constant involving the modified Bessel function of the first kind (detailed derivation is in Appendix).

Max-Margin Targeted Contrastive Loss. Under the vMF assumption, the expectation term required for the kernel density estimation admits a closed-form solution:

$$\mathbb{E}_{\mathbf{z}^c\sim\text{vMF}(\boldsymbol{\mu}_c,\kappa_c)}\left[\exp\left(\mathbf{z}^\top \mathbf{z}^c/\tau\right)\right] = C_d(\kappa_c)/C_d(\tilde{\kappa}_c), \quad (7)$$

where $\tilde{\kappa}_c = \|\kappa_c \boldsymbol{\mu}_c + \mathbf{z}/\tau\|_2$ denotes the refined concentration parameter. Essentially, Eq. (7) measures the alignment between $\mathbf{z}$ and the class-conditional distribution, depicting how likely $\mathbf{z}$ is under the vMF distribution of class c. Thus, we term this quantity the **Spherical Distribution Affinity** (**SDA**), denoted as $\mathcal{A}(\mathbf{z}, c)$. Substituting the closed-form solution into Eq. (5) yields the final MCL loss:

$$\mathcal{L}_{\text{MCL}}(\mathbf{z}_i, y_i, \pi) = -\log \frac{\pi(y_i)\mathcal{A}(\mathbf{z}_i, y_i)}{\sum_{c=1}^{C}\pi(c)\mathcal{A}(\mathbf{z}_i, c)}. \quad (8)$$

Online Parameter Estimation. To optimize the objective in Eq. (8), accurate estimation of each class's prior $\pi(c)$ and distribution parameters $(\boldsymbol{\mu}_c, \kappa_c)$ is essential. We adopt an online estimation strategy: For $\pi(c)$, we maintain separate statistics for source and target domains to account for label distribution shift, i.e., $\pi^S(c)$ is computed from ground-truth labels, while $\pi^T(c)$ is updated using pseudo-labels. For vMF parameters $(\boldsymbol{\mu}_c, \kappa_c)$, we assume a shared latent distribution across domains in the aligned contrastive space, and thus use unified centroids. Specifically, we maintain a momentum embedding $\bar{\mathbf{z}}^c$ per class c, updated via the EMA operation:

$$\bar{\mathbf{z}}^c_{(t)} = \beta \bar{\mathbf{z}}^c_{(t-1)} + (1-\beta)\bar{\mathbf{z}}'^c_{(t)}, \quad (9)$$

where $\bar{\mathbf{z}}'^c_{(t)}$ is the mean pixel features for class c via batch-wise updating and $\beta \in (0, 1)$ is the momentum coefficient. The

centroid is the normalization form: $\boldsymbol{\mu}_c = \bar{\mathbf{z}}^c / \|\bar{\mathbf{z}}^c\|_2$. Then we estimate the norm $R_c = \|\bar{\mathbf{z}}_c\|_2$ as a sufficient statistic for κ_c via the numerical approximation of [Sra, 2012]:

$$\kappa_c = R_c(d - R_c^2)/(1 - R_c^2). \tag{10}$$

3.3 Reliability-Boosted Self-Training

Although MCL constructs a discriminative feature space by enforcing strict margins, it cannot fully eliminate the pseudo-label noise caused by domain shift. In particular, this noise often manifests as class imbalance, causing pseudo-labels of tail classes to be overwhelmed by head classes. Therefore, as a complement to the feature space constraints, we propose the *Reliability-aware Filter (RaF)* to boost pseudo-label-based self-training. Specifically, inspired by the widely established small-loss criterion [Gui *et al.*, 2021] in noisy label learning, we utilize the contrastive loss and SDA calculated for each target pixel (indexed by set Ω) as the metrics for reliability.

First, we consider pixels exhibiting lower loss are more likely to be correctly classified. Considering that a global threshold is suboptimal [Zou *et al.*, 2018] due to the varying learning dynamics across classes and the inherent long-tailed distribution, we employ a class-wise adaptive strategy. Specifically, we define a dynamic threshold ξ_c for the loss metric, determined by the loss value at the ρ-th percentile of pixels predicted as class c within a recent temporal window. Given that tail classes may be scarce or entirely absent within a single mini-batch, we aggregate pseudo-labeled pixels across multiple steps and compute ξ_c once per window to ensure more reliable statistics, especially for tail classes. Consequently, we define the loss-derived reliable subset as a subset of $\Omega = \{i|1, \ldots, HW\}$, where HW quantifies the number of pixels in an image in the form of length and width, as:

$$\Omega_{r\text{-}L} = \{i \in \Omega \mid \mathcal{L}_{\text{MCL}}(\mathbf{z}_i) < \xi_{\hat{y}_i}\}, \tag{11}$$

where $\hat{y}_i$ is the pseudo-label of pixel embedding $\mathbf{z}_i$. ξ_c is the class-adaptive threshold, determined by the loss value at the ρ-th percentile of pixels predicted as class c.

Moreover, to fully exploit potentially reliable samples, we consider pixels with high SDA as reliable ones. Therefore, we also define the affinity-derived reliable subset as $\Omega_{r\text{-}A}$:

$$\Omega_{r\text{-}A} = \{i \in \Omega \mid \mathcal{A}(\mathbf{z}_i, \hat{y}_i) > \tau_{SDA}\}, \tag{12}$$

where $\hat{y}_i$ is the pseudo-label of pixel embedding $\mathbf{z}_i$ and τ_{SDA} is the fixed threshold. Ultimately, RaF separates the target pixels into a reliable subset Ω_r and an unreliable subset Ω_u as:

$$\Omega_r = \Omega_{r\text{-}L} \cup \Omega_{r\text{-}A}, \quad \Omega_u = \Omega \setminus \Omega_r. \tag{13}$$

To fully leverage the target pixels while mitigating potential noise, we apply differentiated supervision strategies for the identified reliable and unreliable subsets. Let $q_i = \sigma(F_\Theta(\mathbf{x}_i))$ denote the prediction from the teacher model, and $p_i = \sigma(F_\theta(\mathbf{x}_i))$ denote the prediction from the student model, we formulate distinct objectives for the separated sets:

$$\begin{aligned} \mathcal{L}_r &= \frac{1}{|\Omega_r|} \sum\nolimits_{i \in \Omega_r} \eta_{n(i)} \cdot D_{KL}(\hat{y}_i || p_i), \\ \mathcal{L}_u &= \frac{1}{|\Omega_u|} \sum\nolimits_{i \in \Omega_u} \eta_{n(i)} \cdot D_{KL}(\text{Sharp}(q_i, \tau_{shp}) || p_i), \end{aligned} \tag{14}$$

where $\text{Sharp}(q_i, \tau_{\text{shp}}) = \frac{(q_i^c)^{1/\tau_{\text{shp}}}}{\sum_{c=1}^{C} (q_i^c)^{1/\tau_{\text{shp}}}}$ for class c and τ_{shp} is the sharpening temperature. Moreover, to mitigate the inherent pseudo-label noise, the weighting factor $\eta_{n(i)}$ computes the pixel ratio in $\mathbf{x}_i$'s image $X_{n(i)}^T$ whose maximum class confidence exceeds the pre-defined threshold τ_{conf} as:

$$\eta_{n(i)} = \frac{1}{HW} \sum\nolimits_{j=1}^{HW} \mathbb{I}\left(\max_c \left[F_\Theta(X_{n(i)}^T) \right]^{(j,c)} > \tau_{\text{conf}} \right). \tag{15}$$

Overall, by leveraging the MCL-derived metrics to identify reliable and unreliable pixels, RaF boosts pseudo-label-based supervision for segmentation training.

3.4 Overall Objective

The total objective function of MarCon is a weighted combination of segmentation losses and contrastive regularization terms across both domains. We minimize, for the source domain $\mathcal{D}_s$, the supervised cross-entropy loss $\mathcal{L}_{\text{seg}}^s$ and the contrastive loss $\mathcal{L}_{\text{MCL}}^s$. For the target domain $\mathcal{D}_t$, we use the self-training loss $\mathcal{L}_{\text{seg}}^t = \mathcal{L}_r + \mathcal{L}_u$ and the contrastive loss $\mathcal{L}_{\text{MCL}}^t$. The overall objective is:

$$\mathcal{L} = \mathcal{L}_{\text{seg}}^s + \mathcal{L}_{\text{seg}}^t + \lambda_c(\mathcal{L}_{\text{MCL}}^s + \mathcal{L}_{\text{MCL}}^t), \tag{16}$$

where λ_c is a trade-off factor. By jointly optimizing these losses, MarCon simultaneously achieves discriminative pixel classification and robust domain-invariant feature alignment.

4 Experiments

4.1 Experimental Setup

Datasets. Following the standard protocol [Hoyer *et al.*, 2023; Khan *et al.*, 2025] in UDA-SS, we evaluate our method on two widely used benchmarks: GTA $\rightarrow$ Cityscapes and SYNTHIA $\rightarrow$ Cityscapes, both of which involve adaptation from synthetic to real-world urban scenes. The source domain GTA [Richter *et al.*, 2016] consists of 24,966 synthetic images automatically extracted from the video game *Grand Theft Auto V*. The SYNTHIA [Ros *et al.*, 2016] dataset provides 9,400 synthetic images. The target domain Cityscapes [Cordts *et al.*, 2016] contains 2,975 training and 500 validation images of real street scenes.

Implementation Details. We implement our method using the MMSegmentation framework. We adopt the same architecture as DAFormer with a MiT-B5 [Xie *et al.*, 2021] backbone, which is pre-trained on ImageNet. The models are trained using the AdamW optimizer with a weight decay of 0.01. The initial learning rates are set to 6×10^{-5} for the encoder and 6×10^{-4} for the decoder. We apply a linear learning rate warm-up for the first 1.5k iterations, and linear decay afterwards. We set the teacher EMA coefficient $\alpha = 0.999$ and SDA threshold $\tau_{SDA} = 0.968$ following DAFormer. Besides, we set MCL loss weight $\lambda_c = 0.4$ and MCL temperature $\tau = 0.07$. Following the typical methods DAFormer and HRDA, we adopt random rescaling and cropping to generate input image with resolutions of 512×512 and 1024×1024, respectively. Accordingly, training is performed on one or two NVIDIA A5000 GPUs.

Method	Road	Sidewalk	Building	Wall	Fence	Pole	Light	Sign	Veg	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	Motor	Bike	mIoU
CRST [Zou *et al.*, 2019]	91.7	45.1	80.9	29.0	23.4	43.8	47.1	40.9	84.0	20.0	60.6	64.0	31.9	85.8	39.5	48.7	25.0	38.0	47.0	49.8
PLCA [Kang *et al.*, 2020]	84.0	30.4	82.4	35.3	24.8	32.2	36.8	24.5	85.5	37.2	78.6	66.9	32.8	85.5	40.4	48.0	8.8	29.8	41.8	47.7
DACS [Tranheden *et al.*, 2021]	93.0	52.0	87.8	29.4	38.3	37.7	45.0	53.3	87.9	46.3	90.2	67.8	38.0	89.0	51.1	51.1	0.0	10.7	19.4	52.1
ProDA [He *et al.*, 2021]	87.8	56.0	79.7	46.3	44.8	45.6	53.5	53.5	88.6	45.2	82.1	70.7	39.2	88.8	45.5	50.4	1.0	48.9	56.4	57.5
CPSL [Li *et al.*, 2022]	92.3	59.5	84.9	45.7	29.7	52.8	61.5	59.5	87.9	41.6	85.0	73.0	35.5	90.4	48.7	73.9	26.3	53.8	53.9	60.8
DAFormer [Hoyer *et al.*, 2022a]	95.7	70.2	89.4	53.5	48.1	49.6	55.8	59.4	89.9	47.9	92.5	72.2	44.7	92.3	74.5	78.2	65.1	55.9	61.8	68.3
+Ours	95.4	69.7	**90.0**	**56.9**	**49.8**	**51.3**	**58.0**	**64.7**	**90.3**	**49.5**	92.4	**72.7**	**47.8**	**93.2**	**80.9**	**85.4**	**75.4**	53.0	**66.2**	**70.7**
HRDA [Hoyer *et al.*, 2022b]	96.5	74.4	91.0	61.6	51.5	57.1	63.9	69.3	91.3	48.4	94.2	79.0	52.9	93.9	84.1	85.7	75.9	63.9	67.5	73.8
+Ours	96.4	**74.5**	**91.4**	60.7	**56.8**	57.1	**66.0**	**72.8**	**91.7**	**54.5**	**94.3**	**79.2**	51.4	**94.8**	**87.2**	**87.6**	75.8	**65.6**	**69.1**	**75.1**
MIC [Hoyer *et al.*, 2023]	97.4	80.1	91.7	61.2	56.9	59.7	66.0	71.3	91.7	51.4	94.3	79.8	56.1	94.6	85.4	90.3	80.4	64.5	68.5	75.9
+CoPT [Mata *et al.*, 2024]	97.6	80.9	91.6	62.1	55.9	59.3	66.7	70.5	91.9	53.0	94.4	80.0	55.6	94.7	87.1	88.6	82.1	65.0	68.8	76.1
+AFR* [Khan *et al.*, 2025]	97.6	81.1	91.7	62.2	57.5	60.7	64.8	75.4	91.6	51.7	93.4	79.7	56.4	94.7	87.0	89.6	80.7	65.5	68.8	76.3
+Ours	**97.6**	**80.6**	**92.0**	**64.3**	**60.2**	**60.2**	**68.3**	**73.2**	**92.0**	**53.6**	94.0	**81.3**	**59.0**	**95.0**	**86.5**	**91.5**	**83.2**	**65.9**	**69.5**	**77.3**

Table 1: UDA-SS performance (mIoU) on GTA.→CS. We mark the improvement by our MarCon as **bold**. $*$ denotes the reproduced result.

Method	Road	Sidewalk	Building	Wall	Fence	Pole	Light	Sign	Veg	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	Motor	Bike	mIoU
CRST [Zou *et al.*, 2019]	67.7	32.2	73.9	10.7	1.6	37.4	22.2	31.2	80.8	-	80.5	60.8	29.1	82.8	-	25.0	-	19.4	45.3	43.8
PLCA [Kang *et al.*, 2020]	82.6	29.0	81.0	11.2	0.2	33.6	24.9	18.3	82.8	-	82.3	62.1	26.5	85.6	-	48.9	-	26.8	52.2	46.8
DACS [Tranheden *et al.*, 2021]	80.6	25.1	81.9	21.5	2.9	37.2	22.7	24.0	83.7	-	90.8	67.6	38.3	82.9	-	38.9	-	28.5	47.6	48.3
ProDA [He *et al.*, 2021]	87.8	45.7	84.6	37.1	0.6	44.0	54.6	37.0	88.1	-	84.4	74.2	24.3	88.2	-	51.1	-	40.5	45.6	55.5
CPSL [Li *et al.*, 2022]	87.2	43.9	85.5	33.6	0.3	47.7	57.4	37.2	87.8	-	88.5	79.0	32.0	90.6	-	49.4	-	50.8	59.8	57.9
DAFormer [Hoyer *et al.*, 2022a]	84.5	40.7	88.4	41.5	6.5	50.0	55.0	54.6	86.0	-	89.8	73.2	48.2	87.2	-	53.2	-	53.9	61.7	60.9
+Ours	**87.1**	**45.7**	**88.7**	**47.4**	**7.1**	**52.6**	53.3	**56.9**	**88.2**	-	**92.3**	**73.8**	**49.7**	**87.8**	-	**60.2**	-	53.9	**62.9**	**63.0**
HRDA [Hoyer *et al.*, 2022b]	85.2	47.7	88.8	49.5	4.8	57.2	65.7	60.9	85.3	-	92.9	79.4	52.8	89.0	-	64.7	-	63.9	64.9	65.8
+Ours	**88.9**	**50.9**	**90.4**	**51.4**	**7.3**	**57.6**	**66.9**	60.8	**88.3**	-	**94.5**	**80.3**	**57.0**	**90.2**	-	64.2	-	**64.3**	64.8	**67.4**
MIC [Hoyer *et al.*, 2023]	86.6	50.5	89.3	47.9	7.8	59.4	66.7	63.4	87.1	-	94.6	81.0	58.9	90.1	-	61.9	-	67.1	64.3	67.3
+CoPT [Mata *et al.*, 2024]	83.4	44.3	90.0	50.4	8.0	60.0	67.0	63.0	87.5	-	94.8	81.1	58.6	89.7	-	66.5	-	68.9	65.0	67.4
+AFR* [Khan *et al.*, 2025]	88.5	54.8	88.9	48.4	8.7	59.3	66.4	63.7	87.4	-	93.9	81.8	59.2	91.2	-	66.7	-	68.4	65.4	68.3
+Ours	**89.9**	**56.6**	**89.9**	**49.0**	**9.4**	**61.3**	**69.9**	63.4	**89.0**	-	**94.7**	**81.6**	**61.3**	90.1	-	**67.7**	-	**69.4**	**66.8**	**69.4**

Table 2: UDA-SS performance (mIoU) on SYN.→CS. We mark the improvement by our MarCon as **bold**. $*$ denotes the reproduced result.

4.2 Experimental Results

This section presents quantitative results on GTA. → CS. and SYN. → CS., along with qualitative visualizations.

GTA → Cityscapes. As shown in Table 1, we compare our method with several current leading approaches on the GTA → Cityscapes benchmark. Specifically, we integrate our approach into multiple Transformer-based models and consistently achieve performance improvements: a gain of 2.4 mIoU with DAFormer, 1.3 mIoU with HRDA and 1.4 mIoU with MIC. Notably, our method achieves improvements over baselines on some tail classes, such as Fence and Bike. It demonstrates that our proposed MarCon effectively mitigates the bias induced by long-tailed distributions.

SYNTHIA → Cityscapes. As shown in Table 2, MarCon also demonstrates strong performance on the SYNTHIA → Cityscapes benchmark. Compared to the baseline DAFormer and HRDA, our method improves mIoU by 2.1 and 1.6, respectively. Furthermore, when integrated into MIC, MarCon yields consistent gains of 2.1 mIoU, highlighting its generalizability across different transformer-based architectures.

Qualitative results. To intuitively demonstrate the superiority of our proposed method, we visualize the segmentation maps and compare them with the strong baseline, DAFormer, on the GTA → Cityscapes benchmark, as shown in Fig 3. The baseline model struggles to handle complex scenarios, as indicated by the red dashed boxes. Specifically, in the first row, DAFormer fails to distinguish the reflective building from the sky due to appearance ambiguity. In the second row, it misses fine-grained details like the small pole and exhibits poor internal consistency on large objects (e.g., the truck). In the third row, the baseline misclassifies a building with unique textures. By integrating our method into DAFormer, we effectively alleviate these failure cases. As highlighted by the green dashed boxes, our approach successfully rectifies the misclassified regions and produces predictions that are highly consistent with the Ground Truth.

GTA. → CS.			SYN. → CS.		
MCL	RaF	mIoU	MCL	RaF	mIoU
-	-	68.3	-	-	60.9
✓	-	70.2	✓	-	62.6
✓	✓	70.7	✓	✓	63.0

Table 3: Ablation studies on GTA. → CS. and SYN. → CS.

4.3 Ablation Studies

Quantitative Ablation Results. To investigate the contribution of individual modules, we conduct ablation studies on GTA → Cityscapes and SYNTHIA → Cityscapes benchmarks. As reported in Table 3, the progressive integration of our proposed components yields consistent performance improvements.The baseline achieves 68.3 and 60.9 mIoU. Applying MCL loss boosts the performance to 70.2 and 62.6, mIoU respectively. Building upon this, incorporating RaF to

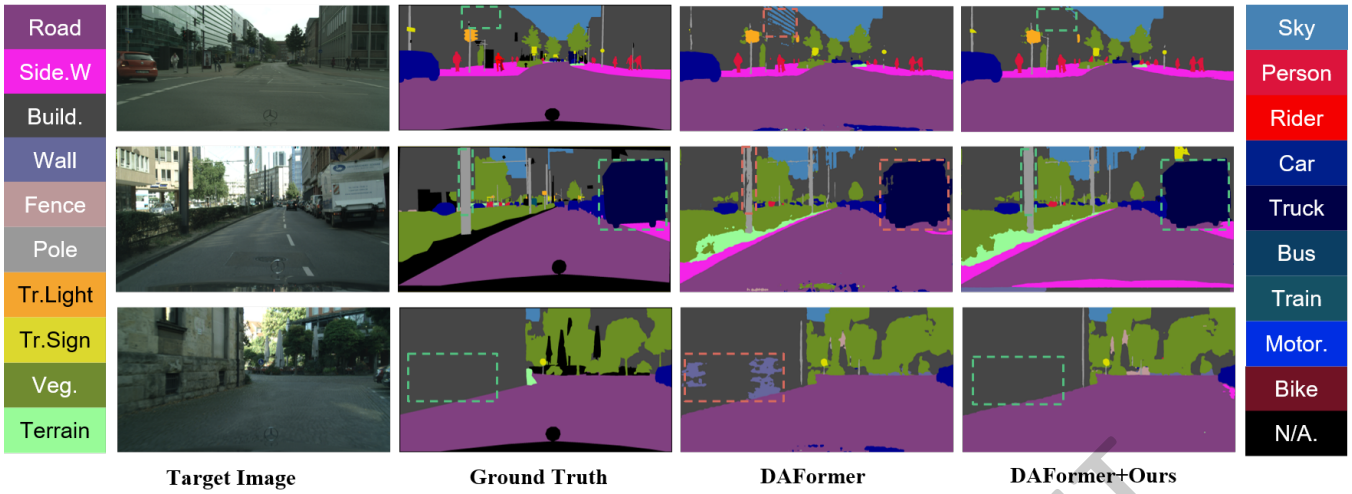


Figure 3: Qualitative comparisons on the GTA → Cityscapes benchmark. We compare the segmentation maps generated by the baseline DAFormer and our method. Red dashed boxes highlight failure cases in the baseline predictions, while green dashed boxes demonstrate that our method successfully corrects these errors, yielding predictions that are well-aligned with the Ground Truth.

ensure cleaner supervision further lifts the mIoU to 70.7 and 63.0. This verifies the effectiveness of proposed modules.

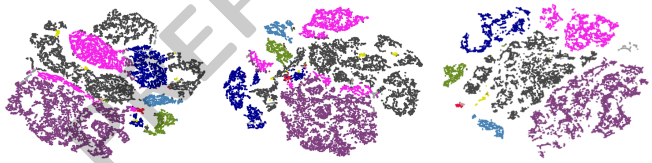
Effect of MCL. To evaluate MCL's effect, we extract pixel features from a target image using baseline DAFormer, Prototype-based CL (PCL) method, and our proposed MarCon. The corresponding t-SNE visualizations are presented in Fig 4. The baseline DAFormer exhibits ambiguous decision boundaries due to its sole reliance on CE supervision, while PCL improves separation via pixel-prototype contrast. However, our **MarCon** achieves the most distinct boundaries and highly compact clusters. This confirms our vMF-based modeling explicitly optimizes for the maximum margin.

Unleashing the Potential of FAF with MCL. As posited in Sec 3.1, we hypothesize that MCL can amplify the efficacy of the FAF module. To verify this, we conduct a detailed ablation study in Fig. 5a. We plot the MCL+FAF (Theo.) curve (grey dashed line) to represent the hypothetical performance assuming the gains from FAF and MCL are independent and simply additive. Observing the results, the actual combined MCL+FAF (Comb.) curve (red line) outperforms this theoretical additive baseline. This margin between the red and grey curves empirically validates our claim. It demonstrates that by constructing a discriminative feature space, MCL effectively mitigates the adverse impact of domain shift on FAF, thereby amplifying the benefits yielded by FAF.

Effect of RaF. Finally, to assess the impact of RaF, we also track the mIoU of generated pseudo-labels on a target image subset across different training stages. As depicted in Fig 5b, the integration of RaF yields higher pseudo-label quality compared to the baseline by mitigating the noisy supervision of the target domain pseudo label.

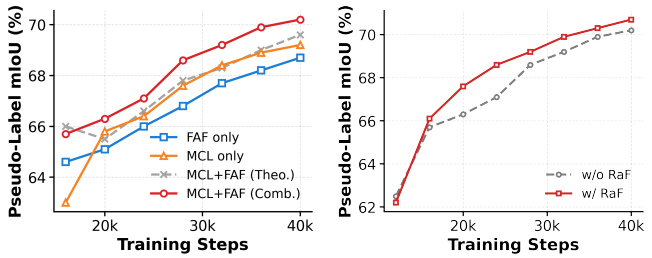
5 Conclusion

In this paper, we presented MarCon, a framework designed to tackle the inherent long-tailed class distribution in Unsupervised Domain Adaptation Semantic Segmentation, thereby mitigating the consequent feature collapse and pseudo-label noise. Specifically, we focus on breaking the vicious cycle where poor tail-class representation and noisy pseudo-labels mutually degrade adaptation performance. To this end, MarCon introduces a synergistic approach combining representation modeling and reliability assessment. Specifically, our Max-Margin Contrastive Learning (MCL) reformulates the feature space using von Mises-Fisher distributions to enforce strict margins, while the Reliability-aware Filter (RaF) dynamically mitigates the impact of unreliable supervision. Extensive experiments on standard benchmarks verify that MarCon consistently outperforms current leading methods.



(a) DAFormer-based. (b) PCL-based. (c) MarCon-based.

Figure 4: t-SNE of features extracted by multiple methods.



(a) MCL amplifies gain of FAF. (b) Impact of RaF.

Figure 5: Effects of our proposed components.

Acknowledgements

This paper is supported by the Ningbo Yongjiang Talent Introduction Programme (No. 2023A-399-G), CCF-NetEase ThunderFire Innovation Research Funding (NO.202516) and the Fundamental Research Funds for the Central Universities (226-2025-00065).

References

[Chen *et al.*, 2018] 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 Trans. Pattern Anal. Mach. Intell.*, pages 834–848, 2018.

[Chen *et al.*, 2020] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *Proc. of ICML*, pages 1597–1607, 2020.

[Chen *et al.*, 2023] Mu Chen, Zhedong Zheng, Yi Yang, and Tat-Seng Chua. Pipa: Pixel- and patch-wise self-supervised learning for domain adaptative semantic segmentation. In *Proc. of ACM MM*, pages 1905–1914, 2023.

[Chen *et al.*, 2024a] Linwei Chen, Ying Fu, Lin Gu, Chenggang Yan, Tatsuya Harada, and Gao Huang. Frequency-aware feature fusion for dense image prediction. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 10763–10780, 2024.

[Chen *et al.*, 2024b] Mu Chen, Zhedong Zheng, and Yi Yang. Transferring to real-world layouts: A depth-aware framework for scene adaptation. In *Proc. of ACM MM*, pages 399–408, 2024.

[Cordts *et al.*, 2016] 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 *Proc. of CVPR*, pages 3213–3223, 2016.

[Fan *et al.*, 2024] Qizhe Fan, Xiaoqin Shen, Shihui Ying, and Shaoyi Du. OTCLDA: optimal transport and contrastive learning for domain adaptive semantic segmentation. *IEEE Trans. Intell. Transp. Syst.*, pages 14685–14697, 2024.

[Ghiasi *et al.*, 2021] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D. Cubuk, Quoc V. Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proc. of CVPR*, pages 2918–2928, 2021.

[Gui *et al.*, 2021] Xian-Jin Gui, Wei Wang, and Zhang-Hao Tian. Towards understanding deep learning from noisy labels with small-loss criterion. In *Proc. of IJCAI*, pages 2469–2475, 2021.

[Gupta *et al.*, 2019] Agrim Gupta, Piotr Dollár, and Ross B. Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *Proc. of CVPR*, 2019.

[He *et al.*, 2021] Ruifei He, Jihan Yang, and Xiaojuan Qi. Re-distributing biased pseudo labels for semi-supervised semantic segmentation: A baseline investigation. In *Proc. of ICCV*, pages 6910–6920, 2021.

[Hoffman *et al.*, 2018] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A. Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *Proc. of ICML*, pages 1994–2003, 2018.

[Hoyer *et al.*, 2022a] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *Proc. of CVPR*, pages 9914–9925, 2022.

[Hoyer *et al.*, 2022b] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. HRDA: context-aware high-resolution domain-adaptive semantic segmentation. In *Proc. of ECCV*, pages 372–391, 2022.

[Hoyer *et al.*, 2023] Lukas Hoyer, Dengxin Dai, Haoran Wang, and Luc Van Gool. MIC: masked image consistency for context-enhanced domain adaptation. In *Proc. of CVPR*, pages 11721–11732, 2023.

[Jiang *et al.*, 2022] Zhengkai Jiang, Yuxi Li, Ceyuan Yang, Peng Gao, Yabiao Wang, Ying Tai, and Chengjie Wang. Prototypical contrast adaptation for domain adaptive semantic segmentation. In *Proc. of ECCV*, pages 36–54, 2022.

[Kang *et al.*, 2020] Guoliang Kang, Yunchao Wei, Yi Yang, Yueting Zhuang, and Alexander G. Hauptmann. Pixel-level cycle association: A new perspective for domain adaptive semantic segmentation. In *Proc. of NeurIPS*, 2020.

[Khan *et al.*, 2025] Md. Al-Masrur Khan, Durgakant Pushp, and Lantao Liu. AFRDA: attentive feature refinement for domain adaptive semantic segmentation. *IEEE Robotics Autom. Lett.*, pages 9573–9580, 2025.

[Lee *et al.*, 2021] Suhyeon Lee, Junhyuk Hyun, Hongje Seong, and Euntai Kim. Unsupervised domain adaptation for semantic segmentation by content transfer. In *Proc. of AAAI*, pages 8306–8315, 2021.

[Li and Li, 2021] Wei Li and Zhixin Li. Domain adaptative semantic segmentation by alleviating long-tail problem. In *Proc. of IJCNN*, pages 1–8, 2021.

[Li *et al.*, 2022] Ruihuang Li, Shuai Li, Chenhang He, Yabin Zhang, Xu Jia, and Lei Zhang. Class-balanced pixel-level self-labeling for domain adaptive semantic segmentation. In *Proc. of CVPR*, pages 11583–11593, 2022.

[Li *et al.*, 2023] Tianyu Li, Subhankar Roy, Huayi Zhou, Hongtao Lu, and Stéphane Lathuilière. Contrast, stylize and adapt: Unsupervised contrastive learning framework for domain adaptive semantic segmentation. In *Proc. of CVPR*, pages 4869–4879, 2023.

[Li *et al.*, 2025a] Wangkai Li, Rui Sun, Bohao Liao, Zhaoyang Li, and Tianzhu Zhang. Balanced learning for domain adaptive semantic segmentation. In *Proc. of ICML*, 2025.

[Li *et al.*, 2025b] Wangkai Li, Rui Sun, Huayu Mai, and Tianzhu Zhang. Towards unsupervised domain bridging via image degradation in semantic segmentation. In *Proc. of NeurIPS*, 2025.

[Long *et al.*, 2015] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proc. of CVPR*, pages 3431–3440, 2015.

[Luo *et al.*, 2022a] Cheng Luo, Qinliang Lin, Weicheng Xie, Bizhu Wu, Jinheng Xie, and Linlin Shen. Frequency-driven imperceptible adversarial attack on semantic similarity. In *Proc. of CVPR*, pages 15294–15303, 2022.

[Luo *et al.*, 2022b] Yawei Luo, Ping Liu, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Category-level adversarial adaptation for semantic segmentation using purified features. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 3940–3956, 2022.

[Mata *et al.*, 2024] Cristina Mata, Kanchana Ranasinghe, and Michael S. Ryoo. Copt: Unsupervised domain adaptive segmentation using domain-agnostic text embeddings. In *Proc. of ECCV*, pages 424–440, 2024.

[Ren *et al.*, 2024] Qinghua Ren, Qirong Mao, and Shijian Lu. Prototypical bidirectional adaptation and learning for cross-domain semantic segmentation. *IEEE Trans. Multim.*, pages 501–513, 2024.

[Richter *et al.*, 2016] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *Proc. of ECCV*, pages 102–118, 2016.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III*, pages 234–241, 2015.

[Ros *et al.*, 2016] Germán Ros, Laura Sellart, Joanna Materzynska, David Vázquez, and Antonio M. López. The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proc. of CVPR*, pages 3234–3243, 2016.

[Sra, 2012] Suvrit Sra. A short note on parameter approximation for von mises-fisher distributions: and a fast implementation of $I_s(x)$. *Comput. Stat.*, pages 177–190, 2012.

[Toldo *et al.*, 2020] Marco Toldo, Andrea Maracani, Umberto Michieli, and Pietro Zanuttigh. Unsupervised domain adaptation in semantic segmentation: a review. *CoRR*, 2020.

[Tranheden *et al.*, 2021] Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. DACS: domain adaptation via cross-domain mixed sampling. In *Proc. of WACV*, pages 1378–1388, 2021.

[Vayyat *et al.*, 2022] Midhun Vayyat, Jaswin Kasi, Anuraag Bhattacharya, Shuaib Ahmed, and Rahul Tallamraju. CLUDA : Contrastive learning in unsupervised domain adaptation for semantic segmentation. *CoRR*, 2022.

[Vu *et al.*, 2019] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. ADVENT: adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proc. of CVPR*, pages 2517–2526, 2019.

[Wang *et al.*, 2021] Jiaqi Wang, Wenwei Zhang, Yuhang Zang, Yuhang Cao, Jiangmiao Pang, Tao Gong, Kai Chen, Ziwei Liu, Chen Change Loy, and Dahua Lin. Seesaw loss for long-tailed instance segmentation. In *Proc. of CVPR*, pages 9695–9704, 2021.

[Xie *et al.*, 2021] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, José M. Álvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Proc. of NeurIPS*, pages 12077–12090, 2021.

[Xie *et al.*, 2023] Binhui Xie, Shuang Li, Mingjia Li, Chi Harold Liu, Gao Huang, and Guoren Wang. Sepico: Semantic-guided pixel contrast for domain adaptive semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 9004–9021, 2023.

[Zhang *et al.*, 2018] Hang Zhang, Kristin J. Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *Proc. of CVPR*, pages 7151–7160, 2018.

[Zhang *et al.*, 2021] 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 *Proc. of CVPR*, pages 12414–12424, 2021.

[Zhao *et al.*, 2024] Xingchen Zhao, Niluthpol Chowdhury Mithun, Abhinav Rajvanshi, Han-Pang Chiu, and Supun Samarasekera. Unsupervised domain adaptation for semantic segmentation with pseudo label self-refinement. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2024, Waikoloa, HI, USA, January 3-8, 2024*, pages 2388–2398, 2024.

[Zheng *et al.*, 2021] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H. S. Torr, and Li Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proc. of CVPR*, pages 6881–6890, 2021.

[Zou *et al.*, 2018] Yang Zou, Zhiding Yu, B. V. K. Vijaya Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proc. of ECCV*, pages 297–313, 2018.

[Zou *et al.*, 2019] Yang Zou, Zhiding Yu, Xiaofeng Liu, B. V. K. Vijaya Kumar, and Jinsong Wang. Confidence regularized self-training. In *Proc. of ICCV*, pages 5981–5990, 2019.