Energy-Based Domain Adaptation Without Intermediate Domain Dataset for Foggy Scene Segmentation

Donggon Jang, Sunhyeok Lee, Gyuwon Choi, Yejin Lee, Sanghyeok Son, and Dae-Shik Kim, *Member, IEEE*

***Abstract*— Robust segmentation performance under dense fog is crucial for autonomous driving, but collecting labeled real foggy scene datasets is burdensome in the real world. To this end, existing methods have adapted models trained on labeled clear weather images to the unlabeled real foggy domain. However, these approaches require intermediate domain datasets (e.g. synthetic fog) and involve multi-stage training, making them cumbersome and less practical for real-world applications. In addition, the issue of overconfident pseudo-labels by a confidence score remains less explored in self-training for foggy scene adaptation. To resolve these issues, we propose a new framework, named DAEN, which Directly Adapts without additional datasets or multi-stage training and leverages an ENergy score in self-training. Notably, we integrate a High-order Style Matching (HSM) module into the network to match high-order statistics between clear weather features and real foggy features. HSM enables the network to implicitly learn complex fog distributions without relying on intermediate domain datasets or multi-stage training. Furthermore, we introduce Energy Score-based Pseudo-Labeling (ESPL) to mitigate the overconfidence issue of the confidence score in self-training. ESPL generates more reliable pseudo-labels through a pixel-wise energy score, thereby alleviating bias and preventing the model from assigning pseudo-labels exclusively to head classes. Extensive experiments demonstrate that DAEN achieves state-of-the-art performance on three real foggy scene datasets and exhibits a generalization ability to other adverse weather conditions. Code is available at https://github.com/jdg900/daen**

***Index Terms*— Unsupervised domain adaptation, foggy scene segmentation, energy score, features statistics matching.**

I. Introduction

SEMANTIC segmentation is a fundamental task in computer vision, with utmost importance in real-world scenarios such as autonomous driving and robotics. While semantic segmentation demonstrates commendable performance in normal weather conditions, it encounters significant performance degradation under poor visibility conditions such as dense fog. The straightforward way is to establish the labeled dataset for foggy scene segmentation. However, collecting a huge set of images and pixel-wise annotations for foggy scenes is extremely challenging and burdensome.

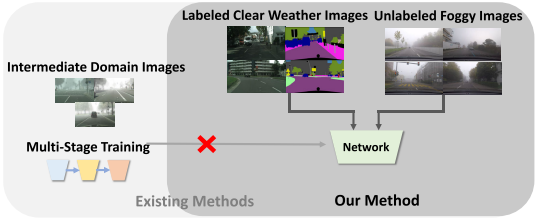

Fig. 1. Existing methods require additional intermediate domain datasets or multi-stage training. However, our method directly adapts the network without the intermediate domain datasets or multi-stage training.

To mitigate the difficulty of collecting foggy images and a scarcity of ground-truth labels, several unsupervised domain adaptation methods [1], [2], [3], [4] have been studied to adapt a model trained on labeled clear weather domain to an unlabeled real foggy domain. However, some of these methods [1], [2], [3] require an additional intermediate domain dataset (e.g. synthetic fog) or multi-stage training to alleviate the substantial domain discrepancy between the clear weather and the real foggy domains, which is cumbersome in real-world scenarios. To address the above limitation, self-training, which has shown promising performance in conventional unsupervised domain adaptation, can be an alternative for direct adaptation to foggy scenes. The self-training approaches [5], [6], [7] use a model trained on the source domain to predict pseudo-labels for the target domain. This process leverages a supervised signal with the predicted pseudo-labels to guide the model in acquiring knowledge from the target domain. However, a notable challenge in self-training lies in the potential accumulation of errors stemming from noisy predictions. To prevent error accumulation caused by noisy predictions, confidence score-based pseudo-labeling assigns pseudo-labels only to high-confident pixels by filtering out unreliable pixels. In self-training for foggy scene adaptation, existing works [3], [4] utilize pseudo-labels produced by confidence score thresholding. They propose pseudo label diffusion [3] and candidate label set [4] in an attempt to obtain high-quality pseudo-labels. Nevertheless, it is important to note that confidence score-based pseudo-labeling tends to generate overly confident

Received 8 December 2023; revised 24 September 2024; accepted 14 October 2024. Date of publication 24 October 2024; date of current version 29 October 2024. The associate editor coordinating the review of this article and approving it for publication was Prof. Hichem Sahbi. *(Corresponding author: Dae-Shik Kim.)*

The authors are with the School of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, Republic of Korea (e-mail: jdg900@kaist.ac.kr; lee.sunhyeok@kaist.ac.kr; gyuwonchoi@kaist.ac.kr; yejin@kaist.ac.kr; ssh816@kaist.ac.kr; daeshik@kaist.ac.kr).

Digital Object Identifier 10.1109/TIP.2024.3483566

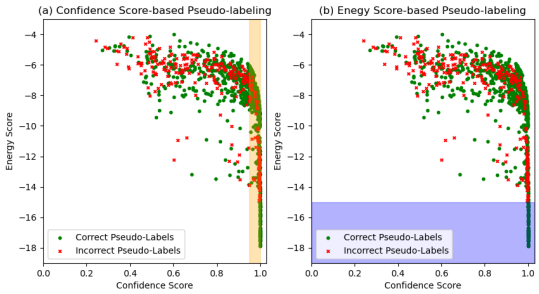

Fig. 2. **X-axis:** Confidence Score, **Y-axis:** Energy Score, **Left:** The orange-shaded region shows the pixel-wise pseudo-labels assigned by the confidence score. As observed, most pseudo-labels generated by the confidence score are overconfident and include many incorrect labels. **Right:** The blue-shaded region displays the pseudo-labels generated by the energy score. As shown, energy score-based pseudo-labeling effectively filters out false positives, providing more reliable pseudo-labels. These results are obtained from a sampled image in Foggy Zurich-test (FZ) [8].

pseudo-labels for unlabeled target images. As shown in Fig. 2 left, the pseudo-labels generated by the confidence score are overconfident and could include both correct and incorrect labels. This poses a potential risk of conveying inaccurate target domain knowledge to the model.

To address these issues, we propose a new framework named DAEN, which Directly Adapts without an additional intermediate domain dataset and multi-stage training, and leverages ENergy score in self-training for foggy scene segmentation. Our method eliminates the need for the intermediate domain dataset and multi-stage training and mitigates unreliable pseudo-labels caused by overconfidence in self-training. Firstly, we propose High-order Style Matching (HSM) to relieve the large domain gap without relying on an extra intermediate domain dataset. By considering the relationship between feature statistics and style, HSM augments clear weather features to emulate the real fog style through high-order feature statistics matching. HSM leverages high-order feature statistics to capture the real fog style that is complex to model through low-order statistics, such as mean and standard deviation, enabling the network to utilize intermediate domain features during training implicitly. Secondly, we propose Energy Score-based Pseudo-Labeling (ESPL) to mitigate the issue of overconfident pseudo-labels generated by confidence score. ESPL determines whether to assign a pseudo-label to each pixel by utilizing pixel-wise energy scores which assess whether a pixel belongs to the inlier or outlier distribution. As illustrated in Fig. 2 right, ESPL effectively filters out false positives, providing more reliable pseudo-labels. This helps the model learn more accurate target domain knowledge.

Extensive experiments show that the proposed DAEN outperforms existing methods on various real foggy domain datasets. Moreover, DAEN exhibits a generalization capability to other adverse weather conditions. The main contributions of this paper are summarized as follows:

- We propose High-order Style Matching (HSM) to obtain intermediate domain features, which allows the network to learn effectively complex real fog style without an additional intermediate domain dataset and multi-stage training.
- We propose Energy Score-based Pseudo-Labeling (ESPL) to alleviate the overconfidence issue in self-training and produce more reliable pseudo-labels.
- The proposed DAEN not only outperforms existing domain adaptation methods for foggy scenes but also achieves generalization capability on rainy and snow scenes.

II. Related Work

A. UDA for Semantic Segmentation

Unsupervised Domain Adaptation (UDA) aims to transfer knowledge learned from a labeled source domain to an unlabeled target domain. Early UDA methods align the source and target distributions at the input- [9], [10], [11], feature- [12], [13], [14], and output-level [6], [15] with adversarial training. Recently, self-training methods [5], [6], [7] have been proposed to overcome the training instability of adversarial training. These methods obtain pseudo-labels for target samples, either online [16], [17], [18] or offline [6], [7]. These pseudo-labels are utilized to allow the network to learn target domain knowledge in a supervised manner, which shows significant performance improvement in UDA. While early works have mainly focused on synthetic-to-real domain adaptation, recent research has drawn attention to clear-to-adverse weather domain adaptation [1], [2], [3], [4], [19], [20], [21], [22], [23], [24]. In this paper, we focus on adaptation from the clear weather domain to real foggy domain.

B. Semantic Foggy Scene Understanding

The initial work for semantic foggy scene understanding is SFSU [25]. SFSU synthesizes a simulated fog to clear the Cityscapes dataset [26] and trains the network using a labeled synthetic fog dataset. However, this method shows limited performance in real fog conditions due to the domain gap between synthetic fog and real fog. To address this limitation, CMAda [27] performs model adaptation from light synthetic fog to dense real fog in an easy-to-hard manner through curriculum learning. FIFO [1] introduces a fog pass filter to narrow the gap between images under different fog conditions (synthetic fog and real fog), facilitating adaptation under real foggy scenes. CuDA-Net [2] defines a dual domain gap (style and fog), progressively reducing the cumulative dual gap via a curriculum multi-stage training scheme. Recently, influenced by the promising solution of self-training in domain adaptation, some methods [3], [4] utilizing pseudo labels for real foggy domain adaptation have also been proposed.

C. Style Representation

The task of transferring the style of one image to another can be viewed as a distribution matching problem [28], [29], [30], [31], [32] in the feature space. Gatys et al. [29] propose to use the gram matrix as style representation and align the gram matrix between the style and content images. AdaIN [30] suggests that image style can be represented by first- and second-order feature statistics, and performs style transfer by re-normalizing the feature maps with channel-wise mean and

standard deviation of style image. However, using low-order feature statistics to represent style in real-world scenes would lead to inappropriate feature matching due to the complexity of the real-world distributions. Recently, several works [33], [34] have explored high-order central moments or histogram matching to achieve more accurate matching. In this paper, we aim to use high-order feature statistics to represent the style of clear weather and real foggy domains.

D. Energy-Based Models (EBM)

LeCun et al. [35] demonstrate the relationship between discriminative machine learning models and energy-based models. Based on the energy-based models, we can calculate the energy score for input samples. These scores provide a criterion for determining whether the samples belong to the inlier or outlier. Inspired by this characteristic, there have recently been attempts to apply EBM to deep learning [36], [37], [38]. In the out-of-distribution (OOD) detection task, Liu et al. [37] utilize the energy score to detect the OOD samples. Tian et al. [39] adopt a pixel-wise energy score to learn inlier pixel distributions and detect anomaly objects in urban-scene semantic segmentation. Moreover, there have been recent endeavors [40], [41], [42], [43] in leveraging EBM for domain adaptation. In [40], they employ the energy score as a regularization term, akin to entropy minimization, for the self-training in Unsupervised Domain Adaptation (UDA). Xie et al. [41] constrain the free energy score of the target domain to be closer to that of the source domain by introducing a novel free energy alignment loss for active domain adaptation. Li et al. [42] propose to extract more informative active samples using the free energy score in source-free active domain adaptation. In this work, we introduce the utilization of EBM to generate pseudo-labels for self-training in UDA. We calculate an energy score for each pixel using the *logsum-exp* operator [36], [37], and use the pixel-wise energy score to determine whether to assign pseudo-labels. To the best of our knowledge, this work marks the first exploration of EBM for unsupervised domain-adaptive semantic segmentation.

III. Proposed Method

In this section, we introduce two methods, High-order Style Matching (HSM) and Energy Score-based Pseudo-Labeling (ESPL), to directly adapt a network without any intermediate domain dataset and alleviate the overconfidence issue of confidence-based pseudo-labeling. To provide a better understanding of our method, we first describe the background in Sec. III-A, and then introduce the detail of HSM and ESPL in Sec III-B and III-C. An overview of our framework is illustrated in Fig. 3.

A. Background

1) Self-Training for UDA: Unsupervised Domain Adaptation (UDA) aims to train a model that can generalize well on the target domain. We represent labeled source domain dataset $D^s = \{(x^s, y^s)\}$ and unlabeled target domain dataset $D^t = \{x^t\}$, where $x^s, x^t \in \mathbb{R}^{H\times W\times 3}$ represent the images, $y^s \in \mathbb{R}^{H\times W\times C}$ are the ground-truth labels, and C is the number of categories. We denote the semantic segmentation network as g_θ. In UDA, the network g_θ is trained using pixel-wise cross-entropy loss on source domain dataset D^s as follows:

$$L^s_{seg} = -\sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{c=1}^{C} y^s \log g_\theta(x^s). \tag{1}$$

However, deploying a source-only trained model on the target domain often leads to a significant drop in performance due to the inherent differences between the two domains.

To resolve this issue, self-training [6], [7], [16], [17] has emerged as a promising solution and has achieved state-of-the-art performance in UDA. To generate pseudo-labels for unlabeled target domain images, a momentum teacher network g_ϕ which obtains more stable parameters in a temporal ensembling manner is typically involved for the self-training. The teacher network g_ϕ produces pseudo-labels as follows:

$$p^t = [c = \arg\max_c g_\phi(x^t)], \tag{2}$$

where $[\cdot]$ denotes the Iverson bracket.

While self-training has shown promising performance in UDA, using generated pseudo-labels directly may be unreliable, resulting in providing incorrect and noisy supervision signals to the segmentation network g_θ. Therefore, confidence score-based thresholding is widely used to filter out unreliable pseudo-labels:

$$q^t = [\max_c g_\phi(x^t) > \tau_c]. \tag{3}$$

We can obtain filtered pseudo-labels exceeding the confidence threshold τ_c and train g_θ using pixel-wise cross-entropy loss with target domain images and pseudo-labels as follows:

$$L^t_{seg} = -\sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{c=1}^{C} q^t p^t \log g_\theta(x^t). \tag{4}$$

During training, the weights of the momentum teacher network g_ϕ are updated by the Exponential Moving Average (EMA) with the weights of the segmentation network g_θ on each training iteration t [44]:

$$\phi_{t+1} \leftarrow \alpha\phi_t + (1-\alpha)\theta_t. \tag{5}$$

In this work, we have a labeled source clear weather domain dataset $D^{cw} = \{(x^{cw}, y^{cw})\}$ and unlabeled target real foggy domain dataset $D^{rf} = \{x^{rf}\}$, same as the setup described above.

2) Energy Function: The discriminative neural classifier f which maps an input x to real-valued logit values $f(x)$ can be interpreted by the energy-based model [35]. From this perspective, we can derive the free energy score using the *logsumexp* operator as follows:

$$E(x, f(x)) = -T \cdot \log(\sum_{c=1}^{C} e^{f_c(x)/T}), \tag{6}$$

where $f_c(x)$ is the logit value corresponding to the c-th class label, C is the number of classes, and T is the temperature parameter.

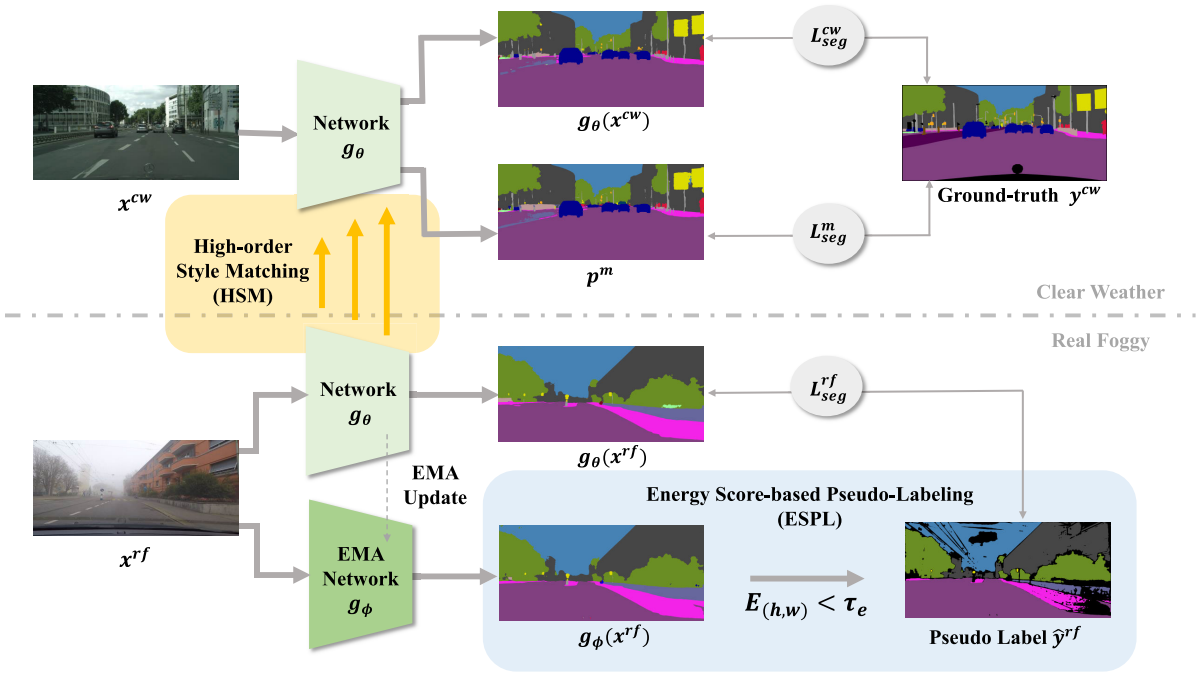

Fig. 3. Overview of DAEN with High-order Style Matching (HSM) and Energy Score-based Pseudo-Labeling (ESPL). Clear weather image x^{cw} and real foggy image x^{rf} are fed into a segmentation network g_θ. At each layer l of an encoder in g_θ, intermediate domain features f_l^m are obtained by matching the high-order feature statistics between clear weather feature f_l^{cw} and real foggy feature f_l^{rf} through HSM. From f_l^{cw} and f_l^m, we can obtain the prediction maps $g_\theta(x^{cw})$ and p^m, thereby a segmentation losses L^{cw}_{seg} and L^m_{seg} are computed with ground-truth label y^{cw}. While real foggy image x^{rf} is also fed into an EMA teacher network g_ϕ, we can obtain more reliable pseudo-label $\hat{y}^{rf}$ with ESPL. Likewise, a segmentation loss L^{rf}_{seg} is computed with prediction map $g_\theta(x^{rf})$ and pseudo-label $\hat{y}^{rf}$.

A lower/higher energy score usually indicates that the input is likely to be an inlier/outlier distribution sample, respectively [37] and [39]. In this work, we adopt an energy score to produce reliable pseudo-labels for unlabeled target samples.

B. High-Order Style Matching (HSM)

It is generally known that the style information of an image can be represented by feature statistics such as mean and standard deviation [29], [30]. Style difference is one of the factors that cause a domain gap, thus matching feature statistics between two domains is essential. Previous methods [30], [45], [46] assume that feature distribution follows a Gaussian distribution, they solely utilize low-order feature statistics (mean and standard deviation) to represent the style of the domain. However, the first- and second-order feature statistics are insufficient to represent complex real fog style. In this work, we further utilize high-order feature statistics (third standardized moment-skewness and fourth standardized moment-kurtosis) to better capture complex real fog style.

Let f_l^{cw} and f_l^{rf} be the clear weather and real foggy features from l-th layer of segmentation network g_θ when clear weather image x^{cw} and real foggy image x^{rf} are fed into the network g_θ. We aim to obtain intermediate domain features by matching high-order feature statistics between f_l^{cw} and f_l^{rf}. Inspired by [33], we first sort two features via Sort-Matching algorithm [47] as:

$$\begin{aligned} \boldsymbol{f_l^{cw}} \in \mathbb{R}^{C_l \times H_l W_l} &: \mu = (\mu_1 \quad \mu_2 \quad \cdots \quad \mu_n), \\ \boldsymbol{f_l^{rf}} \in \mathbb{R}^{C_l \times H_l W_l} &: \nu = (\nu_1 \quad \nu_2 \quad \cdots \quad \nu_n), \end{aligned} \tag{7}$$

where $\{(f_l^{cw})_{\mu_i}\}_{i=1}^{n=H_l W_l}$ and $\{(f_l^{rf})_{\nu_i}\}_{i=1}^{n=H_l W_l}$ are sorted feature values of f_l^{cw} and f_l^{rf} in ascending order. μ_i and ν_i are the indices of the i-th smallest elements of feature vectors, and C_l, H_l, W_l indicate channel dimension, height, and width of features from l-th layer, respectively. Then, we apply the high-order feature statistics matching in a channel-wise manner and obtain the intermediate domain features f_l^m as:

$$(f_l^m)_{\mu_i} = (f_l^{cw})_{\mu_i} + (f_l^{rf})_{\nu_i} - \langle (f_l^{cw})_{\mu_i} \rangle, \tag{8}$$

where $\langle \cdot \rangle$ denotes the stop gradient operation and $(f_l^{cw})_{\mu_i} - \langle (f_l^{cw})_{\mu_i} \rangle$ is introduced to aid back-propagation.

By Eq. 8, HSM ensures a congruence in feature statistics between distinct f_l^{rf} and f_l^m. Hence, intermediate domain features f_l^m elaborate the complex real fog style of f_l^{rf} while maintaining the original content of f_l^{cw}. When the f_l^m is fed into subsequent layers and a classification head, we can obtain a final prediction map p^m. Since the prediction map p^m shares the same semantic scene as the clear weather image x^{cw}, we can apply the pixel-wise cross-entropy loss to prediction map p^m with ground-truth label y^{cw} as follows:

$$L^m_{seg} = -\sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{c=1}^{C} y^{cw} \log(p^m). \tag{9}$$

C. Energy Score-Based Pseudo-Labeling (ESPL)

In the self-training for UDA, pseudo-labels compensate for the absence of ground-truth labels of target samples and enable the network to effectively learn target domain knowledge. Confidence score-based thresholding [48] is widely used to

produce high-quality pseudo-labels. However, it tends to be biased towards high-confident pixels and fails to consider classes with relatively low confidence values.

To address this drawback, we adopt an energy score as an alternative. It is known that the energy score can distinguish whether a sample is likely to be an inlier or an outlier [37], [39]. Based on this perspective, the source and target domains in UDA can be viewed as inlier and outlier distributions, respectively. When we compute the pixel-wise energy scores for unlabeled target images, the pixels with relatively low-energy score can be considered closer to the source domain distribution (inlier) and easier to adapt. Moreover, the softmax confidence score can be decoupled as the sum of the energy score and the maximum logit value. As the softmax confidence score is shifted by the maximum logit value, it generally generates overconfident pseudo-labels. On the other hand, the energy score can prevent such bias since it remains unaffected by this shift. Therefore, we propose Energy Score-based Pseudo-Labeling (ESPL) using the peculiarity of the energy score to produce more reliable pseudo-labels for unlabeled target images and mitigate the overconfidence issue of the confidence score.

When the unlabeled target foggy image x^{rf} is fed into the teacher network g_ϕ, it outputs the pixel-wise logit map $g_\phi(x^{rf}) \in \mathbb{R}^{H\times W\times C}$. We calculate the pixel-wise energy score with the *logsumexp* operator for the unlabeled target samples as follows:

$$E_{(h,w)} = -\log\Big(\sum_{c=1}^{C} e^{(g_\phi(x^{rf})_c)_{(h,w)}}\Big), \tag{10}$$

where $E_{(h,w)}$ is the pixel-wise free energy score at pixel position (h, w) and T is set to 1.

We only assign pseudo-labels to pixels with an energy score lower than the pre-defined energy threshold τ_e. The model is trained with obtained pseudo-labels using the pixel-wise cross-entropy loss based on Eq. 4 as follows:

$$L^{rf}_{seg} = -\sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{c=1}^{C} \mathbb{1}[E_{(h,w)} < \tau_e] p^{rf} \log g_\theta(x^{rf}), \tag{11a}$$

$$= -\sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{c=1}^{C} \hat{y}^{rf} \log g_\theta(x^{rf}), \tag{11b}$$

where $\mathbb{1}$ is an indicator function and $\hat{y}^{rf}$ is a pseudo-label filtered by the pixel-wise energy score $E_{(h,w)}$.

As the g_ϕ updates, the target foggy domain aligns with the source clear weather domain, resulting in lower energy scores of unlabeled target foggy images and more involvement in self-training. We empirically validate that ESPL can generate more reliable pseudo-labels compared to confidence-based thresholding in Sec. IV-E.

D. Optimization

The segmentation network g_θ is also trained using pixel-wise cross-entropy loss on source clear weather domain dataset D^{cw} as follows:

$$L^{cw}_{seg} = -\sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{c=1}^{C} y^{cw} \log g_\theta(x^{cw}). \tag{12}$$

In summary, the total loss L_{total} is expressed as:

$$L_{total} = L^{cw}_{seg} + L^{m}_{seg} + \lambda_{rf} L^{rf}_{seg}, \tag{13}$$

where λ_{rf} is the hyper-parameter to balance the influence of self-training. The whole training framework and the pseudo-code of the method are presented in Algorithm 1.

Algorithm 1 DAEN Algorithm

Input: Labeled clear weather domain dataset D^{cw}, unlabeled real foggy domain dataset D^{rf}, segmentation network g_θ, EMA teacher network g_ϕ, maximum iteration T.
Output: Optimized segmentation network g_θ.

1: Initialize the parameter of g_ϕ with g_θ
2: **for** $t = 0$ to $T - 1$ **do**
3: Update the parameter of g_ϕ with g_θ using EMA update: $\phi_{t+1} \leftarrow \alpha\phi_t + (1-\alpha)\theta_t$.
4: Sample the data $\{x^{cw}, y^{cw}\} \in D^{cw}, x^{rf} \in D^{rf}$.
5: Feed x^{cw} and x^{rf} into segmentation network g_θ.
6: High-order Style Matching (HSM) is applied between f^{cw}_l and f^{rf}_l at the layer l of g_θ and obtain intermediate domain feature f^m_l.
7: Obtain final prediction maps $g_\theta(x^{cw})$ and p^m from g_θ.
8: Feed x^{rf} into EMA teacher network g_ϕ and produce filtered pseudo-label $\hat{y}^{rf}$ using Energy Score-based Pseudo Labeling (ESPL) for self-training.
9: Feed x^{rf} into segmentation network g_θ and obtain the prediction map $g_\theta(x^{rf})$.
10: Optimize the parameter of network g_θ via Eq. 13.
11: **end for**

IV. Experiments

A. Datasets

Cityscapes [26] is a high-resolution (2048 × 1024) dataset containing 3,450 urban scene images collected from 50 different cities. It is split into 2,975 and 500 images for training and validation, respectively. We adopt Cityscapes as the source clear weather domain dataset.

Foggy Zurich [8] is a high-resolution (1920 × 1080) dataset consisting of 3,808 real-world foggy urban scene images collected from Zurich and its suburbs. It is split into 1,552 light foggy images and 1,498 medium foggy images according to fog density. Furthermore, it provides the Foggy Zurich-test which contains 40 real-world foggy scenes for evaluation. It shares the 19 evaluation classes of the Cityscapes. We adopt Foggy Zurich as the target real foggy domain dataset.

Foggy Driving [49] contains 101 real-world foggy urban scene images of which 51 images are captured by a cell phone in Zurich and 50 images are collected from the website, respectively. We only use Foggy Driving for evaluation.

ACDC [50] contains 2,406 real-world urban scene images collected under four adverse conditions (rain, snow, fog, and

night). It is comprised of 1,600 training, 406 validation, and 2,000 test. We adopt a validation set containing three adverse conditions (rain, snow, and fog) to verify the generalization ability of our method. We only utilize this set for evaluation.

B. Implementation Details

Following FIFO [1], we use RefineNet-lw [51] with ResNet-101 [52], initialized with a Cityscapes pre-trained model. Our network is trained by the Stochastic Gradient Descent (SGD) optimizer with a momentum of 0.9 and a weight decay of 5×10^{-4}. The initial learning rate for the encoder and the decoder is set to 6×10^{-4} and 6×10^{-3}, respectively, and the learning rate is decreased gradually using the polynomial policy with a power of 0.9. We apply image resizing, random cropping of 600×600, and random horizontal flipping for data augmentation. For the hyper-parameters, we set the energy score threshold τ_e to -15 and λ_{rf} to 1. All experiments are run on a single RTX 3090 GPU. To ensure reliable results, we evaluate our network with three random seeds and report the average values of the measurements. The mean Intersection over Union (mIoU) is used as the evaluation metric.

C. Baseline Methods

We comprehensively compare our method with various approaches in the field. As a *Baseline*, we utilize RefineNet-lw [51] which has been fully trained on the Cityscapes [26].

1) Conventional UDA Methods: We utilize AdSegNet [13], AdvEnt [53], and FDA [54]. AdSegNet and AdvEnt reduce the domain gap via adversarial learning and entropy minimization, respectively. FDA swaps low-frequency components between source and target images, followed by self-training.

2) Image-to-Image (I2I) Translation Methods: We evaluate CycleGAN [55], CUT [56], and StyTr2 [57]. CycleGAN and CUT preserve content while translating images between domains, using cycle consistency loss and contrastive learning. StyTr2 leverages a Transformer architecture for flexible style transformations. We first train the models to translate Cityscapes [26] images to the style of Foggy Zurich [8]. The resulting synthetic foggy Cityscapes images are then used as an intermediate domain dataset. For domain adaptation, self-training is followed for the unlabeled target domain dataset.

3) Transformer-Based UDA Methods: We select DAFormer [58], HRDA [59], and MIC [60], which excel in synthetic-to-real domain adaptation. DAFormer bridges domain gaps via improved training strategies, HRDA enhances adaptation with high-resolution features, and MIC adds masked image consistency for improved scene understanding. For the comparison, we adopt the results from these transformer-based UDA models trained on Cityscapes $\rightarrow$ Foggy Zurich.

4) Specialized UDA Methods for Foggy Scene Segmentation: We adopt CMAda3+ [27], FIFO [1], CUDA-Net+ [2], and TDo-Dif [3], which represent the state-of-the-art UDA approaches for foggy scene segmentation, as baseline methods. These methods address the domain gap between clear weather and foggy conditions through different techniques, making them suitable for a comprehensive comparison with DAEN. CMAda3+ uses curriculum learning to gradually adapt from light to dense fog, FIFO learns fog-invariant features with a fog pass filter, CUDA-Net+ defines a dual domain (style and fog) gap and addresses the domain gaps utilizing cumulative relationship, and TDo-Dif leverages spatial and temporal similarities with confidence-based pseudo labels. Additionally, we outline the similarities and differences with DAEN.

a) Network perspective: CMAda3+ and TDo-Dif use RefineNet with ResNet-101 as the backbone, while FIFO employs a lighter version, RefineNet-lw. CUDA-Net+ is based on DeepLab-v2. DAEN also utilizes RefineNet-lw with ResNet-101, similar to FIFO. It offers improved efficiency over CMAda3+ and TDo-Dif.

b) Dataset perspective: Unlike CMAda3+, FIFO, and CUDA-Net+, which require intermediate domain datasets, DAEN directly reduces the domain gap with High-order Style Matching (HSM), enabling the model to learn complex fog distributions without the additional datasets. TDo-Dif also reduces the domain gap directly but relies on a model trained on synthetic foggy domain datasets, whereas DAEN starts with a model trained on the clear weather domain dataset, making it more challenging.

c) Training perspective: Unlike CMAda3+, FIFO, CUDA-Net+, and TDo-Dif, DAEN is end-to-end trainable without requiring multi-stage training, which makes it much simpler and more efficient in terms of training. Moreover, while adversarial learning methods like FIFO and CUDA-Net+ struggle with optimization challenges, CMAda3+, TDo-Dif, and DAEN all use a self-training scheme for domain alignment. However, CMAda3+ and TDo-Dif rely on a confidence score for pseudo-label generation, whereas DAEN produces pseudo-labels using an energy score, marking a key difference.

D. Comparison With Other UDA Methods

1) Quantitative Comparison: Tables I, II, and III present the quantitative results on the Foggy Zurich-test, Foggy Driving, and ACDC-Fog validation set. The proposed DAEN consistently outperforms all baselines across these datasets. As highlighted in Table I for Foggy Zurich-test, DAEN demonstrates notable improvements over both *conventional UDA* and *I2I translation methods*. Unlike *I2I translation methods* which require pre-training and additional networks for image translation, DAEN enables the network to learn the real fog style without such cumbersome prerequisites implicitly. When compared to *specialized UDA methods for foggy scene segmentation*, DAEN achieves mIoU improvements of 5.8%, 5.1%, and 5.1% over state-of-the-art methods like FIFO, CUDA-Net+, and TDo-Dif, respectively. For a fair comparison, we also compare DAEN with *FIFO and *CuDA-Net, both trained without intermediate domain datasets. The results show that DAEN achieves better performance, demonstrating DAEN's ability to effectively reduce the domain gap without requiring intermediate domain datasets. Given the strong performance of transformer-based methods in UDA tasks, we compare DAEN with recent *transformer-based UDA methods*. DAEN outperforms these methods on the Foggy Zurich-test by a significant margin. Notably, it surpasses

TABLE I

PERFORMANCE COMPARISON OF MIOU(%) BETWEEN DAEN AND EXISTING STATE-OF-THE-ART METHODS ON FOGGY ZURICH-TEST. OUR RESULTS ARE AVERAGED OVER THREE RANDOM SEEDS. SINCE CLASS *Train* IS NOT INCLUDED IN FOGGY ZURICH-TEST, WE EXCLUDE THE CLASS *Train*. THE BEST AND SECOND-BEST RESULTS ARE HIGHLIGHTED IN **BOLD** AND UNDERLINE. † DENOTE REPRODUCED RESULTS. M.S AND I.D.D DENOTE MULI-STAGE TRAINING AND INTERMEDIATE DOMAIN DATASET, RESPECTIVELY

Method	M.S	I.D.D	Road	S.Walk	Build.	Wall	Fence	Pole	Tr.Light	Tr.Sign	Veget.	Terrain	Sky	Person	Rider	Car	Truck	Bus	M.bike	Bike	mIoU
Baseline	-	-	53.2	51.6	35.1	28.6	19.4	36.8	54.0	56.1	24.5	34.7	57.7	0.0	3.6	79.7	0.0	0.0	0.0	5.7	28.5
*AdSegNet	✗	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	25.0
*ADVEnt	✗	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	39.7
*FDA	✗	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	22.2
†CycleGAN	✓	✓	91.0	57.7	59.3	44.2	27.3	38.8	59.4	61.1	72.3	53.3	89.7	5.2	39.1	85.6	0.0	**57.2**	48.4	16.5	50.3
†CUT	✓	✓	90.3	60.0	61.8	43.5	25.6	39.4	59.4	60.5	72.9	56.1	90.7	6.0	36.0	85.9	0.0	49.7	50.5	16.3	50.3
†StyTr2	✓	✓	86.2	61.3	54.9	38.0	18.8	32.6	57.3	56.9	66.5	51.7	90.1	3.4	41.8	83.6	0.0	40.7	27.8	22.9	45.6
*CMAda3+	✓	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	46.8
†FIFO	✓	✓	65.5	57.7	54.3	49.9	25.8	44.1	60.2	61.6	73.1	63.6	70.4	**6.3**	45.5	84.9	0.0	46.7	46.7	15.1	48.4
*CuDA-Net+	✓	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	49.1
*FIFO	✓	✗	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	37.7
*CuDA-Net	✓	✗	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	48.2
†TDo-Dif	✓	✗	89.8	54.4	**69.0**	**56.7**	**45.3**	36.5	51.7	58.9	**74.5**	59.7	90.1	4.2	44.0	84.8	0.0	40.5	**59.4**	14.0	49.1
*DAFormer	✗	✗	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	40.8
*MIC (DAFormer)	✗	✗	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	43.5
*HRDA	✗	✗	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	46.0
*MIC (HRDA)	✗	✗	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	49.7
DAEN (Ours)	✗	✗	**92.3**	**63.9**	67.8	56.4	28.7	**44.4**	**61.8**	**62.9**	74.1	**68.0**	**91.4**	3.8	**48.2**	**87.9**	0.0	45.8	48.4	**29.6**	**54.2**

TABLE II

PERFORMANCE COMPARISON OF MIOU(%) BETWEEN DAEN AND EXISTING STATE-OF-THE-ART METHODS ON FOGGY DRIVING. OUR RESULTS ARE AVERAGED OVER THREE RANDOM SEEDS. THE BEST AND SECOND-BEST RESULTS ARE HIGHLIGHTED IN **BOLD** AND UNDERLINE. † DENOTE REPRODUCED RESULTS. M.S AND I.D.D DENOTES MULI-STAGE TRAINING AND INTERMEDIATE DOMAIN DATASET, RESPECTIVELY

Method	M.S	I.D.D	Road	S.Walk	Build.	Wall	Fence	Pole	Tr.Light	Tr.Sign	Veget.	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	M.bike	Bike	mIoU
Baseline	-	-	88.0	26.6	68.1	28.5	14.6	42.5	44.3	54.5	63.0	9.1	86.9	64.5	46.7	65.0	6.8	13.0	27.5	28.7	49.2	43.6
*AdSegNet	✗	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	29.7
*ADVEnt	✗	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	46.8
*FDA	✗	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	21.8
†CycleGAN	✓	✓	92.2	35.3	68.1	27.4	14.1	**46.7**	49.8	58.7	71.2	11.9	88.1	66.0	41.0	66.8	7.0	19.1	60.9	30.8	53.0	47.8
†CUT	✓	✓	92.1	34.2	68.2	20.8	13.8	46.5	49.2	57.0	70.2	**12.5**	87.6	66.2	41.3	66.9	9.8	16.9	61.3	30.5	52.1	47.2
†StyTr2	✓	✓	91.8	33.7	66.6	**36.2**	10.5	33.0	46.4	53.0	71.7	7.3	87.6	66.3	46.1	65.4	18.1	11.6	33.8	**35.0**	50.6	45.5
*CMAda3+	✓	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	49.8
†FIFO	✓	✓	90.8	39.1	**72.9**	24.2	20.0	42.3	51.0	59.1	72.0	9.4	**90.2**	64.7	48.5	71.0	25.4	65.8	43.4	24.8	49.1	50.7
*CuDA-Net+	✓	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	53.5
†TDo-Dif	✓	✗	91.7	33.0	72.6	22.7	**21.1**	43.1	46.1	57.2	70.9	11.4	87.3	66.1	41.1	**72.3**	**37.8**	15.5	30.0	25.0	44.2	46.8
DAEN (Ours)	✗	✗	**92.5**	**43.2**	70.1	32.0	19.2	44.8	**51.4**	**60.6**	**73.2**	12.4	89.5	**68.3**	**51.1**	71.3	16.4	**78.0**	**64.9**	30.8	**55.6**	**54.0**

TABLE III

PERFORMANCE COMPARISON OF MIOU(%) BETWEEN DAEN AND EXISTING STATE-OF-THE-ART METHODS ON ACDC-FOG VALIDATION SET. OUR RESULTS ARE AVERAGED OVER THREE RANDOM SEEDS. THE BEST AND SECOND-BEST RESULTS ARE HIGHLIGHTED IN **BOLD** AND UNDERLINE. † DENOTE REPRODUCED RESULTS. M.S AND I.D.D DENOTES MULI-STAGE TRAINING AND INTERMEDIATE DOMAIN DATASET, RESPECTIVELY

Method	M.S	I.D.D	Road	S.Walk	Build.	Wall	Fence	Pole	Tr.Light	Tr.Sign	Veget.	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	M.bike	Bike	mIoU
Baseline	-	-	79.8	65.2	67.8	18.7	26.5	30.5	59.8	54.7	71.2	49.2	93.6	41.6	49.8	69.9	20.9	25.2	79.3	12.1	56.4	51.2
†CycleGAN	✓	✓	94.9	75.5	74.0	36.7	32.7	37.3	65.6	62.3	82.6	**56.1**	94.3	46.9	57.0	83.6	43.2	60.0	78.7	40.1	60.0	62.2
†CUT	✓	✓	94.4	74.1	76.2	41.1	32.2	36.8	65.4	61.5	82.6	54.4	96.0	43.0	58.0	83.3	46.6	71.0	82.4	59.0	57.2	64.0
†StyTr2	✓	✓	87.0	73.8	72.5	42.9	28.9	31.9	61.1	54.1	77.9	52.5	85.5	**49.2**	51.0	80.0	30.1	45.5	43.4	33.6	59.6	55.8
†FIFO	✓	✓	51.3	64.9	71.2	22.0	26.8	27.7	49.2	52.3	73.2	46.1	60.3	48.5	52.5	74.8	16.9	54.1	67.4	16.2	60.6	49.3
†TDo-Dif	✓	✗	94.3	73.8	82.3	**50.7**	**44.3**	**44.8**	60.1	59.0	**85.9**	46.3	95.1	42.1	**65.9**	84.6	**48.3**	70.6	87.6	**66.3**	41.2	65.4
DAEN (Ours)	✗	✗	**95.5**	**81.4**	**82.7**	48.2	35.5	44.4	**68.8**	**66.5**	85.7	55.3	**97.3**	46.7	58.3	**85.6**	41.5	**74.2**	**88.3**	27.8	**61.7**	**65.6**

MIC (HRDA), which has shown state-of-the-art results on traditional UDA benchmarks, reaffirming DAEN's effectiveness in foggy scene segmentation. In terms of class-wise performance, DAEN not only shows advantages in large classes such as *road, sidewalk, terrain*, and *sky*, but also demonstrates notable improvements in fine-grained segmentation classes like *bicycle* and *rider*. As shown in Tables II and III, DAEN also outperforms FIFO, CUDA-Net+, and TDo-Dif by 1.6%, 0.5%, and 6.7% mIoU on Foggy Driving, and by 16.3%, -, and 0.2% on ACDC-Fog validation set. Additionally, DAEN achieves either the best or second-best class-wise performance across most categories.

2) Qualitative Comparison: We conduct a qualitative comparison of DAEN with Baseline, FIFO, and TDo-Dif on three distinct datasets: Foggy Zurich-test, Foggy Driving, and ACDC-Fog validation set. As illustrated in Fig. 4, 5, and 6, the results reveal that the Baseline, FIFO, and TDo-Dif exhibit inferior results and struggle to discern scene context when compared to DAEN. For instance, FIFO tends to misclassify regions such as the *sky* or *sidewalk* as belonging to the *road* class. TDo-Dif faces challenges in distinguishing between a *wall* and a *fence*, and it has trouble precisely delineating the boundaries of a *sidewalk*. In contrast, our proposed DAEN correctly classifies the *sky* region and assigns appropriate classes to ambiguous areas. These qualitative results align well with the class-wise performance, underscoring the effectiveness of Energy Score-based Pseudo Labeling (ESPL) in generating reliable pseudo-labels that enhance the network's ability to understand the contextual information of the target domain.

E. Ablation Studies

1) Effectiveness of HSM and ESPL: We first conduct the experiment on the Foggy-Zurich test (FZ) to validate the effectiveness of the two proposed components: HSM and ESPL. In Table IV, the Baseline shows poor performance on

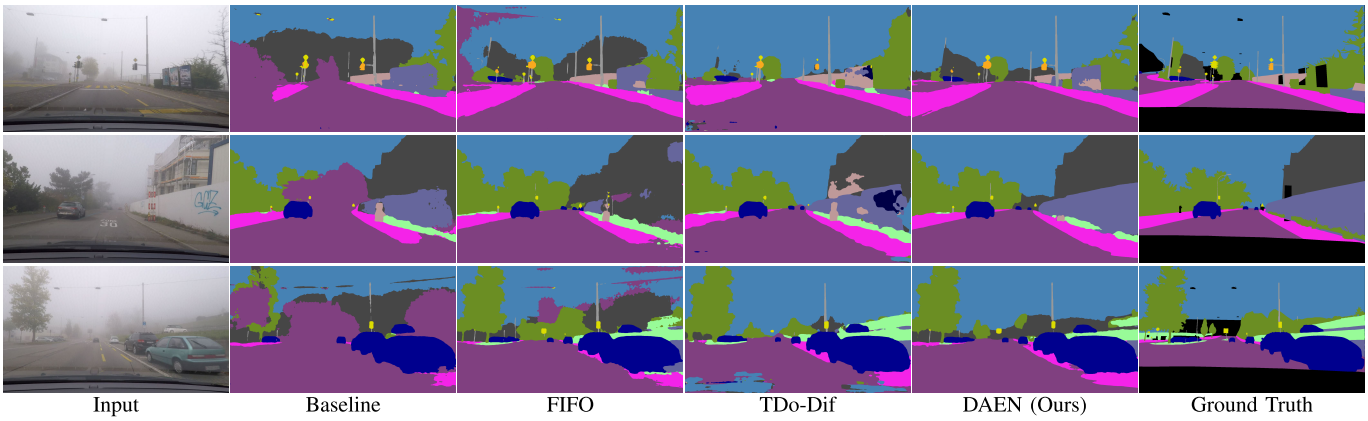

Fig. 4. Qualitative comparison between the proposed DAEN and existing state-of-the-art methods on Foggy Zurich-test.

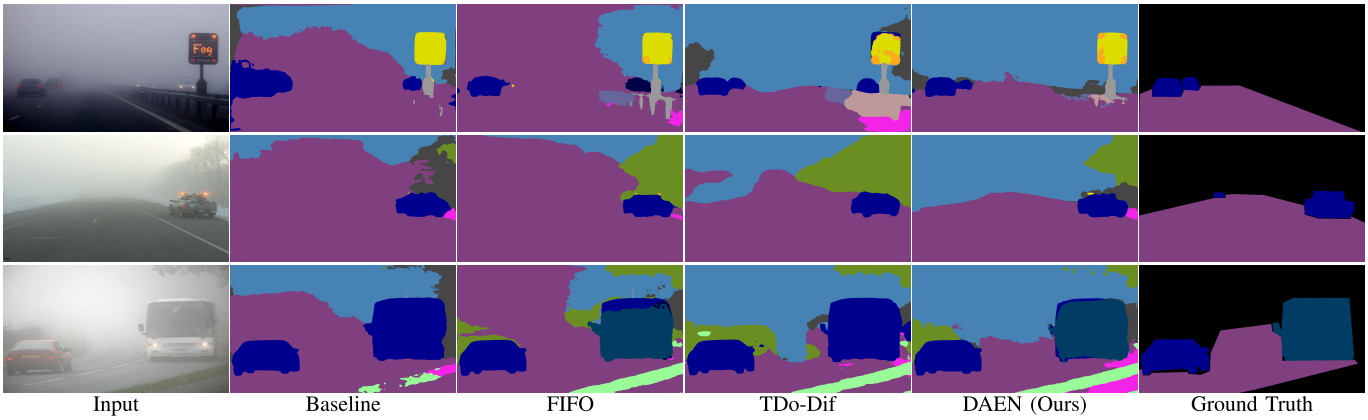

Fig. 5. Qualitative comparison between the proposed DAEN and existing state-of-the-art methods on Foggy Driving.

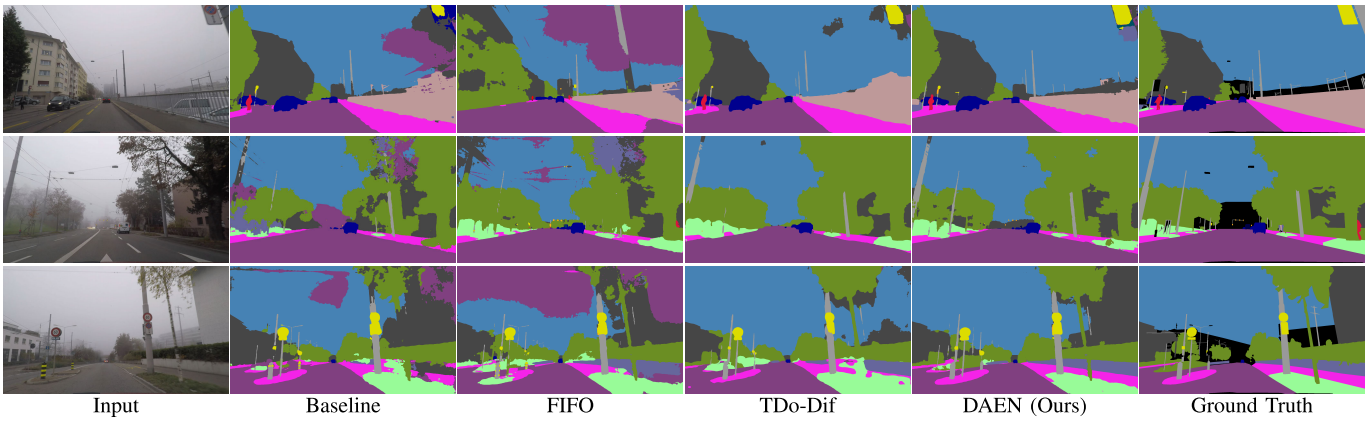

Fig. 6. Qualitative comparison between the proposed DAEN and existing state-of-the-art methods on ACDC-Fog validation set.

the real foggy domain due to overfitting on the clear weather domain. When we combine the Baseline with the ESPL, ESPL significantly improves the performance by 17.1%. This indicates that self-training through the ESPL enables the network to learn the real foggy domain knowledge using reliable pseudo-labels. Additionally, we observe that the HSM can boost the considerable performance by 19.9%, which means that the intermediate domain feature from HSM helps to reduce the large gap between clear weather and real foggy domains. Finally, combining both components contributes to a mutually complementary improvement in performance.

TABLE IV
ABLATION STUDIES OF OUR FRAMEWORK'S TWO COMPONENTS: HSM AND ESPL

Methods	L_{seg}^{cw}	L_{seg}^{m}	L_{seg}^{rf}	FZ	Δ
Baseline	✓			28.5	+0.0
+ESPL	✓		✓	45.6	+17.1
+HSM	✓	✓		48.4	+19.9
DAEN	✓	✓	✓	**54.2**	+25.7

2) The Impact of Layers to Which HSM Is Applied: Table V shows the impact according to layers HSM is applied in the network. Surprisingly, applying HSM only after the

TABLE V

INFLUENCE OF THE LOCATION OF HSM. L0 DENOTES INSERTING HSM AFTER THE FIRST CONVOLUTION LAYER AND L1-3 DENOTES INSERTING HSM AFTER THE CORRESPONDING n-TH RESNET LAYER. FZ DENOTES THE FOGGY ZURICH-TEST

L0	L1	L2	L3	FZ	Δ
				28.5	+0.0
✓				51.7	+23.2
✓	✓			51.1	+22.6
✓	✓	✓		52.4	+24.7
✓	✓	✓	✓	**54.2**	+25.7

TABLE VI

(LEFT): INFLUENCE OF EXPONENTIAL MOVING AVERAGE (EMA) UPDATE RATIO α AND LOSS WEIGHT λ_{rf} FOR SELF-TRAINING. **(RIGHT):** INFLUENCE OF LOSS WEIGHT λ_{rf} FOR SELF-TRAINING. FZ DENOTES THE FOGGY ZURICH-TEST

α	FZ
0.9	44.9
0.99	49.5
0.999	**54.2**
0.9999	51.5

λ_{rf}	FZ
0.1	51.8
1	**54.2**
5	49.8

first convolution layer (L0) yields a significant performance improvement. Besides, as we gradually insert HSM after each residual block (L1-3), a gradual improvement in segmentation results is observed. The best performance is achieved when HSM is applied to L0-3. This is because applying HSM to the shallow layer allows the network to better consider the real fog style without impairing the semantic information. Therefore, we select this as our final model.

3) The Impact of Exponential Moving Average (EMA) Update Ratio α and λ_{rf} for Self-Training: Table VI (left) shows the impact of Exponential Moving Average (EMA) ratio α for updating the momentum teacher network g_ϕ. We observe that $\alpha = 0.999$ shows the best performance on Foggy Zurich-test (FZ) [8]. Thus, we set α to 0.999 for the EMA update. Furthermore, we compare the performance varying the loss weight λ_{rf} to verify the influence of λ_{rf} for self-training. As shown in Table VI (right), $\lambda_{rf} = 1$ shows the best performance on Foggy Zurich-test (FZ) [8] compared to $\lambda_{rf} = 0.1$ and $\lambda_{rf} = 5$. So, we set λ_{rf} to 1.

4) The Impact of Energy Score Threshold Value τ_e: We investigate the selection of an optimal energy score threshold for Energy Score-based Pseudo Labeling (ESPL) and analyze its sensitivity. To achieve this, we first generate pixel-wise energy score histograms for all images in the Foggy Zurich-test (FZ) using the Baseline model (Fig. 9b). Based on the distribution of energy scores, we select candidate thresholds τ_e at specific percentiles of the pixel-wise energy scores: -19 (top 0.02% of pixels with the lowest energy scores), -17 (top 1%), -15 (top 5%), -13 (top 10%), -11 (top 20%), and -9 (top 40%). We then evaluate the sensitivity of these thresholds. As shown in Table VII, $\tau_e = -15$ strikes the best balance between effective adaptation to the target domain (FZ) and maintaining performance in the source domain (CW). Lower thresholds, such as $\tau_e = -19$ and $\tau_e = -17$, are overly restrictive, limiting the number of pixels available for self-training, leading to learning target domain knowledge insufficiently. Conversely, higher thresholds like $\tau_e = -13$, $\tau_e = -11$, and $\tau_e = -9$ allow more noisy pseudo-labels, which degrades performance. A threshold of -15 minimizes the impact of noisy pseudo-labels by involving only reliable pixels early in the process. As training progresses, more target domain pixels achieve lower energy scores, allowing more pixels to contribute to learning (Fig. 9d). Thus, we select $\tau_e = -15$ as the optimal threshold, ensuring robust performance across both real foggy domains (FZ, FD, ACDC-Fog) and the source clear-weather domain (CW).

TABLE VII

COMPARISON OF MODEL PERFORMANCE BASED ON VARYING THE VALUE OF THE ENERGY SCORE THRESHOLD τ_e. FZ, FD, ACDC-FOG AND CW DENOTE THE FOGGY ZURICH-TEST, FOGGY DRIVING, ACDC-FOG VALIDATION SET, AND THE CITYSCAPES VALIDATION SET, RESPECTIVELY

τ_e	FZ	FD	ACDC-Fog	CW
-19	52.2	51.4	65.4	**72.7**
-17	53.3	51.4	**65.6**	67.6
-15	**54.2**	**54.0**	**65.6**	66.5
-13	53.2	53.5	65.3	65.6
-11	54.0	51.8	64.4	65.1
-9	52.8	51.7	64.6	65.0

TABLE VIII

COMPARISON OF ADAIN, MIXSTYLE, AND HSM FOR OBTAINING INTERMEDIATE DOMAIN FEATURES. WE MEASURE THE DISTANCE BETWEEN TWO FEATURE STATISTICS USING FEATURES, f_l^{rf} AND f_l^m, FROM THE FIRST RESIDUAL BLOCK OF THE NETWORK. Δ DENOTES THE EUCLIDEAN DISTANCE BETWEEN TWO FEATURE STATISTICS FROM f_l^{rf} AND f_l^m. FZ AND BDD DENOTE THE FOGGY ZURICH-TEST AND BDD VALIDATION SET, RESPECTIVELY

Methods	FZ	BDD	Δ_μ	Δ_σ	$\Delta_{\widetilde{\mu_3}}$	Δ_{Kurt}
Baseline	28.5	45.1	-	-	-	-
AdaIN	47.5	52.8	0	0	0.0023	0.0068
MixStyle	49.7	52.7	0	0	0.0007	0.0079
HSM	**54.2**	**53.2**	0	0	0	0

F. Further Evaluation

1) AdaIN, MixStyle vs HSM: To validate the effectiveness of High-order Style Matching (HSM) in capturing high-order feature statistics, we compare it with AdaIN [30] and MixStyle [46] which only utilize low-order statistics such as mean and standard deviation. For this comparison, we compute the mean (μ), standard deviation (σ), third standardized moment-skewness ($\widetilde{\mu_3}$), and fourth standardized moment-kurtosis ($Kurt$) from the features, f_l^{rf} and f_l^m, and measure their differences. As shown in Table VIII, while AdaIN and MixStyle only match low-order statistics, HSM accurately aligns both low-order and high-order statistics.

To further evaluate the impact of HSM, we replace it with AdaIN and MixStyle in our framework and compared their performance on foggy scene segmentation. The results in Table VIII show that HSM outperforms both AdaIN and MixStyle. While AdaIN and MixStyle assume that real-world feature distributions follow a Gaussian distribution and thus rely solely on low-order statistics (mean and standard deviation) for feature distribution matching, these assumptions

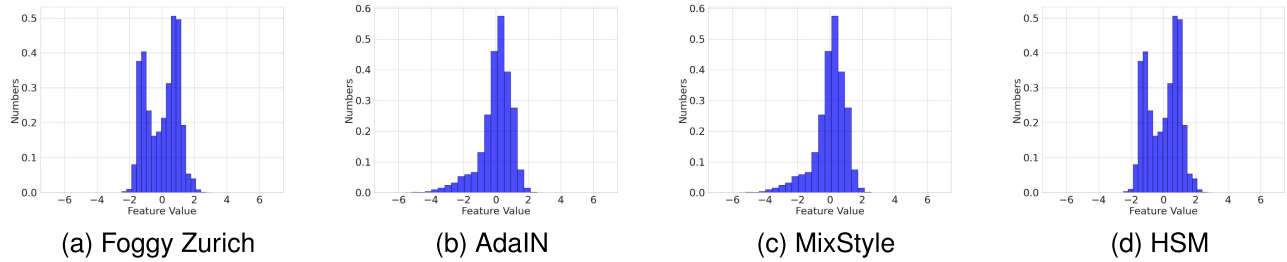

Fig. 7. **(a)**: Histograms of feature values from the target real foggy domain features, f^{rf}. **(b), (c), (d)**: Histograms of feature values from the intermediate domain features, f^{m}, corresponding to each matching method. All features are extracted from the first residual block of RefineNet-lw, which is trained on the Cityscapes dataset. To extract f^{rf} and f^{m}, we use the Foggy Zurich-test (FZ) and Cityscapes validation set.

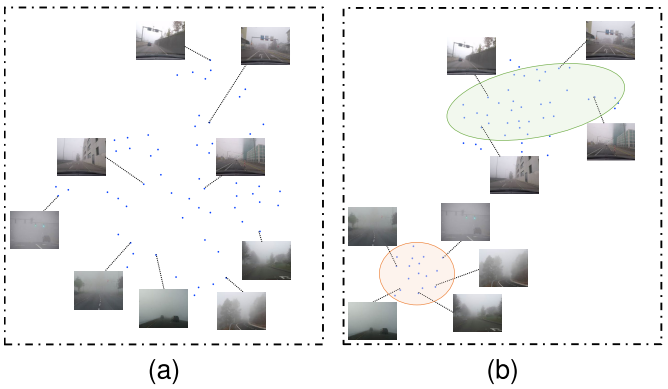

Fig. 8. **(a), (b)**: t-SNE [61] visualization of the low-order feature statistics (mean, μ) and the high-order feature statistics (third standardized momen-t-skewness, $Kurt$), respectively. For the visualization, we use two real foggy domain datasets: Foggy Zurich-test (FZ) and Foggy Driving (FD). The visualized features are extracted from the first residual block of ResNet-101.

are insufficient for capturing the complexities of real-world distributions, such as the ones present in foggy datasets. In contrast, HSM addresses these complexities by integrating both low-order and high-order statistics.

To illustrate this, we visualize the histograms of the feature values from intermediate domain features, f^{m} for each matching method. As depicted in Fig. 7, AdaIN and MixStyle fail to accurately match the feature values from the real foggy domain features, f^{rf}. Conversely, as shown in Fig. 7a and Fig. 7d, HSM achieves a more precise match by incorporating high-order statistics. This demonstrates that HSM more effectively captures complex real fog distributions, enabling the network to represent real fog styles compared to existing methods.

To further verify that high-order statistics effectively capture the complexity of real fog styles, we visualize t-SNE [61] plots for both low-order and high-order statistics. As shown in Fig. 8b, high-order statistics, such as the third standardized moment-skewness ($Kurt$), effectively capture real fog features related to fog density. In contrast, Fig. 8a illustrates that low-order statistics struggle to differentiate between various fog densities within complex fog patterns. These findings highlight the significance of leveraging high-order feature statistics to understand the intricate real fog characteristics.

In addition to evaluating HSM's effectiveness for weather changes, we investigate its applicability to scene changes. For this evaluation, we use the BDD dataset, a real-world dataset featuring diverse urban driving scenes, as the target domain and compare the adaptation performance from Cityscapes to BDD against other style alignment methods, such as AdaIN and MixStyle. As shown in Table VIII, HSM achieves an 8.1 mIoU improvement over the baseline, along with a 0.4 mIoU gain over AdaIN and a 0.5 mIoU gain over MixStyle. These results demonstrate that HSM is not only effective for weather changes but also capable of adapting to scene changes. This comparison highlights the versatility of HSM in handling various types of domain shifts, further validating its robustness beyond weather-specific adaptations.

2) Confidence Score vs Energy Score: We aim to verify whether Energy Score-based Pseudo-Labeling (ESPL) can alleviate the overconfidence issue of confidence-based pseudo-labeling [48]. Fig. 9a and 9b show the distributions of confidence score and energy score over predictions from the Baseline [51] on Foggy Zurich-test (FZ) [8]. As shown in Fig. 9a, most predictions are biased towards high-confidence values. In contrast, the distribution of the energy score is more spread out compared to the confidence score distribution. We also compare these distributions for predictions from DAEN on the FZ. As shown in Fig. 9c, while the confidence score is more highly confident, the energy score results in a more evenly distributed histogram. The bias in the confidence score can be problematic in class-imbalanced driving scene datasets. Even if pixels belong to tail classes (e.g. *sign*, *pole*), the model may predict high confidence for head classes (e.g. *road*, *vegetation*), leading to incorrect pseudo labels. However, the relatively even distribution of energy scores helps mitigate this bias and generate more reliable pseudo-labels.

To further validate that ESPL can produce reliable pseudo-labels for the target foggy domain, we visualize the pseudo-labels generated by the confidence score and ESPL. As shown in the first row of Fig. 10, confidence-based labeling frequently assigns the incorrect pseudo-label *vegetation* to an overpass region. It also fails to provide sufficient information for tail classes such as *train* (first row), *traffic sign* (second row), and *pole* (third row). In contrast, ESPL generates more comprehensive pseudo-labels for these classes, allowing the model to better understand their semantics. These results demonstrate that ESPL effectively addresses the overconfidence and mitigates bias towards head classes (e.g., *vegetation, road*).

Furthermore, we replace ESPL in our framework with confidence-based pseudo-labeling [48] and vary the confidence threshold value. As shown in Table IX, although

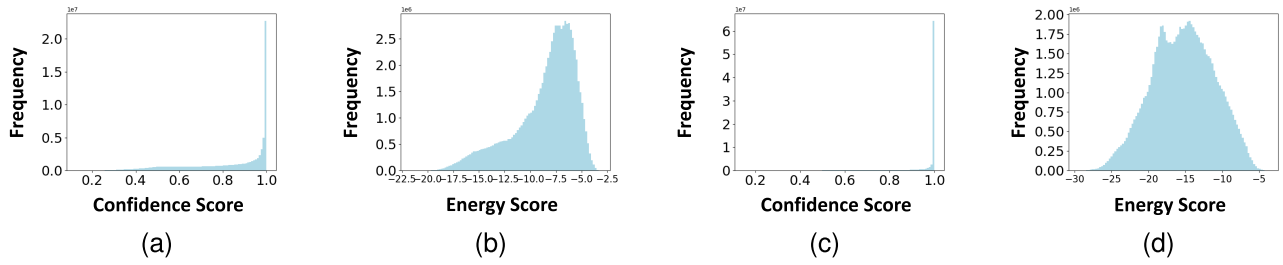

Fig. 9. **(a), (b)**: Confidence & energy score distributions over all predictions from Baseline on Foggy Zurich-test, respectively. **(c), (d)**: Confidence & energy score distributions over all predictions from DAEN on Foggy Zurich-test (FZ), respectively.

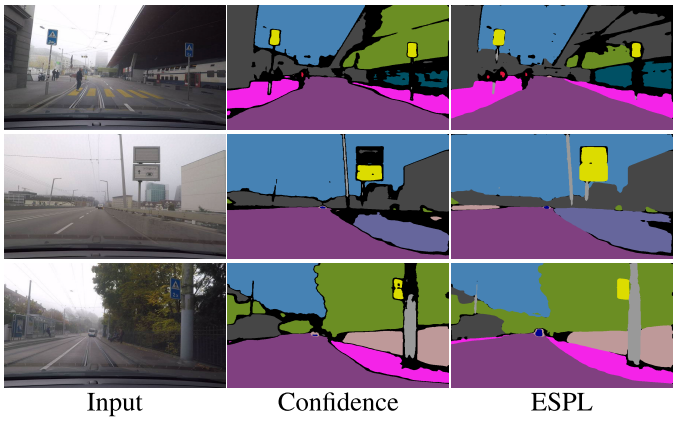

Fig. 10. Qualitative comparison between pseudo labels generated by confidence score and energy score. The threshold value for each method filters the black regions.

TABLE IX

QUANTITATIVE COMPARISON BETWEEN CONFIDENCE SCORE AND ESPL. FZ AND FD DENOTE THE FOGGY ZURICH-TEST AND FOGGY DRIVING, RESPECTIVELY

Method	FZ	FD
Baseline	28.5	43.6
Confidence $= 0.7$	47.9	49.7
Confidence $= 0.8$	48.4	50.0
Confidence $= 0.9$	46.9	50.0
ESPL (Ours)	**54.2**	**54.0**

confidence-based pseudo-labeling improves performance over the baseline, it still falls short of ESPL's performance. ESPL surpasses the confidence-based pseudo-labeling by 1.1% on the Foggy Zurich-test (FZ) and 3.6% on Foggy Driving (FD). This improvement is due to ESPL's ability to provide more reliable pseudo-labels, enabling the model to learn scene context information more effectively in foggy environments.

3) Adaptive Threshold Adjustment for ESPL: We explore adaptive methods for dynamically adjusting the energy score threshold during training. To better understand how pixel-wise energy scores evolve, we track the minimum, mean, and maximum energy scores of target domain batches throughout DAEN's training process. As shown in Fig. 11, energy scores consistently decrease over time, indicating that the model gradually incorporates more target domain pixels into self-training. This trend aligns with the observations in Fig. 9, where pixel-wise energy scores shift to lower values.

Given this evolving distribution of pixel-wise energy scores, we propose and evaluate three strategies for adaptive threshold

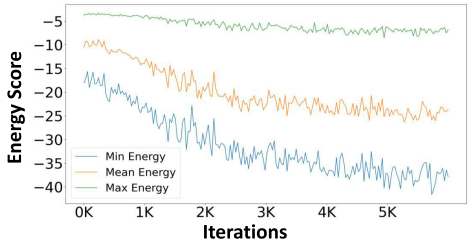

Fig. 11. Changes in the minimum, mean, and maximum pixel-wise energy scores of target domain batches over iterations during the training of DAEN.

TABLE X

QUANTITATIVE PERFORMANCE COMPARISON BASED ON VARIOUS ADAPTIVE THRESHOLD ADJUSTMENT STRATEGIES. FZ AND FD DENOTE THE FOGGY ZURICH-TEST AND FOGGY DRIVING, RESPECTIVELY

Method	FZ	FD
Linear Decay Thresholding	53.8	50.0
Exponential Decay Thresholding	53.6	51.3
Linear Ascend Thresholding	52.9	51.1
Fixed Thresholding (Ours)	**54.2**	**54.0**

adjustment: 1) linear decay, 2) exponential decay, and 3) linear ascend. Each strategy begins with an initial energy threshold of -15, which is adjusted as training progresses. The results of these strategies are presented in Table X. Both the linear and exponential decay methods gradually lower the threshold, allowing more reliable pseudo-labels to be included as the energy scores decrease. However, these methods also reduce the number of pixels involved in training, potentially limiting the model's exposure to diverse data. On the other hand, the linear ascend method raises the threshold over time, which involves a greater number of pixels in training, but at the risk of introducing noisy pseudo-labels, which could negatively impact the model's adaptation to the target domain.

After evaluating these strategies, we opt for a fixed threshold approach for ESPL, as it strikes the best balance between incorporating reliable pseudo-labels and ensuring sufficient pixel diversity during adaptation. As shown in Table X, the fixed threshold approach delivers superior performance across multiple datasets, demonstrating its effectiveness in maintaining robust learning while adapting to the target domain.

4) Theoretical Analysis of Energy Score: To theoretically understand the relationship between softmax confidence score and energy score, we derive a mathematical connection by referring to the analysis of prior work [37]. For data x

TABLE XI
QUANTITATIVE COMPARISON WITH THE BASELINE, FIFO, AND DAEN ON ACDC-SNOW- AND ACDC-RAIN VALIDATION SETS

Methods	ACDC-Snow	ACDC-Rain
Baseline	43.2	42.4
FIFO	41.2	41.2
CuDA-Net	47.2	48.5
TDo-Dif	53.5	50.5
Baseline + HSM	50.1	48.0
Baseline + ESPL	50.8	54.2
DAEN (Ours)	**54.6**	**54.5**

TABLE XII
COMPUTATIONAL COMPLEXITY COMPARISON WITH THE EXISTING METHODS. TRAINING TIME FOR TDO-DIF IS EXCLUDED DUE TO AN ISSUE WITH THE PSEUDO-LABEL GENERATION IN THE PUBLICLY AVAILABLE CODE. TO MEASURE THE GMACS, WE UTILIZE THE IMAGES (1920 × 1080 RESOLUTION) FROM THE FOGGY ZURICH-TEST. F AND M DENOTE THE FOG PASS FILTER AND MAIN NETWORK OF FIFO, RESPECTIVELY

Methods	# of Trainable Params	Training Time	GMACs
FIFO	F: 138.27M / M: 46.34M	F: 20h / M: 13h	410.7
TDo-Dif	118.5M	-	2,085.8
DAEN (Ours)	46.24M	15h	410.7

and corresponding label y, the energy score and softmax confidence score can be expressed as follows:

$$E(x, f(x)) = -T \cdot \log(\sum_{c=1}^{C} e^{f_c(x)/T}), \tag{14}$$

where $f_c(x)$ is the logit value corresponding to the c-th class label, C is the number of classes, and T is the temperature parameter.

$$p(y|x) = \frac{e^{f_y(x)}}{\sum_{c=1}^{C} e^{f_c(x)}}, \tag{15}$$

where $f_y(x)$ denotes the y^{th} index of the logits $f(x)$.

$$\begin{aligned} \max_y p(y|x) &= \max_y \frac{e^{f_y(x)}}{\sum_{c=1}^{C} e^{f_c(x)}} \\ &= \frac{e^{f^{max}(x)}}{\sum_{c=1}^{C} e^{f_c(x)}}. \end{aligned} \tag{16}$$

Based on the Eq. 1 and 3, we can decompose the logarithm of the max confidence score into the sum of the energy score ($T = 1$) and the max logit value:

$$\begin{aligned} \log \max_y p(y|x) &= \log \frac{e^{f^{max}(x)}}{\sum_{c=1}^{C} e^{f_c(x)}} \\ &= f^{max}(x) - \log(\sum_{c=1}^{C} e^{f_c(x)}) \\ &= f^{max}(x) + E(x, f(x)) \end{aligned} \tag{17}$$

Through Eq. 4, the confidence score can be seen as a specific case where the energy score is shifted by the maximum logit value $f^{max}(x)$. On the other hand, the energy score is free from such bias. Therefore, we adopt the energy score as a criterion to improve the quality of pseudo-labels instead of the confidence score.

5) Generalization on Diverse Weather Conditions: To assess the generalization capability of DAEN across various adverse weather conditions, we conducted experiments on the ACDC-Snow and ACDC-Rain validation sets [50]. DAEN, trained on the Cityscapes → Foggy Zurich task, is evaluated against several baseline methods, including the Baseline, FIFO, CuDA-Net, and TDo-Dif. As indicated in Table XI, DAEN consistently outperforms all existing methods, showcasing its robust generalization and applicability across diverse real-world scenarios. Furthermore, even when only HSM or ESPL is integrated with the Baseline, substantial improvements in generalization are observed.

The superior performance of DAEN can be attributed to the effectiveness of its core components, HSM and ESPL. HSM facilitates the alignment of high-order feature statistics, enabling the model to capture intricate real-world distributions beyond foggy weather patterns, thus improving its adaptability to different weather conditions. ESPL, on the other hand, generates more reliable pseudo-labels, allowing the model to learn robust features across a wide range of categories. In contrast to existing baseline methods, which depend on intermediate domain datasets (e.g., synthetic fog) or models pre-trained on synthetic foggy datasets, DAEN directly adapts without reliance on synthetic fog data. This approach enables DAEN to learn comprehensive real-world distributions, enhancing its performance across various adverse weather conditions.

Qualitative analysis further substantiates DAEN's effectiveness under other adverse weather conditions. As shown in Fig. 12, DAEN accurately identifies the *sky* region compared to other methods. Moreover, while TDo-Dif, the second-best performer, often struggles to differentiate distant *poles* and tends to misclassify snow-covered *sidewalks* as *roads*, DAEN performs exceptionally well in these areas. This is evident in Fig. 12 and Fig. 13, where DAEN consistently identifies fine *poles* and *sidewalks* under both snowy and rainy conditions.

6) Computational Complexity: To assess the computational complexity of DAEN, we compare it against FIFO [1] and TDo-Dif [3], both of which have publicly available code. As presented in Table XII, FIFO involves a two-stage training process consisting of a fog pass filter and a main network. The fog pass filter has 138.27M trainable parameters, while the main network includes 46.34M parameters, with training times of 20 and 13 hours, respectively. Although TDo-Dif does not require additional parameters for its multi-stage pseudo-label generation, its main network utilizes 118.5M trainable parameters. In contrast, DAEN reduces the number of trainable parameters to 46.24M and requires only 15 hours of training time. Furthermore, DAEN maintains a similar GMAC count (410.7 GMACs) to FIFO while requiring significantly fewer GMACs than TDo-Dif (2,085.8 GMACs). These results highlight that DAEN not only delivers effective performance but also offers superior computational efficiency, providing notable advantages in both training time and resource utilization.

7) Failure Cases: Although DAEN achieves superior performance in foggy scene segmentation, it still encounters specific failure cases. Fig. 14 illustrates examples of these failure cases on Foggy Zurich-test [8]. DAEN often misclassifies

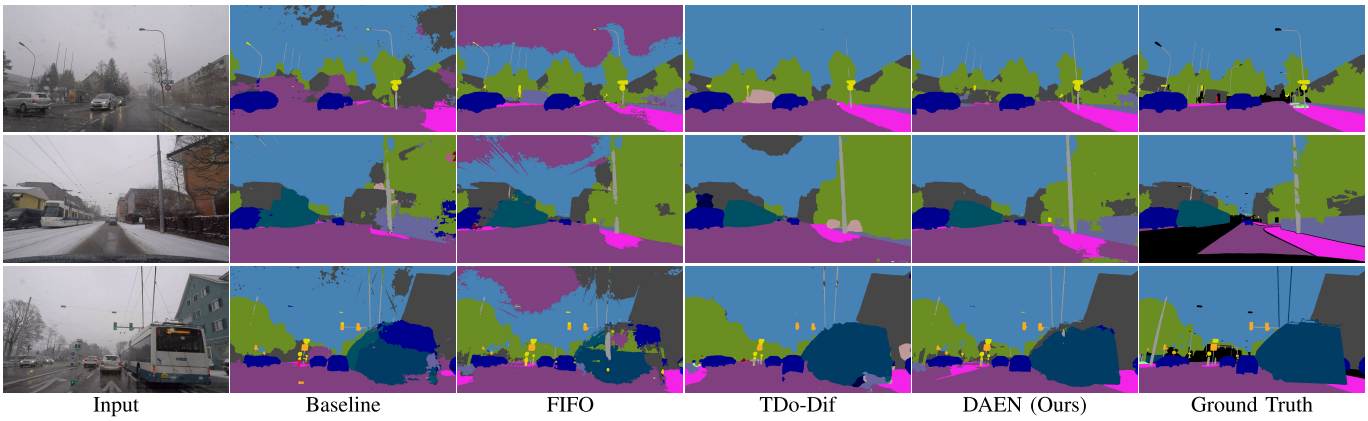

Fig. 12. Qualitative comparison between the proposed DAEN and existing state-of-the-art methods on ACDC-Snow validation set.

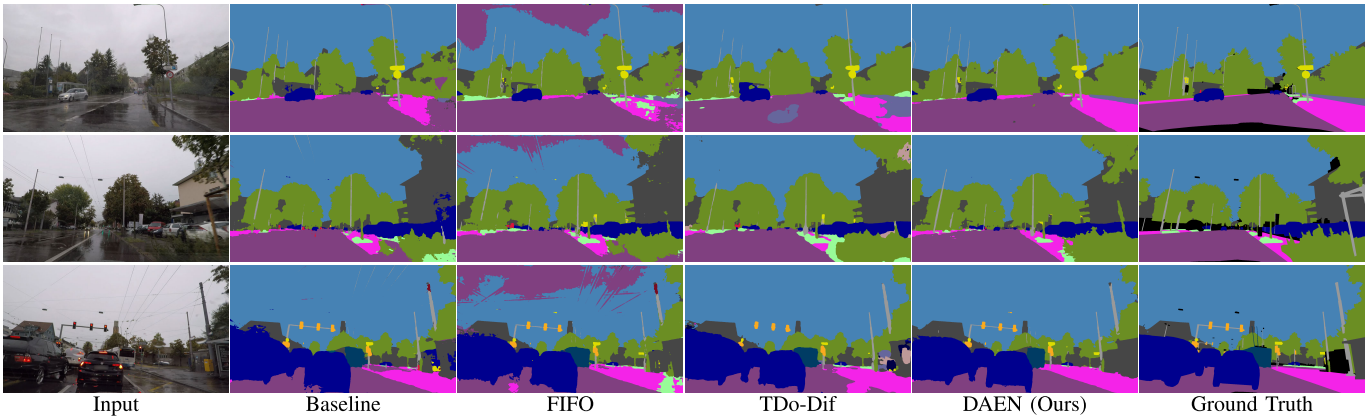

Fig. 13. Qualitative comparison between the proposed DAEN and existing state-of-the-art methods on ACDC-Rain validation set.

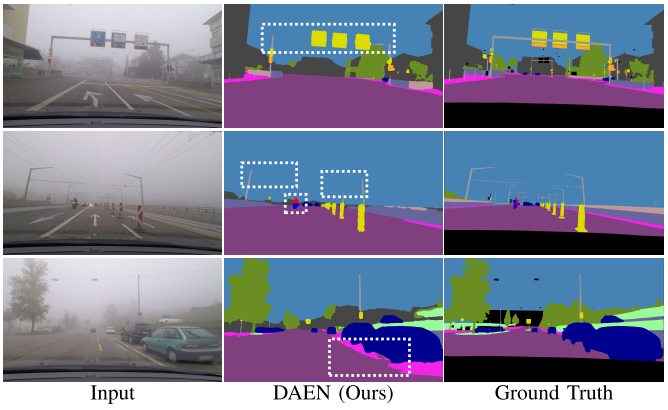

Fig. 14. Failure cases of DAEN on Foggy Zurich-test.

horizontal *poles* as *sky* and struggles with fine-grained segmentation for classes such as *motorbike* and *person*. Additionally, the model often confuses the boundaries between *sidewalk* and *road*. We conjecture that these issues arise from the relatively low occurrence of classes such as *motorbike, person*, and *pole* in the target domain dataset compared to more frequent classes like *road, sky,* and *vegetation*. Moreover, the spatial and visual similarity between *sidewalk* and *road* complicates precise boundary delineation.

To address these issues, employing re-weighted loss or focal loss could help mitigate the model's bias toward more frequent classes by increasing the weight assigned to less frequent ones. This would enhance their representation during training.

TABLE XIII

PERFORMANCE COMPARISON OF DAEN APPLIED TO TRANSFORMER-BASED NETWORKS. THE CITYSCAPES PRE-TRAINED SEGFORMER IS USED AS THE BASELINE

Methods	FZ
SegFormer	48.3
SegFormer + DAEN	**57.0**

Furthermore, incorporating class-wise prototypical contrastive learning could improve class discrimination by fostering more distinct class representations and reducing confusion between visually similar classes.

8) Applicability to Transformer-Based Networks: We evaluate the applicability of DAEN, originally designed for CNN-based networks, to transformer-based architectures by using Segformer [62], a state-of-the-art transformer model for semantic segmentation, as the baseline. Segformer's encoder comprises four transformer blocks, and its decoder utilizes an MLP. In this adaptation, HSM is applied after the first three transformer blocks, analogous to its application after the first three residual blocks in CNN-based networks. For ESPL, we perform an energy score threshold search as detailed in Sec IV-E4, selecting a threshold of -5. The results, shown in Table XIII, demonstrate an 8.7 mIoU improvement when DAEN is applied to Segformer compared to the Cityscapes pre-trained Segformer. This indicates that despite being tailored for CNN-based networks, DAEN can also be effectively applied to transformer-based networks. These findings suggest

the potential for further optimization of DAEN's components for various transformer architectures in unsupervised domain adaptation, which will be explored in future work.

V. Conclusion

In this paper, we propose DAEN, a novel framework for domain adaptive foggy scene semantic segmentation. Our framework directly adapts a network from the clear weather domain to the foggy domain without additional intermediate domain datasets or multi-stage training, and it mitigates the overconfident pseudo-labels by a confidence score in self-training. To this end, we propose two key components: HSM and ESPL. HSM matches the high-order feature statistics between clear weather and the real foggy features to obtain intermediate domain features, allowing the network to learn real fog style implicitly. ESPL provides more reliable pseudo-labels through a pixel-wise energy score, alleviating the bias and preventing the model from assigning pseudo-labels exclusively to head classes. Extensive experiments show that our method outperforms existing state-of-the-art methods on various benchmark datasets, showing a generalization ability in diverse weather conditions.

References

[1] S. Lee, T. Son, and S. Kwak, "FIFO: Learning fog-invariant features for foggy scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18889–18899.

[2] X. Ma et al., "Both style and fog matter: Cumulative domain adaptation for semantic foggy scene understanding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18900–18909.

[3] L. Liao, W. Chen, J. Xiao, Z. Wang, C.-W. Lin, and S. Satoh, "Unsupervised foggy scene understanding via self spatial–temporal label diffusion," *IEEE Trans. Image Process.*, vol. 31, pp. 3525–3540, 2022.

[4] L. Liao et al., "Only a few classes confusing: Pixel-wise candidate labels disambiguation for foggy scene understanding," in *Proc. AAAI Conf. Artif. Intell.*, 2023, vol. 37, no. 2, pp. 1558–1567.

[5] I. Shin, S. Woo, F. Pan, and I. S. Kweon, "Two-phase pseudo label densification for self-training based domain adaptation," in *Proc. Eur. Conf. Comput. Vis.* Glasgow, U.K. Cham, Switzerland: Springer, 2020, pp. 532–548.

[6] Y. Zou, Z. Yu, X. Liu, B. Kumar, and J. Wang, "Confidence regularized self-training," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 5982–5991.

[7] Y. Zou, Z. Yu, B. Vijaya Kumar, and J. Wang, "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 289–305.

[8] C. Sakaridis, D. Dai, S. Hecker, and L. V. Gool, "Model adaptation with synthetic and real data for semantic dense foggy scene understanding," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 687–704.

[9] J. Hoffman et al., "CyCADA: Cycle-consistent adversarial domain adaptation," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1989–1998.

[10] G. Kang, Y. Wei, Y. Yang, Y. Zhuang, and A. Hauptmann, "Pixel-level cycle association: A new perspective for domain adaptive semantic segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 3569–3580.

[11] Z. Murez, S. Kolouri, D. Kriegman, R. Ramamoorthi, and K. Kim, "Image to image translation for domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4500–4509.

[12] Y. Luo, P. Liu, T. Guan, J. Yu, and Y. Yang, "Significance-aware information bottleneck for domain adaptive semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 6778–6787.

[13] Y. Tsai, W. Hung, S. Schulter, K. Sohn, M. Yang, and M. Chandraker, "Learning to adapt structured output space for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7472–7481.

[14] H. Wang, T. Shen, W. Zhang, L. Duan, and T. Mei, "Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, Glasgow, U.K. Cham, Switzerland: Springer, Aug. 2020, pp. 642–659.

[15] Y. Luo, L. Zheng, T. Guan, J. Yu, and Y. Yang, "Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2507–2516.

[16] N. Araslanov and S. Roth, "Self-supervised augmentation consistency for adapting semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 15384–15394.

[17] W. Tranheden, V. Olsson, J. Pinto, and L. Svensson, "DACS: Domain adaptation via cross-domain mixed sampling," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Jan. 2021, pp. 1379–1389.

[18] Q. Zhou et al., "Context-aware mixup for domain adaptive semantic segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 2, pp. 804–817, Feb. 2023.

[19] M. Li, B. Xie, S. Li, C. Harold Liu, and X. Cheng, "VBLC: Visibility boosting and logit-constraint learning for domain adaptive semantic segmentation under adverse conditions," 2022, *arXiv:2211.12256*.

[20] X. Wu, Z. Wu, H. Guo, L. Ju, and S. Wang, "DANNet: A one-stage domain adaptation network for unsupervised nighttime semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15769–15778.

[21] H. Gao, J. Guo, G. Wang, and Q. Zhang, "Cross-domain correlation distillation for unsupervised domain adaptation in nighttime semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9903–9913.

[22] D. Brüggemann, C. Sakaridis, P. Truong, and L. Van Gool, "Refign: Align and refine for adaptation of semantic segmentation to adverse conditions," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 3173–3183.

[23] F. Li et al., "Parsing all adverse scenes: Severity-aware semantic segmentation with mask-enhanced cross-domain consistency," in *Proc. AAAI Conf. Artif. Intell.*, 2024, vol. 38, no. 12, pp. 13483–13491.

[24] S. Lee, N. Kim, S. Kim, and S. Kwak, "FREST: Feature RESToration for semantic segmentation under multiple adverse conditions," 2024, *arXiv:2407.13437*.

[25] C. Sakaridis, D. Dai, and L. Van Gool, "Semantic foggy scene understanding with synthetic data," *Int. J. Comput. Vis.*, vol. 126, pp. 973–992, May 2018.

[26] M. Cordts et al., "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.

[27] D. Dai, C. Sakaridis, S. Hecker, and L. Van Gool, "Curriculum model adaptation with synthetic and real data for semantic foggy scene understanding," *Int. J. Comput. Vis.*, vol. 128, no. 5, pp. 1182–1204, May 2020.

[28] V. Dumoulin, J. Shlens, and M. Kudlur, "A learned representation for artistic style," 2016, *arXiv:1610.07629*.

[29] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2414–2423.

[30] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1501–1510.

[31] Y. Li, N. Wang, J. Liu, and X. Hou, "Demystifying neural style transfer," 2017, *arXiv:1701.01036*.

[32] M. Liu, S. Lin, H. Zhang, Z. Zha, and B. Wen, "Intrinsic-style distribution matching for arbitrary style transfer," *Knowledge-Based Syst.*, vol. 296, Jul. 2024, Art. no. 111898.

[33] Y. Zhang, M. Li, R. Li, K. Jia, and L. Zhang, "Exact feature distribution matching for arbitrary style transfer and domain generalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 8035–8045.

[34] N. Kalischek, J. D. Wegner, and K. Schindler, "In the light of feature distributions: Moment matching for neural style transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9377–9386.

[35] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, and F. Huang, "A tutorial on energy-based learning," in *Predicting Structured Data*, vol. 1. MIT Press, 2006, pp. 191–246.

[36] W. Grathwohl, K.-C. Wang, J.-H. Jacobsen, D. Duvenaud, M. Norouzi, and K. Swersky, "Your classifier is secretly an energy based model and you should treat it like one," 2019, *arXiv:1912.03263*.

[37] W. Liu, X. Wang, D. J. Owens, and Y. Li, "Energy-based out-of-distribution detection," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, Dec. 2020, pp. 21464–21475.

[38] E. Nijkamp, M. Hill, S.-C. Zhu, and Y. N. Wu, "Learning non-convergent non-persistent short-run MCMC toward energy-based model," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[39] Y. Tian, Y. Liu, G. Pang, F. Liu, Y. Chen, and G. Carneiro, "Pixel-wise energy-biased abstention learning for anomaly segmentation on complex urban driving scenes," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 246–263.

[40] X. Liu, B. Hu, X. Liu, J. Lu, J. You, and L. Kong, "Energy-constrained self-training for unsupervised domain adaptation," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 7515–7520.

[41] B. Xie, L. Yuan, S. Li, C. H. Liu, X. Cheng, and G. Wang, "Active learning for domain adaptation: An energy-based approach," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, 2022, pp. 8708–8716.

[42] X. Li, Z. Du, J. Li, L. Zhu, and K. Lu, "Source-free active domain adaptation via energy-based locality preserving transfer," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 5802–5810.

[43] S. Herath et al., "Energy-based self-training and normalization for unsupervised domain adaptation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 11653–11662.

[44] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[45] P. Li, L. Zhao, D. Xu, and D. Lu, "Optimal transport of deep feature for image style transfer," in *Proc. 4th Int. Conf. Multimedia Syst. Signal Process.*, May 2019, pp. 167–171.

[46] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang, "Domain generalization with MixStyle," 2021, *arXiv:2104.02008*.

[47] J. P. Rolland, V. Vo, B. Bloss, and C. K. Abbey, "Fast algorithms for histogram matching: Application to texture synthesis," *J. Electron. Imag.*, vol. 9, no. 1, pp. 39–45, 2000.

[48] D.-H. Lee et al., "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2013, vol. 3, no. 2, p. 896.

[49] S. Schneider, E. Rusak, L. Eck, O. Bringmann, W. Brendel, and M. Bethge, "Improving robustness against common corruptions by covariate shift adaptation," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2020, pp. 11539–11551.

[50] C. Sakaridis, D. Dai, and L. Van Gool, "ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10765–10775.

[51] V. Nekrasov, C. Shen, and I. Reid, "Light-weight RefineNet for real-time semantic segmentation," 2018, *arXiv:1810.03272*.

[52] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[53] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "ADVENT: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 2517–2526.

[54] Y. Yang and S. Soatto, "FDA: Fourier domain adaptation for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4085–4095.

[55] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2223–2232.

[56] T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu, "Contrastive learning for unpaired image-to-image translation," in *Proc. Eur. Conf. Comput. Vis.*, Glasgow, U.K. Cham, Switzerland: Springer, 2020, pp. 319–345.

[57] Y. Deng et al., "StyTr2: Image style transfer with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11326–11336.

[58] L. Hoyer, D. Dai, and L. Van Gool, "DAFormer: Improving network architectures and training strategies for domain-adaptive semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9924–9935.

[59] L. Hoyer, D. Dai, and L. Van Gool, "HRDA: Context-aware high-resolution domain-adaptive semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, Jun. 2022, pp. 372–391.

[60] L. Hoyer, D. Dai, H. Wang, and L. Van Gool, "MIC: Masked image consistency for context-enhanced domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 11721–11732.

[61] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.

[62] E. Xie et al., "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Sys. (NIPS)*, vol. 34, Dec. 2021, pp. 12077–12090.