IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 25 July 2022, accepted 21 August 2022, date of publication 26 August 2022, date of current version 2 September 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3202190

RESEARCH ARTICLE

Pseudo Label Rectification via Co-Teaching and Decoupling for Multisource Domain Adaptation in Semantic Segmentation

SO JEONG PARK[1], HAE JU PARK[1], EUN SU KANG[1], BA HUNG NGO[1], (Graduate Student Member, IEEE), HO SUB LEE[2], (Graduate Student Member, IEEE), AND SUNG IN CHO[1], (Member, IEEE)

[1]Department of Multimedia Engineering, Dongguk University, Seoul 04620, South Korea
[2]Department of Electronics and Electrical Engineering, Daegu University, Gyeongsan-si 38453, South Korea

Corresponding author: Sung In Cho (csi2267@dongguk.edu)

This work was supported in part by the National Research Foundation of Korea (NRF) Grant funded by the Korean Government (MSIT) under Grant 2020R1C1C1009662 and in part by the Korea Institute of Police Technology (KIPoT) Grant funded by the Korea Government (KNPA) (AI Driving Ability Test Standardization and Evaluation Process Development) under Grant 092021D75000000.

ABSTRACT Multi-source Unsupervised Domain Adaptation (MUDA) is an approach aiming to transfer the knowledge obtained from multiple labeled source domains to an unlabeled target domain. In this paper, we propose a novel self-training method for MUDA, which includes pseudo label-oriented coteaching and pseudo label decoupling that are attempted for the pseudo label rectification-based MUDA for semantic segmentation. Existing ensemble-based self-training methods which are well-known approaches for MUDA use pseudo labels made from the ensemble of the predictions of multiple models to transfer the knowledge of source domains to the target domain. In these methods, information from multiple models can be contaminated, or errors from incorrect pseudo labels can be propagated. On the other hand, the proposed pseudo label-oriented coteaching trains multiple models by using pseudo labels from the peer model without any integration of pseudo labels. Simultaneously, the pseudo label decoupling method is proposed for rectification of pseudo labels, which updates the models with two pseudo labels only if they disagree. It also alleviates the problem of class imbalance in semantic segmentation, in which dominant classes lead the update for training. The effects of the proposed pseudo label-oriented coteaching and pseudo label decoupling on the performance of semantic segmentation were verified by extensive experiments. The proposed method achieved the best semantic segmentation accuracy compared with the benchmark methods. In addition, we confirmed that the prediction accuracy of small objects was greatly improved by the proposed pseudo label rectification.

INDEX TERMS Multi-source domain adaptation, semantic segmentation, unsupervised learning, self-training.

I. INTRODUCTION

Semantic segmentation is the task of classifying each pixel of an image into a corresponding class. With the recent development of deep learning, deep neural networks (DNNs) [1], [2], [3], [4], [5], [6] are widely used for semantic segmentation. The accuracy of DNNs for semantic segmentation largely depends on the quantity and quality of available training data. However, it requires a lot of time and expertise to build a dataset for semantic segmentation. For this reason, when performing semantic segmentation in different domains, Unsupervised Domain Adaptation (UDA) methods [7], [8], [9], [10], [11], [12], [13], [14], [15], [16] are widely used to increase the accuracy of semantic segmentation for the unlabeled target domain by utilizing the labeled source domain.

Based on the number of available source domains, Unsupervised Domain Adaptation (UDA) can be divided into Single-Source Unsupervised Domain Adaptation (SUDA) or

The associate editor coordinating the review of this manuscript and approving it for publication was Hongjun Su.

Multi-Source Unsupervised Domain Adaptation (MUDA). MUDA can utilize richer information from various source domains, so the performance of MUDA is generally higher than SUDA. As a simple example, MUDA with the simply combined multiple source domains is likely to improve semantic segmentation accuracy for the target domain compared to SUDA. However, it tends to be biased towards a certain source domain having a small domain gap with the target domain. Therefore, an approach that can effectively utilize multiple source domains is required to maximize the advantages of rich information in multiple source domains.

The methods of MUDA for semantic segmentation are roughly divided into adversarial learning-based methods and self-training-based methods. [14], [15], [17], [18] use adversarial learning for the semantic segmentation task. The architecture of these approaches consists of a feature extractor and a discriminator for the domain adaptation task. Specifically, the discriminator is trained to discriminate the representation between source and target domains, while the feature extractor is trained to align the representation of source and target domains in the latent space.

In the self-training-based methods [7], [19], [20], [21], models trained on multiple source domains generate predictions of the unlabeled sample, and the predictions with high confidence are set as pseudo labels. Then, the pseudo labels are used for supervised learning for the unlabeled target domain. According to [22], these methods improve prediction accuracies for the unlabeled target domain by allowing the model's decision boundary to be located in the low-density region. However, incorrect pseudo labels can degrade model performance, and self-training in SUDA cannot alleviate this artifact since it trains a model with its own pseudo labels. On the other hand, self-training in MUDA allows multiple models to be trained collaboratively by using integrated pseudo labels from different predictions of multiple models. Nevertheless, it still has the following problems for semantic segmentation. There is a class imbalance in a semantic segmentation dataset, where some classes occupy most of the labels. Therefore, some dominant classes can easily dominate the training process. Therefore, minor classes can be misclassified as dominant classes, and it causes incorrect pseudo labels. In addition, the correct prediction can be ignored when the confidence of the incorrect prediction is higher. Fig. 1 intuitively describes the problem of pseudo label-based self-training that uses the ensemble of two predictions as pseudo labels. Although (a) is more similar to ground-truth (d) for the ''road'' class, ensemble result (c) is mostly occupied by (b) because the misclassified pseudo label in (b) has higher confidence than the correctly classified pseudo label in (a). In previous ensemble-based self-training methods in MUDA, this incorrect pseudo label was the only label used for target supervised learning. In other words, if the pseudo label is generated incorrectly, there is a fatal problem that all of the multiple source domain-based models are trained incorrectly.

To solve the issues described above, we propose a novel pseudo label rectification-based self-training method

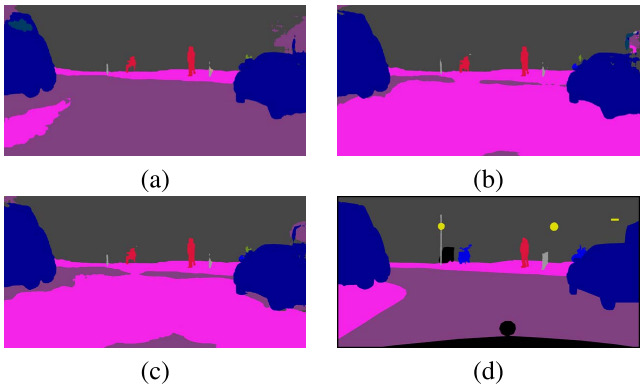

FIGURE 1. For the same image in the target domain (Cityscapes), (a) and (b) are the predictions of the models trained on GTA5 and SYNTHIA datasets, respectively. (c) is the ensemble of the two predictions, and (d) is ground-truth.

for MUDA. While the existing self-training techniques utilize the mixed pseudo labels for MUDA during training, we take a different approach that trains models using multiple pseudo labels that can be rectified by the proposed ''pseudo label-oriented coteaching'' and ''pseudo label decoupling''. In the pseudo label-oriented coteaching method, two models complement mutual errors by exchanging two pseudo labels. In this process, by using weak and strong augmentations, consistency regularization can be achieved. In addition, the proposed pseudo label decoupling, which exchanges the pseudo labels only when the predictions of two models are different, is applied to prevent the models from being overfitted by incorrect pseudo labels. Furthermore, it addresses the class imbalance problem, which occurs frequently, in semantic segmentation task. The comparison between the proposed method (yellow box) and the existing ensemble-based self-training method [7] (purple box) is illustrated in Fig. 2. The red box in Fig. 2 describes the case that the predictions of the two models are the same. In ensemble-based self-training, the incorrect pseudo label is utilized for self-training, so it interferes with obtaining knowledge of ''traffic sign'' from source domains. On the other hand, in the proposed method, training interference is reduced by excluding incorrect pseudo labels by using the pseudo label decoupling, so the knowledge of ''traffic sign'' from the source domain can be successfully utilized. The green box in Fig. 2 describes the case when two models make different predictions. In the proposed method, the part worth updating leads the loss value, thereby the effectiveness of the training is achieved.

The main contributions in this paper can be summarized as follows:

1) We introduce a novel self-training method with pseudo label rectification for the multi-source domain adaptation in semantic segmentation.

2) Pseudo label-oriented coteaching is proposed to compensate for the problems that can be caused by ensembled pseudo labels during the self-training with multiple source domains. For this, we design a self-training technique in

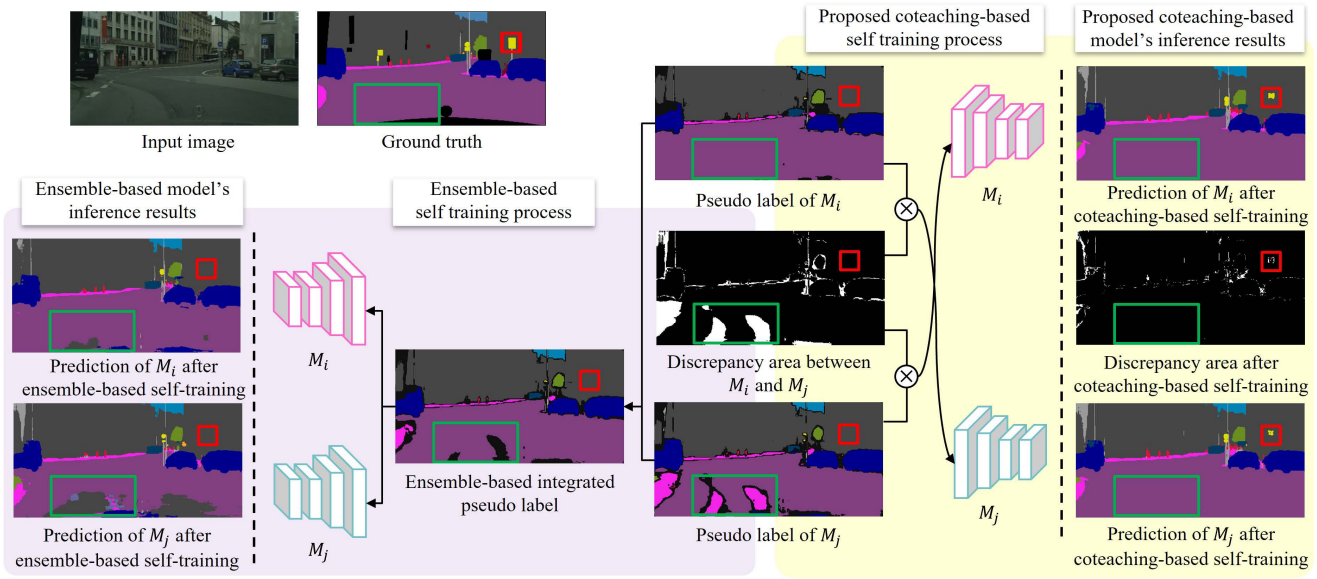

FIGURE 2. Comparison of the process utilizing multiple pseudo labels in the self-training. Left purple box: ensemble-based self-training method, right yellow box: the proposed coteaching-based method. M_i: the model trained on *i*-th source domain, M_j: the model trained on *j*-th source domain.

which two semantic segmentation models interchange their pseudo labels.

3) Pseudo label decoupling is introduced to compensate for the training inefficiency caused by the class imbalance in the data of semantic segmentation. For this, only pseudo labels that disagree with the prediction of the peer model are used for the model update.

4) We demonstrate the effectiveness of the proposed method by comparing it with state-of-the-art methods on various benchmarks including GTA5+SYHTHIA→Cityscapes, GTA5+Synscapes→Cityscapes, and GTA5+Synscapes→Mapillary.

II. RELATED WORK

A. SINGLE-SOURCE UNSUPERVISED DOMAIN ADAPTATION (SUDA)

SUDA methods have been actively studied to improve segmentation performance for the target domain by transferring knowledge of the labeled source domain to the unlabeled target domain. The domain gap for semantic segmentation is caused by differences in various elements such as style and texture between input images of different domains. SUDA methods can be divided into three approaches: approaches to reduce the domain gap in the image level [8], [23], [24], [25], [26], and feature level [9], [10], [11], [12], [27] and approaches based on self-training [13], [16], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38].

CyCADA [8], the representative method to reduce the domain gap in the image level, changes the style of the image in the source domain to be similar to that of the target domain, using Cycle-GAN [39]. FDA [25] replaces the components of the low frequency band of the source domain with the those of the target domain, so it acquires source images in which the characteristics (or styles) of the target domain are reflected. FCAN [26] adds the models for image translation named Appearance Adaptation Network (AAN). The representative methods to reduce the domain gap at the feature level are adversarial learning-based approaches. AdaptSeg [9] uses the feature-level discriminator to reduce the domain gap for semantic segmentation. FADA [12] utilizes the advanced discriminator that classifies the class-wise feature distribution of the source and target domains, unlike previous adversarial learning-based approaches, which classify the entire feature distribution of the source and target domains regardless of the class. SSF-DAN [27] proposes semantic-wise separable discriminator and class-wise adversarial loss reweighting to achieve a balanced class-wise adversarial learning process. ProDA [13] is the self-training-based SUDA method that creates prototypes for each class in the target domain. Then it corrects the noisy pseudo labels by utilizing the distance between the prototypes and the feature vectors of target samples. As a result, it obtains great performance improvement by the revised pseudo labels. Seg-Uncertainty [34] set the uncertainty with the variance of the model predictions and proposes a variance regularization term to rectify the noisy pseudo labels. LSE [35] argues that the segmentation model should generate invariant predictions to the size of the object in the image. Therefore, pseudo labels are generated using patches of variously scaled images and used for the self-training.

B. MULTI-SOURCE UNSUPERVISED DOMAIN ADAPTATION (MUDA)

MDAN [14] is an adversarial learning-based method that includes multiple domain classifiers to find the optimal decision boundary of the target domain. MADAN [15] is another

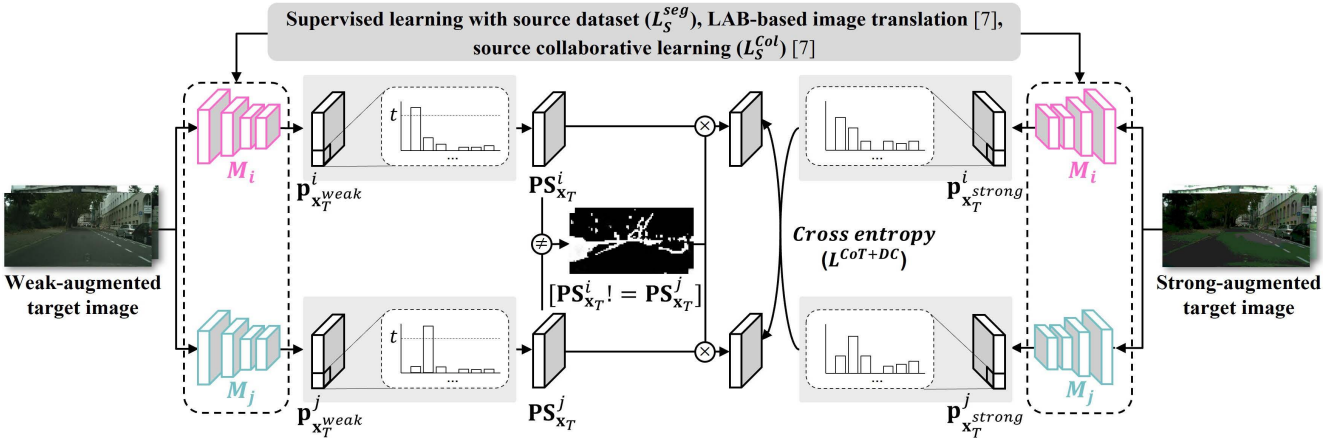

FIGURE 3. Structure of the proposed method. M_i: the model trained on *i*-th source domain, M_j: the model trained on *j*-th source domain.

MUDA method using a single segmentation model, which utilizes Cycle-GAN [39] to translate the style of the image in the source domain similar to that of the target domain, reducing the image-level domain gap. MDACL [7] trains multiple models for each source domain and integrates knowledge of multiple models by collaborative learning. To reduce the image-level domain gap, the input RGB image is transformed to LAB color space, then the mean and standard deviation of the pixel values in the source domain are changed to those of the target domain. After this, by the collaborative learning in the multiple source domains, each model learns semantic knowledge from the peer model. Specifically, multiple models trained on their corresponding source domain make predictions for the same source domain sample and reduce the distribution gap of multiple predictions (views). In addition to this, for the target domain, collaborative learning is conducted based on the ensemble-based self-training method.

C. LEARNING FROM NOISY LABELS

As previous papers [13], [33], [34] have succeeded in improving accuracy by pseudo label rectification on self-training, we explore the data cleaning method that can be effectively applied to the self-training method in MUDA for semantic segmentation.

Co-teaching [40] that is a method for cleaning noisy labels, trains two differently initialized models for the same dataset. Data cleaning is achieved by selecting small loss instances from the peer model and updating the current model only with these selected instances. Noisy labels can be excluded from training by this process, and by reflecting the opinion of the peer model, it is possible to prevent the error propagation caused by the incorrect model's own predictions or noisy labels.

Decoupling [41] is a method that updates models only when predictions for the same input of two differently initialized models disagree with each other. The general DNN updates its parameters based on the errors between model predictions and labels. This general training approach is vulnerable to noisy labels, especially as the model matures. Decoupling avoids this training interference caused by noisy labels by updating the target model only when the disagreement occurs.

III. PROPOSED METHOD

As shown in Fig. 3, the proposed MUDA uses MDACL [7] as a baseline that has two models M_i and M_j supervised from two source domains S_i and S_j and borrows LAB-based image translation and source collaborative learning. On the other hand, unlike the previous method, we propose an advanced pseudo label rectification method in the self-training process with the target domain. The proposed method extracts two pseudo labels from two models supervised from different source domains, and exchanges pseudo labels between peer models during self-training, which is named coteaching. To alleviate the class imbalance during the coteaching and the training interference problem by the incorrect pseudo labels, the proposed pseudo label decoupling is applied.

A. PSEUDO LABEL-ORIENTED COTEACHING FOR THE COMPLEMENTARY SELF-TRAINING

In general, self-training-based domain adaptation (DA) methods [7], [13], [16], [42], [43], [44] use the predictions having high confidence as pseudo labels, and the model is trained for the target domain in supervised-manner by using the pseudo labels as the ground-truth. These pseudo labels are often noisy due to the domain gap between the source and target domains. When a self-training-based model updates only using its own pseudo label, two problems may occur: first, the update of the model can be biased toward learning about easy cases, because only predictions with high confidence will be used as pseudo labels and easy cases lead the training of models. Second, when the pseudo label is incorrect, the efficiency of training is greatly reduced. In [7], the models are trained based on multiple source domains, and the pseudo label is generated by the ensemble of the

predictions from multiple models. This method has the effect of solving the above-mentioned problems by utilizing the ensembled pseudo labels generated by integrating the multiple predictions of multiple models. However, since the ensembled pseudo labels consider only the maximum probability after combining the predictions of multiple models, they may be biased toward the result of a certain model that provides higher confidence as shown in Fig. 1. Even if a model trained on a source domain makes the correct prediction for the target input, the ensembled result can be determined differently by its peer model. And in this case, valuable information for the model update can be ignored. In addition, the incorrectly generated pseudo labels make the errors continuously propagated. To solve this problem, unlike the previous papers [7], [21] that combine multiple predictions for the generation of the pseudo label, we propose a new pseudo label-oriented coteaching that interchanges pseudo labels. In the proposed coteaching method, two types of augmented images $\mathbf{x}_T^{weak}$ and $\mathbf{x}_T^{strong}$ are generated, which are weak and strong augmented versions of the same image of the target domain, respectively. The weak augmentation includes random resizing, random crop, and random horizontal flip. The strong augmented images are created by applying random contrast adjustment, random brightness adjustment, random color balancing, histogram equalization, random posterization, and random sharpness adjustment to the weak augmented image. $\mathbf{x}_T^{weak}$ is used for generating the pseudo label from the prediction of the target domain while $\mathbf{x}_T^{strong}$ is used to extract the prediction for calculating cross-entropy loss with the extracted pseudo label. This gives a consistency regularization effect so that the model makes consistent predictions of the perturbed data and finds the effective manifold space for semantic segmentation [45], [46].

The model M_i which is trained on the i-th source domain extracts two predictions $\mathbf{p}^i_{\mathbf{x}_T^{weak}}$ and $\mathbf{p}^i_{\mathbf{x}_T^{strong}}$ for each pixel of the two augmented versions of an input image. Following (1), the pseudo label $\mathbf{PS}^i_{\mathbf{x}_T}$ can be generated by $\mathbf{p}^i_{\mathbf{x}_T^{weak}}$,

$$\mathbf{PS}^{i,(h,w)}_{\mathbf{X}_T} = \underset{c}{\text{argmax}} \left(\mathbf{p}^{i,(h,w)}_{\mathbf{x}_T^{weak}} \right), \tag{1}$$

where c is the index of the class. h and w denote the position indexes of $\mathbf{x}_T^{weak}$. To exclude unstable pseudo labels from training, we set threshold τ and only use predictions with higher confidence than τ for self-training. To prevent the problem caused by the class imbalance of $\mathbf{PS}_{\mathbf{x}_T}$, the threshold τ is determined by considering the ratio of each class. The probability of the top 50% of each class is referred to τ_{soft} and a constant value is referred to τ_{hard} (selected to 0.9 in the proposed method). The final threshold value τ is determined by $\min\left(\tau_{soft}, \tau_{hard}\right)$ as in [7], [47], and [16].

The pseudo labels of each model are used to update the peer model. In other words, $\mathbf{PS}^i_{\mathbf{x}_T}$ and $\mathbf{PS}^j_{\mathbf{x}_T}$ are used to update M_j and M_i, respectively. The cross-entropy loss for self-training of M_i using the pseudo label is as follows:

$$L^{CoT}_{M_i} = -\sum_{h,w}^{H,W} \sum_{c}^{C} \mathbf{PS}\mathbf{x}_T^{j,(h,w,c)} log \left(\mathbf{p}^{i,(h,w,c)}_{\mathbf{x}_T^{strong}} \right), \tag{2}$$

where H and W are the height and width of the input image, respectively, and C is the number of classes. For the self-training of M_j, $L^{CoT}_{M_j}$ is used by exchanging i and j in (2).

By learning with its peer model's pseudo labels, the proposed coteaching method prevents both i-th and j-th models from being updated equally by incorrect pseudo labels. If two models make different predictions for the same target image, each model has an opportunity to learn different opinions. Therefore, the two pseudo labels continuously have the chance to vary, so that the overfitting problem caused by the incorrect pseudo labels can be relieved.

B. PSEUDO LABEL DECOUPLING FOR AVOIDING BIASED LEARNING

With the coteaching method described in Section III-A, we propose the pseudo label decoupling to rectify the incorrect pseudo label to minimize the impact of the incorrect prediction. In addition, this pseudo label decoupling can also alleviate the class imbalance problem. In Section III-A, two pseudo labels $\mathbf{PS}^i_{\mathbf{x}_T}$ and $\mathbf{PS}^j_{\mathbf{x}_T}$ extracted from two models M_i and M_j can appear in five cases as follows and the update of M_i and M_j is performed only for **cases 3-5**.

Case 1: Both $\mathbf{PS}^i_{\mathbf{x}_T}$ and $\mathbf{PS}^j_{\mathbf{x}_T}$ are correct, $\mathbf{PS}^i_{\mathbf{x}_T} = \mathbf{PS}^j_{\mathbf{x}_T}$
Case 2: Both $\mathbf{PS}^i_{\mathbf{x}_T}$ and $\mathbf{PS}^j_{\mathbf{x}_T}$ are incorrect, $\mathbf{PS}^i_{\mathbf{x}_T} = \mathbf{PS}^j_{\mathbf{x}_T}$
Case 3: $\mathbf{PS}^i_{\mathbf{x}_T}$ is correct, $\mathbf{PS}^j_{\mathbf{x}_T}$ is incorrect, $\mathbf{PS}^i_{\mathbf{x}_T} \neq \mathbf{PS}^j_{\mathbf{x}_T}$
Case 4: $\mathbf{PS}^i_{\mathbf{x}_T}$ is incorrect, $\mathbf{PS}^j_{\mathbf{x}_T}$ is correct, $\mathbf{PS}^i_{\mathbf{x}_T} \neq \mathbf{PS}^j_{\mathbf{x}_T}$
Case 5: Both $\mathbf{PS}^i_{\mathbf{x}_T}$ and $\mathbf{PS}^j_{\mathbf{x}_T}$ are incorrect, $\mathbf{PS}^i_{\mathbf{x}_T} \neq \mathbf{PS}^j_{\mathbf{x}_T}$

Therefore, (2) can be expressed as follows:

$$L^{CoT+DC}_{M_i} = -\sum_{h,w}^{H,W} \left[\mathbf{PS}^{i,(h,w)}_{\mathbf{x}_T} \neq \mathbf{PS}^{j,(h,w)}_{\mathbf{x}_T} \right] \sum_{c}^{C} \mathbf{PS}^{j,(h,w,c)}_{\mathbf{x}_T} log \left(\mathbf{P}^{i,(h,w,c)}_{\mathbf{x}_T^{strong}} \right), \tag{3}$$

where [.] is the indication function.

Case 1 is a situation in which the two models already predict the correct answers. It usually occurs in easy cases, such as large segments with dominant classes. In this case, the knowledge of the correct prediction is already sufficiently obtained from the source domain. On the other hand, there is a possibility that the dominant class can cause a problem in which the model is learned for mainly easy classes. **Case 2** generally occurs when classes with a small area (ex. traffic sign and traffic light) in the input image are overwhelmed by dominant classes (ex. building and vegetation) as shown in Fig. 2. As the two models agree with incorrect predictions of each other, updating using these pseudo labels disturbs the training. By excluding these two cases from the update, the performance degradation of self-training due to the class

imbalance and incorrect pseudo labels can be alleviated. **Case 3** and **Case 4** are the cases where complementary learning by coteaching can be effectively achieved. In **Case 5**, both pseudo labels are incorrect, but they increase the entropy of the two incorrect predictions, preventing training from being biased toward a specific incorrect class. By the coteaching-based update of M_i and M_j only in **Cases 3-5**, the error propagation caused by the incorrect pseudo label can be prevented.

C. NETWORK ARCHITECTURE AND TRAINING METHODOLOGY

The overall process of the proposed method is described in Fig. 3. The weights of two models are initialized to the pre-trained weights based on AdaptSeg [9]. Following the baseline [7], image translation in LAB color space, supervised learning on source domains, and source collaborative learning are adopted.

The cross-entropy loss for supervised learning on the source domain and the Kullback-Leibler Divergence [48] loss for the source collaborative learning are calculated as following (4) and (5), respectively.

$$L_{S_i}^{seg} = -\sum_{h,w}^{H,W}\sum_{c}^{C} \mathbf{y}_{S_i}^{(h,w,c)} log\left(M_i\left(\mathbf{x}_{S_i}^{(h,w,c)}\right)\right), \tag{4}$$

$$L_{S_{i\to j}}^{Col} = -\sum_{h,w}^{H,W} \sigma\left(M_i\left(\mathbf{x}_{S_i}^{(h,w)}\right)\right) log\left(\frac{\sigma\left(M_j\left(\mathbf{x}_{S_i}^{(h,w)}\right)\right)}{\sigma\left(M_i\left(\mathbf{x}_{S_i}^{(h,w)}\right)\right)}\right), \tag{5}$$

where $\mathbf{x}_{S_i}$ and $\mathbf{y}_{S_i}$ denote the input image and ground-truth of the i-th source domain, respectively. H, W, and C are the same as those in (2), and $\sigma(.)$ indicates the softmax function. $L_{S_i}^{seg}$ and $L_{S_{i\to j}}^{Col}$ are calculated for all source domains and optimized by joint learning. The final loss in the source domain is defined as follows:

$$L_{source} = L_{S_i}^{seg} + L_{S_j}^{seg} + L_{S_{i\to j}}^{Col} + L_{S_{j\to i}}^{Col}. \tag{6}$$

Thereafter, two predictions $\mathbf{p}_{\mathbf{x}_T^{weak}}^{i}$ and $\mathbf{p}_{\mathbf{x}_T^{weak}}^{j}$ about the same weak augmented target image $\mathbf{x}_T^{weak}$, are extracted by the two models M_i and M_j. Then, two pseudo labels $P\mathbf{S}_{\mathbf{x}_T}^{i}$ and $\mathbf{PS}_{\mathbf{x}_T}^{j}$ are derived by $\mathbf{p}_{\mathbf{x}_T^{weak}}^{i}$ and $\mathbf{p}_{\mathbf{x}_T^{weak}}^{j}$, as in (1). $\mathbf{p}_{\mathbf{x}_T^{strong}}^{i}$ and $\mathbf{p}_{\mathbf{x}_T^{strong}}^{j}$ are also extracted by M_i and M_j from the strong augmented image $\mathbf{x}_T^{strong}$, and used to calculate (3) with the pseudo labels. $L_{M_i}^{CoT+DC}$ and $L_{M_j}^{CoT+DC}$ of two models M_i and M_j are calculated by (3) and jointly learned in the target domain as follows:

$$L_{target} = L_{M_i}^{CoT+DC} + L_{M_j}^{CoT+DC}, \tag{7}$$

The total loss used to update the model is computed as follows:

$$L^{total} = L^{source} + \lambda L^{target} * \frac{N_{cur}}{N_{max}}, \tag{8}$$

TABLE 1. The key characteristics of prior and proposed methods.

Method	MUDA method	Image level	Feature level	Self training	Multi pseudo labels
AdaptSeg [9]			✓		-
AdaptPatch [10]			✓		-
ADVENT [11]			✓		-
CyCADA [8]		✓			-
FADA [12]			✓		-
FDA [25]		✓			-
FCAN [26]		✓	✓		-
SSF-DAN [27]			✓		-
Seg-Uncertainty [34]				✓	
LSE [35]				✓	✓
PyCDA [36]				✓	
CRST [37]				✓	
MLSL [38]				✓	
MDAN [14]	✓		✓		-
MADAN [15]	✓	✓	✓		-
MDACL [7]	✓	✓	✓	✓	
Ours	✓	✓	✓	✓	✓

where N_{cur} and N_{max} denote the current iteration number and the maximum number of iterations, respectively, and $\frac{N_{cur}}{N_{max}}$ is the weight parameter to increase the influence of L^{target} as the M_i and M_j become mature. This is used because the pseudo labels at the beginning of training are unstable and unreliable. λ is an additional parameter for adjusting the weights of L^{source} and L^{target}. The setting of λ will be described in Section IV.

D. COMPARISON OF KEY CHARACTERISTICS BETWEEN PREVIOUS AND PROPOSED METHODS

The key characteristics of previous and proposed methods are compared in Table 1. The proposed method utilizes multiple source domains, so it gets larger coverage of target representation than SUDA methods [8], [9], [10], [11], [12], [25], [26], [27], [34], [35], [36], [37], [38] that are distinguished by the first column. We compare the proposed method with [7], [14], [15] in detail because the target of our method is MUDA.

There are several MUDA methods, but most are proposed for classification tasks, such as MDAN [14]. MADAN [15] is proposed specifically for semantic segmentation, and it succeeds in domain adaptation for semantic segmentation through image level and feature level domain alignment. However, in the case of MADAN, Cycle-GAN [39] is used for image style translation, which requires the huge complexity of training additional networks. To solve this problem, MDACL [7] proposes a method for translating source domain images into target domain style in the LAB color space, which is more convenient to be applied to the training process. MDACL also proposes the source collaborative learning through Kullback-Leibler Divergence and the target collaborative learning by self-training. However, the self-training of MDACL has a problem with dealing with incorrect labels and class imbalances, so the proposed method solves this problem in the target collaborative learning by using multiple versions of pseudo labels. With two pseudo labels, the two models are trained complementarily and get the peer-review effect through the proposed pseudo label-oriented coteaching,

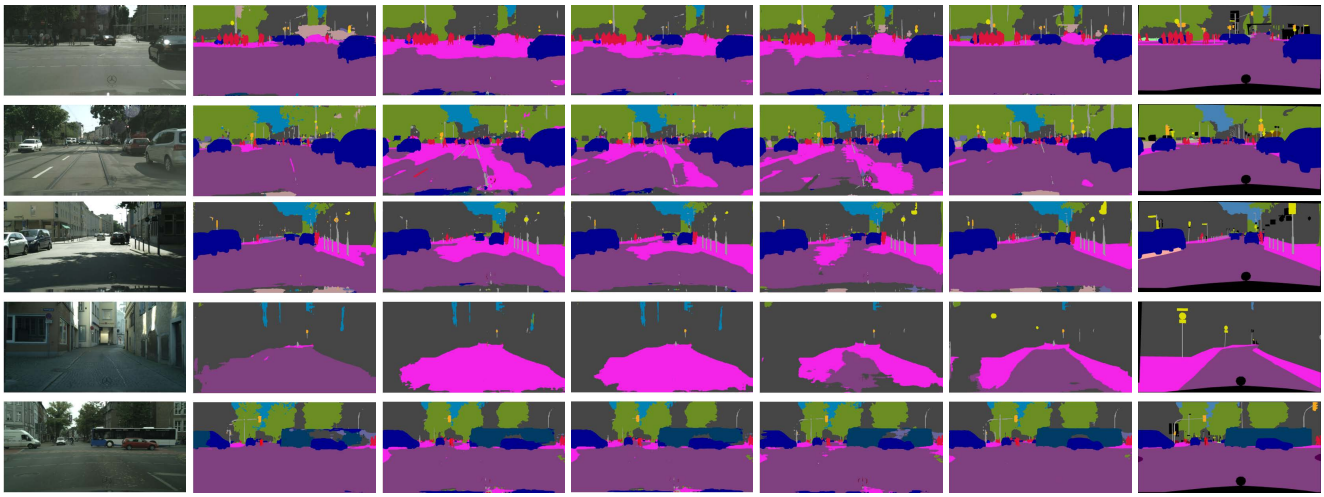

FIGURE 4. Visual comparison of the baseline models and the proposed model on the Cityscapes validation set on the setting of $\lambda = 0.5$.

maximizing the advantage of using multiple source domains. This could be a solution to the common problem of self-training in which models are continually biased by incorrectly generated pseudo labels. In addition, by using multiple pseudo labels, it is possible to obtain information on the disagreement between multiple pseudo labels, which is utilized for the proposed pseudo label decoupling. It helps to solve training difficulties due to class imbalance, which commonly occurs in semantic segmentation.

IV. EXPERIMENTS

A. TRAINING SETTING

1) DATASETS

in this experiment, GTA5 [49], SYNTHIA [50], and Synscapes [51] that are widely used for DA in semantic segmentation, were used as the source domains. The real-world semantic segmentation datasets Cityscapes [52] and Mapillary vistas [53] were used as the target domains. The GTA5 dataset provides 24,966 pixel-level segmentation labels synthesized from an open-world game and follows the class composition of the Cityscapes. The SYNTHIA dataset includes semantic segmentation labels automatically generated from images in the virtual world. The total number of images is 9,400, providing annotations for 16 classes matched with the Cityscapes. The Synscapes dataset includes 25,000 photo-realistic rendered images and segmentation labels. The Cityscapes provides the RGB images and the semantic segmentation labels of 50 different cities in the real world, and the number of samples for the training set and validation set are 2,975 and 500, respectively. The Mapillary vistas dataset consists of 25,000 images of street scenes and manually annotated segmentation labels. Following the settings of the existing SUDA and MUDA methods [7], [8], [9], [10], [11], [12], [13], [14], [15], our implementation utilized all images and labels of the training set of source domain datasets (GTA5, SYNTHIA, and Synscapes) and only images (without labels) of the training set of target domain datasets (Cityscapes or Mapillary) for training. To verify the performance of the proposed and the benchmark methods, only image-label pairs of the validation set of target datasets were used.

TABLE 2. Hyper parameter setting.

SGD optimizer setting	Weight decay	$2.0 \times e^{-4}$
	Momentum	0.9
	Learning rate	$0.5 \times e^{-4}$
λ setting	GTA5+SYNTHIA→Cityscapes	0.5
	GTA5+Synscapes→Cityscapes	1.0
	GTA5+Synscapes→Mapillary	0.5

TABLE 3. Ablation study for the proposed method. The source domain datasets are GTA5 and SYNTHIA, and the target domain dataset is Cityscapes.

Pretrained	Ensemble	Coteaching	Decoupling	Strong aug.	mIoU	Gain
✓					48.4	0.0
✓	✓				51.3	+2.9
✓	✓		✓		53.2	+4.8
✓	✓			✓	54.4	+6.0
✓		✓		✓	55.4	+7.0
✓		✓	✓	✓	56.7	+8.3

2) IMPLEMENTATION DETAILS

our method was implemented in Pytorch. For the segmentation network, Deeplabv2 [2] with Resnet-101 [54] was used. The values of the hyperparameters are described in Table 2. The optimizers for the training of segmentation models were unified with Stochastic Gradient Descent (SGD) [55]. The weight decay and the momentum were set to $2.0 \times e^{-4}$ and

TABLE 4. The quantitative comparison of the proposed method with SOTA methods. This table shows the mIoU values for 16 classes of the Cityscapes validation set. In the "source" column, G and Y mean GTA5 and SYNTHIA datasets, respectively. All results of the benchmark methods used for comparison were collected from the previous publications [7], [9], [10], [11], [12], [15].

Method	Source	Target	Road	Sidewalk	Building	Wall	Fence	Pole	Traffic light	Traffic sign	Vegetation	Sky	Person	Rider	Car	Bus	Motorcycle	Bicycle	mIoU
AdaptSeg [9]	G	Cityscapes	91.4	48.4	81.2	27.4	21.2	31.2	**35.3**	16.1	**84.1**	78.2	57.7	28.2	**85.9**	43.5	23.9	16.9	48.2
AdaptPatch [10]			**92.3**	**51.9**	82.1	29.2	25.1	24.5	33.8	**33.0**	82.4	82.2	58.6	27.2	84.3	**46.3**	29.5	32.3	**50.9**
ADVENT [11]			89.4	33.1	81.0	26.6	26.8	27.2	33.5	24.7	83.9	78.8	**58.7**	**30.5**	84.8	44.5	**31.6**	**32.4**	49.2
CyCADA [8]			85.2	37.2	76.5	21.8	15.0	23.8	22.9	21.5	80.5	60.7	50.5	9.0	76.9	28.2	4.5	0.0	38.4
FADA [12]			**92.3**	51.1	**83.7**	**33.1**	**29.1**	**28.5**	28.0	21.0	82.6	**85.3**	55.2	28.8	83.5	37.4	21.1	15.2	48.5
AdaptSeg [9]	Y		84.0	40.5	79.3	**10.4**	0.2	22.7	6.5	8.0	78.3	82.7	56.3	22.4	74.0	33.2	18.9	**34.6**	40.8
AdaptPatch [10]			82.4	38.0	78.6	8.7	**0.6**	26.0	3.9	11.1	75.5	**84.6**	53.3	21.6	71.4	32.6	19.3	31.7	40.0
ADVENT [11]			**85.6**	**42.2**	79.7	8.7	0.4	25.9	5.4	8.1	80.4	84.1	**57.9**	**23.8**	73.3	36.4	14.2	33.0	41.2
CyCADA [8]			66.2	29.6	65.3	0.5	0.2	15.1	4.5	6.9	67.1	68.2	42.8	14.1	51.2	12.6	2.4	20.7	29.2
FADA [12]			84.5	40.1	**83.1**	4.8	0.0	**34.3**	**20.1**	**27.2**	**84.8**	84.0	53.5	22.6	**85.4**	**43.7**	**26.8**	27.8	**45.2**
MDAN [14]	G + Y		64.2	19.7	63.8	13.1	19.4	5.5	5.2	6.8	71.6	61.1	42.0	12.0	62.7	2.9	12.8	8.1	29.4
MADAN [15]			86.2	37.7	79.1	20.1	17.8	15.5	14.5	21.4	78.5	73.4	49.7	16.8	77.8	28.3	17.7	27.5	41.4
AdaptSeg [9]			87.1	46.0	82.1	19.7	1.0	41.3	38.7	19.7	85.9	79.6	**65.1**	**29.8**	87.8	43.7	**23.7**	24.0	48.5
MDACL [7]			-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	54.0
Ours			**94.8**	**66.3**	**86.2**	**36.1**	**22.6**	**43.7**	**46.9**	**40.9**	**88.6**	**88.9**	63.1	23.5	**89.7**	**53.8**	20.9	**42.7**	**56.8**

TABLE 5. The quantitative comparison of the proposed method with SOTA methods on GTA5 + Synscapes → Cityscapes and GTA5 + Synscapes → Mapillary settings. This table shows the mIoU values for 19 classes of the validation sets of target domains. In the 'source' column, G and S mean GTA5 and Synscapes datasets, respectively. All results of the benchmark methods used for comparison were collected from the previous publications [7]. DataComb: the method of simply combining multiple source domains.

Method	Source	Target	Road	Sidewalk	Building	Wall	Fence	Pole	Traffic light	Traffic sign	Vegetation	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	Motorcycle	Bicycle	mIoU
DataComb	G + S	Cityscapes	85.1	36.9	84.1	39.0	33.3	38.7	43.1	40.2	84.8	37.1	82.4	65.2	37.8	69.4	43.4	38.8	34.6	33.2	53.1	51.6
AdaptSeg [9]			89.3	47.3	83.6	40.3	27.8	39.0	44.2	42.5	86.7	45.5	84.5	63.1	38.0	79.4	34.9	48.3	42.1	30.7	52.3	53.7
ADVENT [10]			91.8	49.0	84.6	39.4	31.5	39.9	42.9	43.5	86.3	45.1	84.6	65.3	41.0	87.1	37.9	49.2	31.0	30.3	48.8	54.2
MDAN [14]			92.4	56.1	86.8	42.7	32.9	39.3	48.0	40.3	87.2	47.2	**90.5**	64.1	35.9	87.8	33.8	48.6	39.0	27.6	49.2	55.2
MADAN [15]			94.1	61.0	86.4	43.3	32.1	40.6	49.0	44.4	87.3	**47.7**	89.4	61.7	36.3	87.5	35.5	45.8	31.0	33.5	52.1	55.7
MDACL [7]			93.6	59.6	**87.1**	**44.9**	**36.7**	42.1	49.9	2.5	87.7	47.6	89.9	63.5	**40.3**	88.2	**41.0**	**58.3**	**53.1**	**37.9**	57.7	59.0
Ours			**94.8**	**63.7**	86.9	40.8	34.4	**47.9**	**53.4**	**54.9**	**88.9**	44.1	86.9	**70.9**	35.2	**88.6**	38.4	55.0	40.5	37.3	**64.2**	**59.3**
DataComb		Mapillary	77.7	30.9	75.2	27.0	27.5	33.4	37.2	37.3	76.9	43.1	93.3	55.8	38.0	72.5	38.4	40.2	2.8	36.9	42.3	46.7
AdaptSeg [9]			84.2	33.4	78.0	27.9	34.0	38.0	41.6	39.4	78.6	34.5	92.7	46.9	41.6	81.9	38.3	39.0	3.6	41.5	40.5	48.2
ADVENT [10]			82.7	36.2	78.0	27.1	31.2	38.4	40.8	40.2	80.8	44.2	96.0	47.1	43.5	82.3	39.0	39.3	5.0	42.0	40.3	49.2
MDACL [7]			**88.4**	40.1	81.9	32.4	39.8	41.4	42.2	42.7	80.1	**46.4**	95.6	58.2	**48.5**	84.7	**46.6**	**45.5**	11.7	**46.9**	42.4	53.4
Ours			88.3	**49.1**	**83.0**	**34.0**	**40.2**	**48.9**	**51.5**	**63.8**	**83.7**	45.3	**97.0**	**66.2**	43.4	**86.4**	42.2	43.2	**11.8**	41.6	**50.2**	**56.3**

0.9, respectively. The initial learning rate was $0.5 \times e^{-4}$ for the proposed MUDA framework, and $1.25 \times e^{-4}$ for the pre-training process. On the other hand, the Adam optimizer was utilized for the training of the discriminator in the pre-training process with the initial learning rate of $0.5 \times e^{-4}$ and the betas (0.9, 0.99). The λ values for each experiment were set empirically, and the values are described in Table 2. The mean Intersection over Union (mIoU) was used for the quantitative evaluation of semantic segmentation.

B. ABLATION STUDY

Table 3 shows the results of the ablation study to prove performance improvement by each component in the proposed method. The λ in (8) which was used for

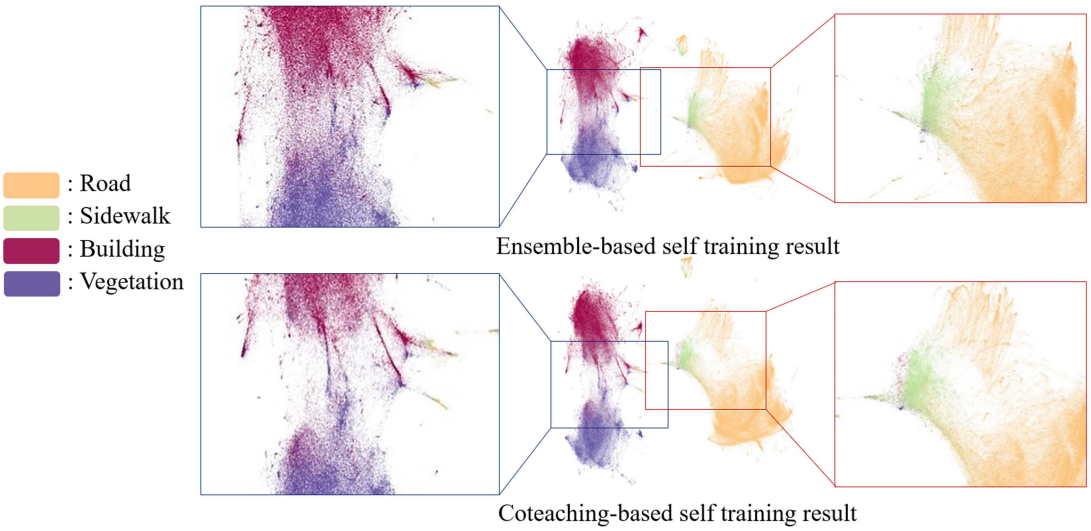

FIGURE 5. Comparison of embedding space between the ensemble-based self-trained model and the proposed model for four classes (road, sidewalk, building, and vegetation) which are described in orange, green, red, and purple points, respectively.

balancing the weights between L^{source} and L^{target} was set to 1.0 in this experiment. The pre-trained model described in Section III-C achieved 48.4% mIoU. After ensemble-based self-training, 2.9% accuracy improvement was obtained while applying the strong augmentation described in III-A to ensemble-based self-training resulted in 6.0% improvement. When we changed the ensemble-based self-training to co-teaching, we obtained 55.4% mIoU which is 1.0% higher than the ensemble-based training result. With the proposed pseudo label decoupling, an additional 1.3% performance improvement was achieved. When we applied the pseudo label decoupling to the ensemble-based self-training, the performance improved to 53.2% mIoU. As a result, the proposed method provided 56.7% of mIoU that is 8.3% higher than that of the pre-trained model [9].

Fig. 4 compares the results of the baseline model (ensemble-based self-trained model) and the proposed model. Overall, the AdaptSeg-GTA5 made more similar predictions to the ground-truth than the AdaptSeg-SYNTHIA. However, the ensemble results (AdaptSeg-ensemble) contained inaccurate results due to the strong influence of the AdaptSeg-SYNTHIA. The results of the baseline (ensemble-based self-trained model) with those pseudo labels seemed to have been biased to the model providing higher confidence, as shown in Fig. 4. On the other hand, the proposed method showed results that are not biased toward either model because the proposed coteaching method utilized both pseudo labels for the complementary self-training. In addition, in the fourth row of Fig. 4, the proposed method succeeded in predicting the ''traffic sign'' class that both AdaptSeg-GTA5 and AdaptSeg-SYNTHIA failed to predict. The reason for the performance improvement is that the pseudo label decoupling could remove the training interference of the dominant class (''building''), allowing the model to obtain information of the small object (''traffic sign'').

C. COMPARISON WITH SOTA METHODS

The performances of various SUDA methods [8], [9], [10], [11], [12] and MUDA methods [7], [9], [14], [15] on GTA5+SYNTHIA → Cityscapes setting were compared in Table 4. In SUDA, AdaptPatch [10] achieved the best results of 50.9% mIoU in the setting of GTA5 → Cityscapes, while FADA [12] achieved the best results of 45.2% mIoU in the setting of SYNTHIA → Cityscapes. In MUDA methods, we used both GTA5 and SYNTHIA as the source domains and the Cityscapes as the target domain. MDAN [14] and MADAN [15] achieved 29.4% and 41.4% mIoUs, respectively. For comparison with MUDA, AdaptSeg [9] that is a SUDA method, was expanded to have two classifiers trained on two source domains GTA5 and SYNTHIA. As a result, 48.5% mIoU was obtained. The performance of MDACL [7] which is the baseline of the proposed method showed 54.0% mIoU. Our method achieved the best score, 56.8% mIoU with the setting of $\lambda = 0.5$. In addition, a noticeable performance improvement was obtained in small object classes such as the traffic light and traffic sign as shown in Table 4.

In Table 5, the IoU scores of 19 classes on GTA5+Synscapes → Cityscapes and GTA5+Synscapes → Mapillary settings were shown. The λ values for the proposed method were 1.0 and 0.5 for the Cityscapes and Mapillary target settings, respectively. For the GTA5+Synscapes → Cityscapes setting, the result of DataComb that utilizes simply combined source domains was 51.6% mIoU. The recent MUDA method, MDACL achieved 59.0% mIoU which was improved by 7.4% compared with DataComb. Our method achieved 59.3% mIoU which was the best score in this setting.

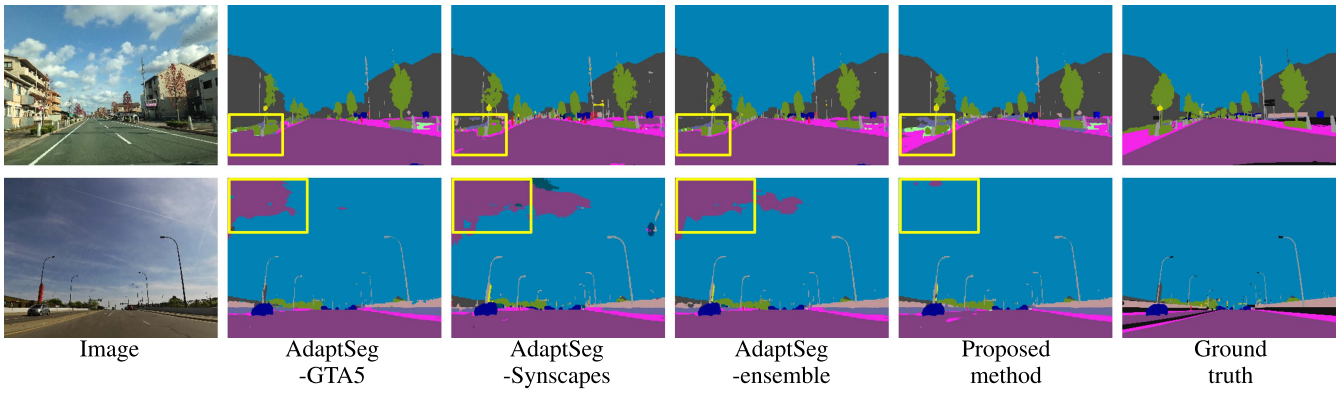

FIGURE 6. Visual analysis for Case 2. In yellow boxes, both AdaptSeg-GTA5 and AdaptSeg-Synscapes made the incorrect pseudo labels as the same class. When these two pseudo labels were integrated (AdaptSeg-ensemble), the incorrect pseudo labels were made. However, by the proposed decoupling method, these incorrect pseudo labels were not utilized for the self-training. As a result, the proposed method successfully induced the two models to obtain the correct knowledge from the source domains without being biased by the incorrect pseudo labels.

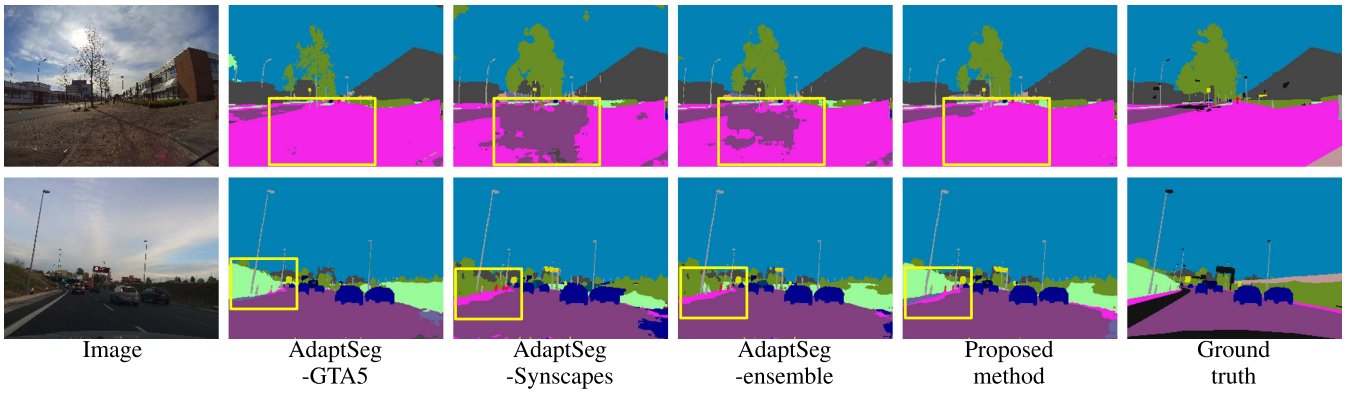

FIGURE 7. Visual analysis for Case 3. In here, AdaptSeg-GTA5 made the correct pseudo labels and AdaptSeg-Synscapes made the incorrect pseudo labels. Since the confidence of AdaptSeg-Synscapes was higher than that of AdaptSeg-GTA5, the ensembled pseudo labels (AdaptSeg-ensemble) included the incorrect pseudo labels. If we update the models only with these wrong pseudo labels, AdaptSeg-Synscapes will continue to be confident about the wrong prediction, and AdaptSeg-GTA5 will be contaminated by the incorrect pseudo labels. On the other hand, the proposed pseudo label-oriented coteaching method utilized both pseudo labels for the self-training and successfully corrected errors through the peer-review effects.

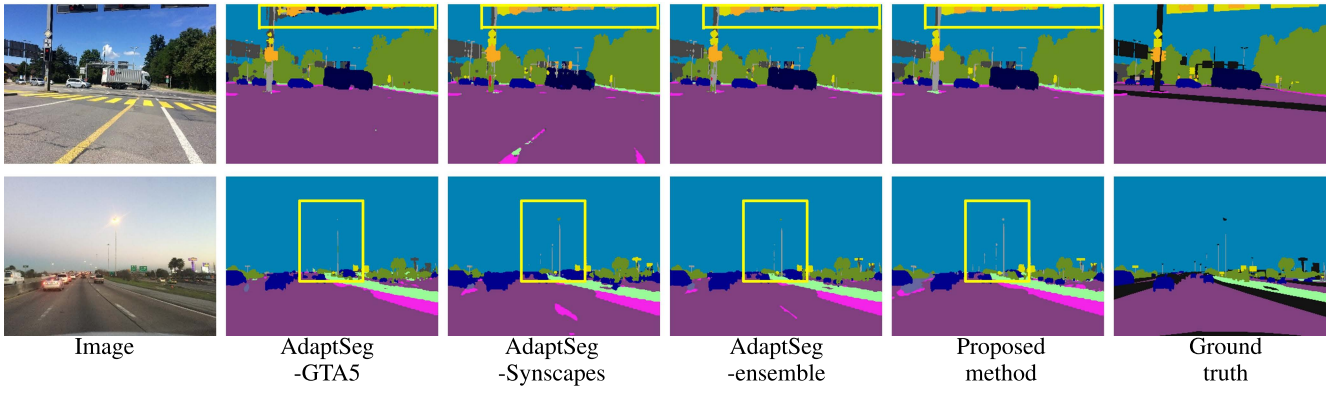

FIGURE 8. Visual analysis for Case 4. AdaptSeg-Synscapes made the correct predictions but AdaptSeg-GTA5 made the incorrect predictions, as opposed to Case 3. Most of the ensembled pseudo labels followed the predictions of AdaptSeg-Synscapes at this time. The proposed method utilized both the correct pseudo labels and the incorrect pseudo labels for the self-training and showed the most ideal prediction results.

For the GTA5+Synscapes → Mapillary setting, DataComb provided 46.7% mIoU. AdaptSeg [9] and ADVENT [11] achieved 48.2% and 49.2% mIoU, respectively, and MDACL obtained 53.4% mIoU. The proposed method exceeded the performance of other benchmarks with 56.3% mIoU. Similar to the results in Table 4, the accuracy of the proposed method

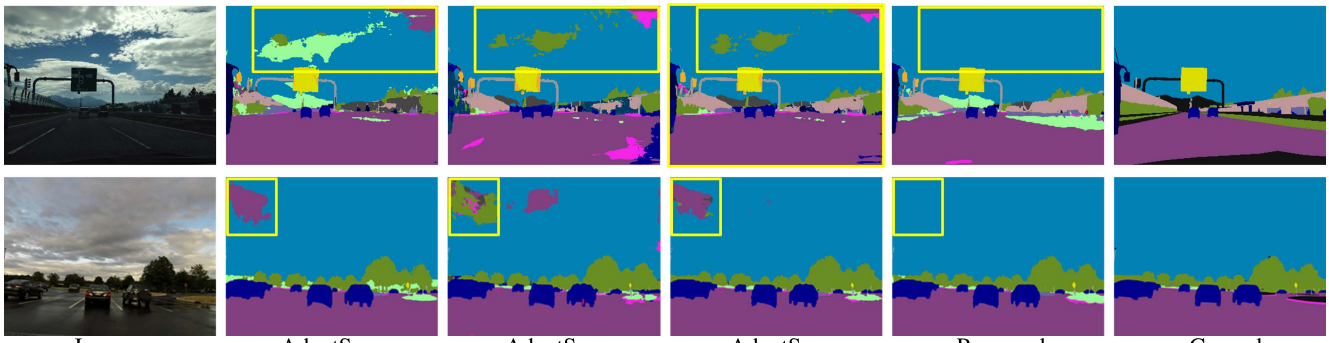

FIGURE 9. Visual analysis for Case 5. AdaptSeg-GTA5 and AdaptSeg-Synscapes made different predictions, and both predictions were incorrect. In this case, both models were updated by the incorrect pseudo labels. However, the models were not biased toward certain incorrect predictions and appeared to successfully revise their errors with the knowledge of the source domains.

about traffic lights and signs, which are small objects, was overwhelmingly high.

D. EMBEDDING SPACE ANALYSIS

We extracted the embedding space utilizing Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) [56] to analyze the effectiveness of the proposed method, as shown in Fig. 5. The 256-dimensional vector prior to the prediction layer of DeepLabv2 [2] was represented as the two-dimensional vector for embedding space visualization. For visual convenience, we selected four classes: road, sidewalk, building, and vegetation. All samples in the embedding space were selected from 500 images of Cityscapes validation set. When comparing building and vegetation classes (red and purple points in the figure), we can see that the proposed method provided a better cluster quality for two classes of embedding space. In the case of road and sidewalk, the proposed method separated the two classes more clearly than the existing ensemble-based self-trained model.

E. VISUAL ANALYSIS

In this section, we provided the visual analysis of the proposed method on the GTA5+ Synscapes → Mapillary setting with the results of AdaptSeg [9] which was utilized for the pre-training process of the proposed method. In this setting, the proposed method achieved 56.3% mIoU, as described in Table 5. In section III-B, we described the five cases in which two pseudo labels from two source domain-based models can appear. We set the source domains i and j to the GTA5 and Synscapes, respectively, and performed the analysis utilizing the predictions from the models for **Cases 2-5** in Figs. 6-9.

Case 1 is when both AdaptSeg-GTA5 and AdaptSeg-Synscapes make correct predictions, and the pseudo labels created in this case are not utilized for the self-training. We assume that the models are already obtaining sufficient information from the source domain for this case, and we focus the update of the model on the decoupled area. Even if the pseudo labels of **Case 1** are excluded from the self-training, there seems to be no decline in performance for the prediction, as shown in Figs. 6-9. This strategy can also prevent the training from being dominated by some dominant classes. A detailed description of each figure is described in the caption.

V. CONCLUSION

In this paper, we proposed a novel self-training-based MUDA to effectively improve semantic segmentation accuracy in the target domain using pseudo label rectification with the optimal training methodology. Specifically, the proposed pseudo label-oriented coteaching induced complementary self-training by leveraging the effect of peer-review without any pseudo label integration process. In addition, the proposed pseudo label decoupling alleviated the problem caused by a class imbalance in the semantic segmentation task and reduced the negative influence of self-training caused by incorrect pseudo labels. We proved the effectiveness of the proposed method through extensive experiments with various settings and analyzed the effects of the proposed pseudo label-oriented coteaching and pseudo label decoupling.

REFERENCES

[1] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," 2014, *arXiv:1412.7062*.

[2] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2017.

[3] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[4] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.

[5] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[6] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*.

[7] J. He, X. Jia, S. Chen, and J. Liu, "Multi-source domain adaptation with collaborative learning for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11008–11017.

[8] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell, "CyCADA: Cycle-consistent adversarial domain adaptation," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1989–1998.
[9] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. 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.
[10] Y.-H. Tsai, K. Sohn, S. Schulter, and M. Chandraker, "Domain adaptation for structured output via discriminative patch representations," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1456–1465.
[11] 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. (CVPR)*, Jun. 2019, pp. 2517–2526.
[12] H. Wang, T. Shen, W. Zhang, L.-Y. Duan, and T. Mei, "Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 642–659.
[13] P. Zhang, B. Zhang, T. Zhang, D. Chen, Y. Wang, and F. Wen, "Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12414–12424.
[14] H. Zhao, S. Zhang, G. Wu, J. M. Moura, J. P. Costeira, and G. J. Gordon, "Adversarial multiple source domain adaptation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–12.
[15] S. Zhao, B. Li, X. Yue, Y. Gu, P. Xu, R. Hu, H. Chai, and K. Keutzer, "Multi-source domain adaptation for semantic segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 7287–7300.
[16] Y. Zou, Z. Yu, B. 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.
[17] H. Wang, W. Yang, Z. Lin, and Y. Yu, "TMDA: Task-specific multi-source domain adaptation via clustering embedded adversarial training," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2019, pp. 1372–1377.
[18] W. Zhang, W. Ouyang, W. Li, and D. Xu, "Collaborative and adversarial network for unsupervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3801–3809.
[19] Y. Pan, Y. Li, Q. Cai, Y. Chen, and T. Yao, "Multi-source domain adaptation and semi-supervised domain adaptation with focus on visual domain adaptation challenge 2019," 2019, *arXiv:1910.03548*.
[20] S. Qiu, C. Zhu, and W. Zhou, "Meta self-learning for multi-source domain adaptation: A benchmark," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1592–1601.
[21] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang, "Domain adaptive ensemble learning," *IEEE Trans. Image Process.*, vol. 30, pp. 8008–8018, 2021.
[22] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Proc. Workshop challenges Represent. Learn. (ICML)*, 2013, vol. 3, no. 2, p. 896.
[23] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, "Unsupervised pixel-level domain adaptation with generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3722–3731.
[24] A. Mathur, A. Isopoussu, F. Kawsar, N. B. Berthouze, and N. D. Lane, "FlexAdapt: Flexible cycle-consistent adversarial domain adaptation," in *Proc. 18th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2019, pp. 896–901.
[25] 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.
[26] Y. Zhang, Z. Qiu, T. Yao, D. Liu, and T. Mei, "Fully convolutional adaptation networks for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6810–6818.
[27] L. Du, J. Tan, H. Yang, J. Feng, X. Xue, Q. Zheng, X. Ye, and X. Zhang, "SSF-DAN: Separated semantic feature based domain adaptation network for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 982–991.
[28] C.-H. Chao, B.-W. Cheng, C. Feng, and C.-Y. Lee, "Semantic segmentation based unsupervised domain adaptation via pseudo-label fusion," in *Proc. ICLR*, 2021. [Online]. Available: https://openreview.net/forum?id=tmMmIimNnp
[29] S. Paul, Y.-H. Tsai, S. Schulter, A. K. Roy-Chowdhury, and M. Chandraker, "Domain adaptive semantic segmentation using weak labels," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 571–587.
[30] 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.* Cham, Switzerland: Springer, 2020, pp. 532–548.
[31] K. Zhang, Y. Sun, R. Wang, H. Li, and X. Hu, "Multiple fusion adaptation: A strong framework for unsupervised semantic segmentation adaptation," 2021, *arXiv:2112.00295*.
[32] Q. Zhang, J. Zhang, W. Liu, and D. Tao, "Category anchor-guided unsupervised domain adaptation for semantic segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 433–443.
[33] Y. Wang, J. Peng, and Z. Zhang, "Uncertainty-aware pseudo label refinery for domain adaptive semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9092–9101.
[34] Z. Zheng and Y. Yang, "Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation," *Int. J. Comput. Vis.*, vol. 129, no. 4, pp. 1106–1120, 2021.
[35] M. N. Subhani and M. Ali, "Learning from scale-invariant examples for domain adaptation in semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 290–306.
[36] Q. Lian, L. Duan, F. Lv, and B. Gong, "Constructing self-motivated pyramid curriculums for cross-domain semantic segmentation: A non-adversarial approach," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6758–6767.
[37] Y. Zou, Z. Yu, X. Liu, B. V. K. V. Kumar, and J. Wang, "Confidence regularized self-training," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5982–5991.
[38] J. Iqbal and M. Ali, "MLSL: Multi-level self-supervised learning for domain adaptation with spatially independent and semantically consistent labeling," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 1864–1873.
[39] 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. (ICCV)*, Oct. 2017, pp. 2223–2232.
[40] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–11.
[41] E. Malach and S. Shalev-Shwartz, "Decoupling 'when to update' from 'how to update,'" in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
[42] K. Mei, C. Zhu, J. Zou, and S. Zhang, "Instance adaptive self-training for unsupervised domain adaptation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 415–430.
[43] B. H. Ngo, J. H. Kim, Y. J. Chae, and S. I. Cho, "Multi-view collaborative learning for semi-supervised domain adaptation," *IEEE Access*, vol. 9, pp. 166488–166501, 2021.
[44] B. H. Ngo, J. H. Park, S. J. Park, and S. I. Cho, "Semi-supervised domain adaptation using explicit class-wise matching for domain-invariant and class-discriminative feature learning," *IEEE Access*, vol. 9, pp. 128467–128480, 2021.
[45] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," 2016, *arXiv:1610.02242*.
[46] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "FixMatch: Simplifying semi-supervised learning with consistency and confidence," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 596–608.
[47] Y. Li, L. Yuan, and N. Vasconcelos, "Bidirectional learning for domain adaptation of semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6936–6945.
[48] J. M. Joyce, "Kullback–Leibler divergence," in *International Encyclopedia of Statistical Science*. Berlin, Germany: Springer, 2011, pp. 720–722.
[49] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for data: Ground truth from computer games," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 102–118.
[50] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3234–3243.
[51] M. Wrenninge and J. Unger, "Synscapes: A photorealistic synthetic dataset for street scene parsing," 2018, *arXiv:1810.08705*.
[52] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.
[53] G. Neuhold, T. Ollmann, S. R. Bulo, and P. Kontschieder, "The mapillary vistas dataset for semantic understanding of street scenes," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4990–4999.

[54] 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.
[55] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Statist.*, vol. 22, no. 3, pp. 400–407, 1951.
[56] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," 2018, *arXiv:1802.03426*.

SO JEONG PARK received the B.S. degree in multimedia engineering from Dongguk University, in 2021, where she is currently pursuing the M.S. degree. Her current research interests include semantic segmentation and domain adaptation.

HAE JU PARK was born in Daejeon, Republic of Korea, in 1999. She received the B.S. degree in multimedia engineering from Dongguk University, in 2022, where she is currently pursuing the M.S. degree. Her current research interests include image processing, 3D data processing, computer vision, and deep learning.

EUN SU KANG was born in Seoul, Republic of Korea, in 1996. He received the B.S. degree in multimedia engineering from Dongguk University, in 2021, where he is currently pursuing the M.S. degree. His research interests include image processing, computer vision, and deep learning networks.

BA HUNG NGO (Graduate Student Member, IEEE) received the B.S. degree in control engineering and automation from the Hanoi University of Mining and Geology, Hanoi, Vietnam, in 2014, and the M.S. degree in control engineering and automation from the Hanoi University of Science and Technology, in 2016. He is currently pursuing the Ph.D. degree with Dongguk University, Seoul, Republic of Korea. His current research interests include computer vision and deep learning, especially deep transfer learning, domain adaptation, and deep learning in medical imaging.

HO SUB LEE (Graduate Student Member, IEEE) received the B.S. degree in electrical and electronic engineering from Kyungpook National University, Republic of Korea, in 2014, and the M.S. and Ph.D. degrees in electrical and electronic engineering from the Pohang University of Science and Technology, in 2016 and 2020, respectively. He is currently an Assistant Professor of electronic engineering at Daegu University, Republic of Korea. His current research interests include image analysis, computer vision, and circuit design for display and multimedia systems.

SUNG IN CHO (Member, IEEE) received the B.S. degree in electronic engineering from Sogang University, Seoul, Republic of Korea, in 2010, and the Ph.D. degree in electrical and computer engineering from the Pohang University of Science and Technology, Pohang, Republic of Korea, in 2015. From 2015 to 2017, he was a Senior Researcher with LG Display. From 2017 to 2019, he was an Assistant Professor of electronic engineering at Daegu University, Gyeongsan, Republic of Korea. He is currently an Assistant Professor of multimedia engineering at Dongguk University, Seoul. His current research interests include image analysis and enhancement, video processing, computer vision, and deep learning.