Multi-Target Domain Adaptation with Class-Wise Attribute Transfer in Semantic Segmentation

Changjae Kim[†1]
kimcj5434@gmail.com
Seunghun Lee[2]
lsh5688@dgist.ac.kr
Sunghoon Im[*2]
sunghoonim@dgist.ac.kr

[1] LG Electronics, Korea
[2] DGIST, Korea

Abstract

In this paper, we present a novel multi-target domain adaptation (MTDA) method that adapts a single model to multiple domains with class-wise attribute transfer. To achieve this, we propose a high-precision pseudo labeling method for target domain images by utilizing cross-domain correspondence matching, which matches a target region to the most similar source region. Then, we propose class-wise image translation using the pseudo labels to avoid the problem of transferring characteristics between different classes and to allow translation between the same classes. Lastly, we introduce cross-domain feature consistency to learn the different characteristics of each target domain. Extensive experiments on the various complex driving scene show that ours achieves better performance than other state-of-the-art methods. The dense ablation study demonstrates the effectiveness of the proposed method.

1 Introduction

Unsupervised domain adaptation (UDA) addresses the domain shift problem caused by the distribution gap between training data and test data. It results in significant performance degradation in test domain inference. Existing works [1, 2, 3, 10, 26, 27] alleviate this problem using pixel and feature level distribution alignment method and self-training method in visual perception tasks. However, the previous works

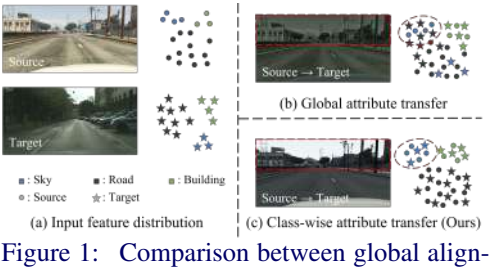

Figure 1: Comparison between global alignment and class-wise alignment.

† Work done in DGIST.
* Corresponding author.

have assumed a single target domain scenario which is incompatible with real-world data having multiple domains, such as the various object appearance, background, and illumination conditions. These single-target domain adaptations (STDA) methods have limitation in that it requires multiple domain-specific models to address multiple target domains effectively. To overcome the limitation, the multi-target domain adaptation (MTDA) methods [7, 13, 30, 32, 35, 41] have recently been proposed for visual perception tasks that adapt a single model from a source domain to multiple target domains.

Some MTDA works [13, 19] adopt image translation with domain alignment methods to deal with multiple target domains and show impressive performance in MTDA tasks. They show that global domain alignment via image translation takes an important role in adapting a model to multiple target distributions. However, [13, 19] transfer global styles of the target domain without considering the unique characteristics of each class in the image translation process. As a result, attributes of all target classes are mixed and transferred to each source class losing the distinctiveness. To this end, we propose class-wise multi-target domain adaptation that clearly distinguishes the attribute features of the class and purely translates the image styles of each class as shown in Fig. 1.

The class-wise image translation transfers the target domain attribute of each class to a source image to align the pixel-level distribution. The key point of class-wise translation is to obtain the label of an unlabeled target image to prevent attribute mixing between different classes. To this end, we propose a high-precision pseudo labeling (HPP) method which generates a pseudo label of the target image using the knowledge of the source domain and the similarity between a source image and the target image. The overall pipeline of image translation and pseudo labeling methods are illustrated in Fig. 2. Additionally, we propose a cross-domain feature consistency, which is a simple yet effective method for target domain distribution alignment without complex adversarial learning with domain discriminators[35]. It enables the model to extract domain-invariant features by creating the features of translated images that have been translated from the same source image to different domains. The proposed method has advantages over previous works on the granularity of distribution matching and the simplicity of training. Extensive experiments demonstrate the effectiveness of the proposed method and our approach achieves state-of-the-art performance.

2 Related Work

2.1 Image Translation

Image translation has received growing attention and remarkably has advanced with the emergence of conditional GANs [14]. Their method is improved by the following studies on adversarial learning [10, 15, 22, 36, 44] and style transfer methods [5, 6, 11, 28, 38]. Some works [12, 15, 17, 18, 23, 31] show that an image can be translated into various styles with disentangled features such as content and style. Based on these studies, [15, 16, 45] propose the method that transfer the style of each class respectively using ground truth information of the target image. However, class annotation is unavailable in unsupervised settings and the annotation cost is expensive, especially for segmentation labels. On the contrary, we propose the class-wise image translation without the ground truth of target data by estimating the pseudo label map of target images.

2.2 Unsupervised Domain Adaptation

Unsupervised domain adaptation aims to train a model on an unlabeled target domain using the information from a labeled source domain. The distribution alignment between source and target domains is the representative approach of unsupervised domain adaptation. [1, 18, 21, 27] align the pixel intensity distribution adopting style transfer and image translation methods. Some works [9, 24, 25] reduce the domain gap by adversarial training in the intermediate feature space of task networks, while others [37, 40] impose an adversarial loss on output space which has less domain gap compared to an image or feature space. Recent studies [20, 26, 42, 43] improve the generalization ability of the task model and use the pseudo labeling technique to adapt the target domain. To enhance the confidence of the pseudo label, target prediction confidence [20, 43], and distance from the centroid features of each class [26, 42] are used as criteria to filter uncertain regions.

2.3 Multi-Target Domain Adaptation

Multi-target domain adaptation is recently proposed that adapts multiple target domains from a source domain. To do so, common model parameter dictionary [41], knowledge distillation [30], and domain-invariant feature extraction methods [32] are proposed. Recent studies propose methods to tackle more complex tasks such as semantic segmentation in a driving situation. [13, 35] adopt knowledge distillation from a domain-specific teacher model to a domain-agnostic student model. Moreover, [35] use many discriminators to align source and target domains, and [13] enforce parameter consistency between teacher models and student model to create a domain-agnostic student model. [19] adapt a model to multiple target domains without domain-specific teachers using an image translation network that generates images with the properties of each target domain. In this work, we adopt the framework of [19] as a baseline and propose an improved image translation method and simple target domain alignment method.

3 Methods

The goal of the MTDA task is to learn a model that works well on K target domains $\mathcal{T}_{k=\{1,\ldots,K\}}$ simultaneously, by utilizing data from the source domain. We adopt ADAS [19] as our baseline. We newly propose the class-wise image translation method which transfers the attribute feature of each class. Our class-wise image translation considers the differences in attributes between classes to improve the target appearance. In addition, we propose a simple yet effective domain alignment method by introducing feature consistency between images converted to each target domain. We describe details of the class-wise image translation and cross-domain feature consistency in the following sections.

3.1 Overall networks

The overall networks for class-aware MTDT are illustrated in Fig. 2-(a). The image translation networks $\theta = \{f^{enc}, f^{dec}, g^{enc}, g^{dec}\}$ are composed of a semantic encoder, semantic decoder, attribute encoder, $f^{enc}, f^{dec}, g^{enc}$ and image decoder g^{dec}. The semantic encoder extracts the semantic features $F^S_{\mathcal{S}}$, $F^S_{\mathcal{T}_k}$ from the source and target images $I_{\mathcal{S}}$, $I_{\mathcal{T}k}$, respectively.

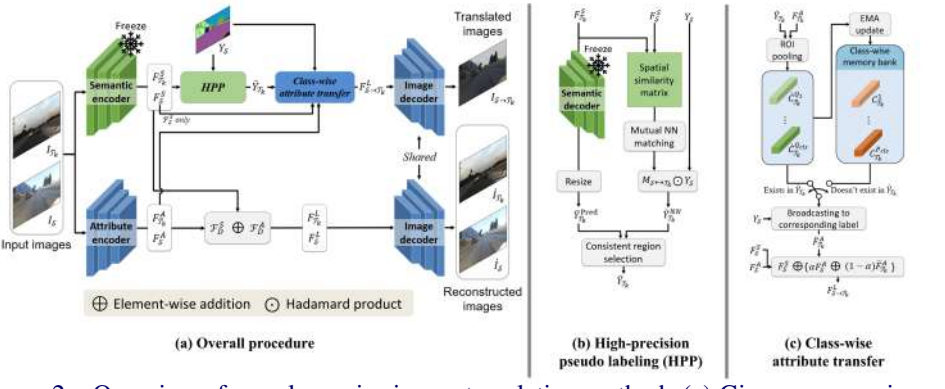

Figure 2: Overview of our class-wise image translation method. (a) Given a source image $I_{\mathcal{S}}$, label map $Y_{\mathcal{S}}$ and target image $I_{\mathcal{T}_k}$, our network produces an attribute-transferred image $I_{\mathcal{S}\to\mathcal{T}_k}$. Latent features $F^L_{\mathcal{S}}$, $F^L_{\mathcal{T}_k}$ are extracted by adding semantic features and attribute features from different encoders. Reconstructed images $\dot{I}_{\mathcal{S}}, \dot{I}_{\mathcal{T}k}$ are only generated during training time. (b) We obtain a high-precision pseudo label $\hat{Y}_{\mathcal{T}_k}$ using two intermediate pseudo labels $\hat{Y}^{Pred}_{\mathcal{T}_k}$ from the semantic decoder and $\hat{Y}^{NN}_{\mathcal{T}_k}$ from correspondence between the source image and the target image. (c) The representative target attributes $C_{\mathcal{T}_k}$, $\dot{C}_{\mathcal{T}_k}$ are broadcast to source label map $Y_{\mathcal{S}}$ to generate target attribute feature $\bar{F}^A_{\mathcal{T}_k}$ to be added to source attribute feature.

The semantic decoder produces the segmentation maps as follows:

$$\hat{Y}^{Pred}_x = f^{dec}(F^S_x),\ F^S_x = f^{enc}(I_x),\ x \in \{\mathcal{S}, \mathcal{T}_k\}, \tag{1}$$

where $\mathcal{S}$, $\mathcal{T}_k$ are the source and target domains, repsectively. We froze the parameter of the semantic encoder and decoder pretrained on the source datasets.

We train the translation network with two processes: class-wise attribute translation and reconstruction for feature space formation. First, reconstruction is a process of learning that ensures the image generated through the encoder and decoder is close to the original input image. This process constructs a feature space capable of generating an image in each domain. To continuously extract semantic information, the parameters of the semantic encoder are fixed. We use these semantic embeddings for pseudo labeling described in Sec. 3.2. The non-semantic attribute features $F^A_{\mathcal{S}}, F^A_{\mathcal{T}_k}$ are encoded by passing the images into the attribute encoder as follows:

$$F^A_x = g^{enc}(I_x), x \in \{\mathcal{S}, \mathcal{T}_k\}. \tag{2}$$

Then, the latent space features F^L_x are constructed by adding semantic features and attribute features, which is the input of image decoder g^{dec} to reconstruct the image $\dot{I}_x$ as follows:

$$\dot{I}_x = g^{dec}(F^L_x),\ F^L_x = F^S_x + F^A_x. \tag{3}$$

3.2 High-Precision Pseudo Labeling

The proposed HPP generates a pseudo label $\hat{Y}_{\mathcal{T}_k}$ of a target image, given the semantic features of source and target images $F^S_{\mathcal{S}}, F^S_{\mathcal{T}_k}$ as illustrated in Fig. 2-(b). We extract two intermediate pseudo labels to increase the precision of pseudo labels We pass the target features $F^S_{\mathcal{T}_k}$ through the segmentation decoder trained on source data as follows:

$$\hat{Y}^{Pred}_{\mathcal{T}_k} = f^{dec}(F^S_{\mathcal{T}_k}). \tag{4}$$

We froze the parameters of the semantic decoder to keep the learned semantic information during HPP training. Then, we extract another pseudo label $\hat{Y}^{NN}_{\mathcal{T}_k}$ from semantic correspondence between the source image and the target image. We compute the matching indicator matrix M using mutual nearest-neighbor matching as follows:

$$M_{(i,j)} = \begin{cases} 1, \text{ if } \phi(M'_{(i,*)}) = j,\ \phi(M'_{(*,j)}) = i \\ 0, \text{ else} \end{cases}, \tag{5}$$

where M' is a spatial similarity matrix of semantic features computed by a cosine function and ϕ is the *argmax* function which outputs the index of maximum value. We obtain the second pseudo label $\hat{Y}^{NN}_{\mathcal{T}_k}$ by using the Hadamard product of M and $Y_{\mathcal{S}}$ as follows:

$$\hat{Y}^{NN}_{\mathcal{T}_k} = M \odot Y_{\mathcal{S}}. \tag{6}$$

Lastly, we extract the high-precision pseudo labels $\hat{Y}_{\mathcal{T}_k}$ where where $\hat{Y}^{Pred}_{\mathcal{T}_k}$ and $\hat{Y}^{NN}_{\mathcal{T}_k}$ are consistent as follows:

$$\hat{Y}_{\mathcal{T}_k(i,j)} = 1(\hat{Y}^{Pred}_{\mathcal{T}_k(i,j)} = \hat{Y}^{NN}_{\mathcal{T}_k(i,j)})\hat{Y}^{Pred}_{\mathcal{T}_k(i,j)}. \tag{7}$$

The generated pseudo label is used to obtain the attribute features of the target classes in the following subsection.

3.3 Class-Wise Attribute Transfer

We transfer the target attribute feature to the area corresponding to the source class using the source label map. To get the attribute features of each target class $\dot{C}_{\mathcal{T}_k}$ from the input target image, we apply RoI pooling using a high-precision pseudo label. These attribute features are updated in corresponding memory attribute features $C_{\mathcal{T}_k}$ in the class-wise memory bank through the exponential moving average. The input attribute features $\dot{C}_{\mathcal{T}_k}$ are broadcast to the corresponding region in the source label to make a target attribute feature map $\bar{F}^A_{\mathcal{T}_k}$ which has the same layout as the source image. If there are source classes that are not in the current target image, memory attribute features $C_{\mathcal{T}_k}$ are used instead of input attribute features.

The broadcast target attribute feature map and the source attribute feature map are aggregated by interpolation with ratio α. We then add semantic embeddings to the latent image features, just like in the reconstruction process, by adding the semantic features of the original image to the aggregated attribute features as follows:

$$F^L_{\mathcal{S}\to\mathcal{T}_k} = F^S_{\mathcal{S}} + \{\alpha F^A_{\mathcal{S}} + (1-\alpha)\bar{F}^A_{\mathcal{T}_k}\}. \tag{8}$$

Finally, we generate attribute-translated images through the image decoder given the generated latent features of the target attribute as follows:

$$I_{\mathcal{S}\to\mathcal{T}_k} = g^{dec}(F^L_{\mathcal{S}\to\mathcal{T}_k}). \tag{9}$$

3.4 Training Loss for Translation Network

We train the image translation network θ and multi-head discriminator D by minimizing the total loss L_{trans} as follows:

$$L_{trans} = \lambda_{rec}L_{rec} + \lambda_{adv}L_{adv} + \lambda_{dom}(L^D_{dom} + L^\theta_{dom}), \tag{10}$$

where $L_{rec}, L_{adv}, L^{D}_{dom}$ and L^{θ}_{dom} are the reconstruction, adversarial, and domain discrimination losses, respectively. The weight terms λ balance the losses. The reconstruction loss is the L1 distance between the original input images and the reconstructed images $\dot{I}_{\mathcal{S}}, \dot{I}_{\mathcal{T}_k}$ generated by an image decoder g^{dec}. The adversarial learning between the image translation network and multi-head discriminator [19] to generate target domain images. The vanilla GAN loss [8] is adopted as the adversarial loss. We train the model to generate images with different characteristics for each domain by imposing domain classification loss. The domain classification layers and image translation networks are trained with the domain classification losses L^{D}_{dom} and L^{θ}_{dom}. We use the typical cross-entropy loss for the domain classification losses. The mathematical details of each loss term are described in the supplementary material.

3.5 Training Loss for Semantic Network with Cross-Domain Feature Consistency

The loss for the semantic network consists of the cross-entropy loss on the predicted segmentation maps and the proposed cross-domain feature consistency as follows:

$$L_{seg} = \lambda_{CE} L_{CE} + \lambda_{con} L_{con}, \tag{11}$$

where λ is weight balancing terms for the losses. We impose the typical cross-entropy loss L_{CE} to learn the translated image and target domain image using the ground truth labels and pseudo labels filtered by BARS [19].

The cross-domain feature consistency loss L_{con} is designed for target domain alignment by imposing constraints on the features from differently translated images to be consistent. Because the converted images have the same layout with different characteristics for each domain, it is possible to learn a model that extracts domain-invariant features by imposing element-wise consistency. We impose the feature consistency loss L_{con} with N set of unordered target domain pairs as follows:

$$L_{con} = \frac{1}{WHN} \sum_{\{\mathcal{T}_n, \mathcal{T}_m\}} L_2(F_{\mathcal{T}_n}, F_{\mathcal{T}_m}), \text{ where } \{\mathcal{T}_n, \mathcal{T}_m \mid n,\ m = 1, 2 \ldots K,\ n \neq m\}, \tag{12}$$

where W, H, and N are the width, the height of the feature map, and the number of pairs in Z, respectively. Additional training details can be found in our supplementary materials.

4 Experiments

In this section, we describe the experimental results of the proposed method. We evaluate our method on a semantic segmentation task in both the synthetic-to-real adaptation in Sec. 4.2 and the real-to-real adaptation in Sec. 4.3 with multiple driving scene datasets. We also conduct ablation studies to demonstrate the effectiveness of the proposed method in Sec. 4.4.

4.1 Datasets

We use GTA5 [34] as the source domain, along with multiple real-world datasets, Cityscapes [4], Indian Driving (IDD) [39], and Mapillary [29] as the target domains for the MTDA experiments. We train our model with labeled source data and unlabeled target data from

	Method	mIoU C	mIoU I	mIoU M	mIoU Avg.
G→ C, I	ADVENT	70.0	64.8	-	67.4
	MTKT	70.4	65.9	-	68.2
	ADAS	**75.4**	66.9	-	71.2
	Ours	74.4	**69.2**	-	**71.8**
G→ C, M	ADVENT	69.1	-	68.7	68.9
	MTKT	71.1	-	70.8	70.9
	ADAS	**75.3**	-	72.6	73.9
	Ours	74.8	-	**73.8**	**74.3**
G→ C, I, M	ADVENT	69.8	65.6	68.0	67.8
	MTKT	70.4	65.9	71.1	69.1
	ADAS	**74.9**	66.7	72.2	71.3
	Ours	74.0	**70.3**	**74.3**	**72.9**

Table 1: Quantitative comparison between our method and state-of-the-art methods with 7 classes setting. **Bold** means the best score.

	Method	mIoU C	mIoU I	mIoU M	mIoU Avg.
G→ C, I	CCL	45.0	46.0	-	45.5
	ADAS	45.8	46.3	-	46.1
	Ours	**46.5**	**46.9**	-	**46.7**
G→ C, M	CCL	45.1	-	48.8	46.8
	ADAS	45.8	-	**49.2**	47.5
	Ours	**47.1**	-	48.9	**48.0**
G→ I, M	CCL	-	44.5	46.4	45.5
	ADAS	-	**46.1**	47.6	46.9
	Ours	-	45.7	**48.7**	**47.2**
G→ C, I, M	CCL	46.7	47.0	49.9	47.9
	ADAS	46.9	47.7	**51.1**	48.6
	Ours	**49.3**	**48.8**	50.2	**49.4**

Table 2: Quantitative comparison between our method and state-of-the-art methods with 19 classes setting.

multiple domains. We use mIoU (%) as an evaluation metric for all domain adaptation experiments.

GTA5 contains 24,966 synthetic images with a resolution of 1914×1052 pixels, collected from the video game GTA5.

Cityscapes is a real-world dataset with 5,000 street scenes taken from cities in Europe and labeled into 19 classes. We used 2,975 images for training and 500 validation images.

IDD is the complex driving scene dataset that captures the Indian roads with diverse objects. It contains a total of 10,003 images, with 6,993 images for training, 981 for validation, and 2,029 for testing.

Mapillary provides 25,000 images collected from all around the world and diverse cameras. It includes 18,000 images for training, 5,000 images for testing, and 2,000 images for validation.

4.2 Synthetic-to-Real Adaptation

We conduct experiments on synthetic-to-real adaptation using GTA5 as the source dataset and Cityscapes, IDD, and Mapillary as the target datasets. First, we show the qualitative results of image-to-image translation from ours and ADAS [19] in Fig. 3. The results show that the transferred images from ADAS lose the unique attribute of each class in the target image (clear sky to the dark sky in the second row) because ADAS is designed for global style transfer. In contrast, our class-wise transfer method preserves the unique characteristics of each object or class (see the red box of the target image).

We report the quantitative semantic segmentation results with 7 superclasses in Tab. 1 and 19 classes in Tab. 2, respectively. The results show that the proposed method consistently outperforms all the competitive methods including ADVENT [40], MTKT [65], CCL [13] and ADAS [19]. ADVENT, an STDA method, records a lower performance than other MTDA methods Because ADVENT assumes the target domain as a single domain, it is difficult to adapt diverse multiple domains. Since both MTKT and CCL globally align the target distributions, it can cause the alignment between different target classes. The results in Tab. 1 and Tab. 2 show that ours outperform competitive methods for both 7 class and 19 class settings. As shown in Fig. 3, ADAS mixes the attributes of classes in a target image because of the global style transfer. On the other hand, the proposed method transfers the attribute of each class clearly and alleviates the listed problems. We believe that this is the key to the proposed method outperforming the previous works.

Figure 3: Qualitative comparison of image translation results from ours and ADAS.

Figure 4: Image translation results of a real-to-real adaptation setting. Our method can synthesize high-quality images for various source domains.

4.3 Real-to-Real Adaptation

Method	C→I,M		I→C,M		M→C,I	
	7	19	7	19	7	19
CCL	-	**52.5**	-	48.3	-	56.3
MTKT	68.8	-	-	-	-	-
ADAS	72.7	50.5	78.7	50.0	78.5	56.4
Ours	**73.3**	51.6	**79.4**	**51.2**	**78.7**	**57.2**

Table 3: Quantitative comparison on real-to-real adaptation.

The real-to-real adaptation experiments show the scalability of our model. We use one of the real-world datasets as a source domain and the other datasets are used as target domains. Fig. 4 shows the translated images using our class-wise image translation method. These images contain distinctive characteristics of objects for each domain. Our method outperforms other MTDA methods most adaptation scenarios in both label mapping settings as shown in Tab. 3. These experiments show that our proposed method generates high fidelity images regardless of the source domain and demonstrates the scalability and reliability of the proposed method.

4.4 Ablation Study

4.4.1 Global translation vs Class-wise translation

Method	mIoU C	mIoU I	mIoU M	mIoU Avg.
Color Transfer	33.8	37.4	42.1	37.8
DRANet	37.3	39.3	43.2	39.9
MTDT-Net	41.4	40.6	44.1	42.0
Ours	**42.7**	**41.3**	**45.3**	**43.1**

Table 4: Comparison with competitive image translation methods.

We conduct the analysis on the effect of the class-wise translation. We compare the proposed class-wise translation method to three different global adaptation methods, Color transfer [33], DRANet [18] and MTDT-Net [19]. The proposed method shows better performance by a large margin than the other methods in Tab. 4. While the other methods transfer the global style of the target image, our class-wise method aligns the pixel-level distribution finely and granularly using class-wise attribute transfer. This is the key to outperforming all these methods because our method transfers the unique attribute of the target class. The translated image makes it easier for the segmentation network to learn the confident features of each class of the target domain.

4.4.2 Pseudo Labeling

	Cityscapes	IDD	Mapillary	Avg.
$\hat{Y}^{NN}_{\mathcal{T}_k}$	0.59	0.57	0.55	0.57
$\hat{Y}^{Pred}_{\mathcal{T}_k}$ + BARS	0.72	0.66	0.73	0.70
Ours	**0.85**	**0.85**	**0.88**	**0.86**

Table 5: Correctness of the pseudo label using pixel precision as the performance metric.

Method	mIoU C	mIoU I	mIoU M	mIoU Avg.
w/o L_{Con}	48.5	48.2	49.6	48.8
w/ L_{Con}	**49.3**	**48.8**	**50.2**	**49.4**

Table 6: Ablation study of cross-domain feature consistency.

We show the qualitative synthetic-to-real image translation results and the pseudo label of the target image used in the translation process in Fig. 5. The building in the blue box of Fig. 5-(b) is converted to include the characteristics of the building marked with red boxes in the same row. The pseudo label corresponding to the area of the red box shows the correct pseudo label for building and other classes. The exact pseudo label is effective for the converted object to contain a distinctive appearance of the target class.

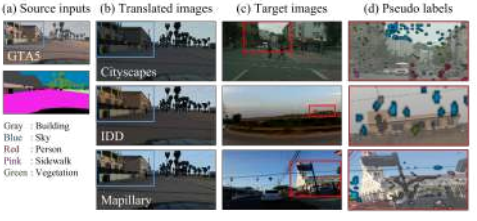

Figure 5: Qualitative results of image translation. (b) The building in the blue box of translated images contains the attribute of the building in the red box of target images. (d) The pseudo label in the red box region shows the generated pseudo label of the building.

We also report the quantitative result about the precision of the proposed pseudo labeling method with baseline methods in Tab. 5. To measure pixel precision, We set the pixel as the true positive if the pixel prediction is correct. Otherwise, the pixels are false positive. We compare the pseudo label computed by the nearest neighbor $\hat{Y}^{NN}_{\mathcal{T}_k}$, prediction $\hat{Y}^{Pred}_{\mathcal{T}_k}$ with BARS and our HPP method. The BARS is a filtering method proposed in [19] to remove the outliers where the prediction of a pixel and the class of the nearest centroid do not match. The NN-based pseudo label shows lower performance by a large margin than the other methods because it depends on the image pair that came into the input. The prediction method with BARS also shows limited performance because it usually filters boundary pixels between classes. On the other hand, our method generates a high-precision pseudo label exploiting the knowledge learned by the segmentation model and matching information between input image pairs. This experiment demonstrates that the proposed class-wise image translation method improves the precision of the pseudo labels.

4.4.3 Cross-Domain Feature Consistency

In this section, we conduct the ablation study on cross-domain feature consistency loss L_{con} to demonstrate the effectiveness of the proposed method. We train the segmentation network with translated images in a supervised manner and with target images in an unsupervised manner. We impose cross-entropy loss in (11) using ground truth source labels for translated images and filtered pseudo labels for target images. It improves the performance for all target domains with an average mIoU margin of 0.6%, as shown in Tab. 6.

5 Conclusion

In this work, we propose a novel class-wise image translation method and a simple yet effective domain alignment method. In the image translation procedure, we present high-precision pseudo labeling to prevent the mixing of attributes between classes, building a framework that allows attribute transfer only between the same classes between each domain. Since the proposed class-wise image translation method imitates the pixel-level distribution of the target domain better than the global image translation method. It not only produces visually pleasing image translation results but also achieves better domain adaptation performance. Extensive experiments demonstrate that the proposed cross-domain feature consistency imposed on the features from differently translated images adequately trains a domain-invariant model. In particular, the ablation study on cross-domain feature consistency provides reliable results. This work contributes to the recent research on domain adaptation toward more practical use cases. Future works may consider more complex and realistic scenarios which adapt to a continuous domain that changes over time, such as different seasons.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00210908).

References

[1] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3722–3731, 2017.

[2] Yuhua Chen, Wen Li, Xiaoran Chen, and Luc Van Gool. Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[3] Yun-Chun Chen, Yen-Yu Lin, Ming-Hsuan Yang, and Jia-Bin Huang. Crdoco: Pixel-level domain transfer with cross-domain consistency. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[4] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[5] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. *arXiv preprint arXiv:1610.07629*, 2016.

[6] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[7] Behnam Gholami, Pritish Sahu, Ognjen Rudovic, Konstantinos Bousmalis, and Vladimir Pavlovic. Unsupervised multi-target domain adaptation: An information theoretic approach. *IEEE Transactions on Image Processing (TIP)*, 29, 2020.

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

[9] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*, 2016.

[10] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International Conference on Machine Learning (ICML)*. PMLR, 2018.

[11] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2017.

[12] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2018.

[13] Takashi Isobe, Xu Jia, Shuaijun Chen, Jianzhong He, Yongjie Shi, Jianzhuang Liu, Huchuan Lu, and Shengjin Wang. Multi-target domain adaptation with collaborative consistency learning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[14] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[15] Somi Jeong, Youngjung Kim, Eungbean Lee, and Kwanghoon Sohn. Memory-guided unsupervised image-to-image translation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[16] Soohyun Kim, Jongbeom Baek, Jihye Park, Gyeongnyeon Kim, and Seungryong Kim. Instaformer: Instance-aware image-to-image translation with transformer. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[17] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2018.

[18] Seunghun Lee, Sunghyun Cho, and Sunghoon Im. Dranet: Disentangling representation and adaptation networks for unsupervised cross-domain adaptation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[19] Seunghun Lee, Wonhyeok Choi, Changjae Kim, Minwoo Choi, and Sunghoon Im. Adas: A direct adaptation strategy for multi-target domain adaptive semantic segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[20] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[21] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[22] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[23] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2019.

[24] Yawei Luo, Ping Liu, Tao Guan, Junqing Yu, and Yi Yang. Significance-aware information bottleneck for domain adaptive semantic segmentation. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2019.

[25] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[26] Haoyu Ma, Xiangru Lin, Zifeng Wu, and Yizhou Yu. Coarse-to-fine domain adaptive semantic segmentation with photometric alignment and category-center regularization. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[27] Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim. Image to image translation for domain adaptation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[28] Hyeonseob Nam and Hyo-Eun Kim. Batch-instance normalization for adaptively style-invariant neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[29] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2017.

[30] Le Thanh Nguyen-Meidine, Atif Belal, Madhu Kiran, Jose Dolz, Louis-Antoine Blais-Morin, and Eric Granger. Unsupervised multi-target domain adaptation through knowledge distillation. In *Proceedings of IEEE Winter Conference on Computer Vision (WACV)*, 2021.

[31] Taesung Park, Jun-Yan Zhu, Oliver Wang, Jingwan Lu, Eli Shechtman, Alexei A Efros, and Richard Zhang. Swapping autoencoder for deep image manipulation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[32] Xingchao Peng, Zijun Huang, Ximeng Sun, and Kate Saenko. Domain agnostic learning with disentangled representations. In *International Conference on Machine Learning (ICML)*. PMLR, 2019.

[33] Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer graphics and applications*, 21(5):34–41, 2001.

[34] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2016.

[35] Antoine Saporta, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Multi-target adversarial frameworks for domain adaptation in semantic segmentation. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2021.

[36] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. *International Conference on Learning Representations (ICLR)*, 2017.

[37] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[38] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

[39] Girish Varma, Anbumani Subramanian, Anoop Namboodiri, Manmohan Chandraker, and CV Jawahar. Idd: A dataset for exploring problems of autonomous navigation in unconstrained environments. In *Proceedings of IEEE Winter Conference on Computer Vision (WACV)*, 2019.

[40] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[41] Huanhuan Yu, Menglei Hu, and Songcan Chen. Multi-target unsupervised domain adaptation without exactly shared categories. *arXiv preprint arXiv:1809.00852*, 2018.

[42] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[43] Zhedong Zheng and Yi Yang. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *International Journal of Computer Vision (IJCV)*, 129(4):1106–1120, 2021.

[44] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2017.

[45] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Sean: Image synthesis with semantic region-adaptive normalization. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.