Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 1848857, 8 pages
https://doi.org/10.1155/2022/1848857

Research Article

Domain Adaptation through Photorealistic Enhanced Images for Semantic Segmentation

Takafumi Katayama,[1] **Tian Song**,[1] **Xiantao Jiang**,[2,3] **Jenq-Shiou Leu**,[4] **and Takashi Shimamoto**[1]

[1] *Graduate School of Technology, Industrial and Social Sciences, Tokushima University, Tokushima 770-8506, Japan*
[2] *Department of Information Engineering, Shanghai Maritime University, Shanghai 201306, China*
[3] *School of Computer Information Engineering, Nanchang Institute of Technology, Nanchang 330044, China*
[4] *Department of Electrical and Computer Engineering, National Taiwan University of Science and Technology, Taipei 10607, Taiwan*

Correspondence should be addressed to Takafumi Katayama; t.katayama@tokushima-u.ac.jp

Received 15 March 2022; Accepted 20 June 2022; Published 15 July 2022

Academic Editor: Adriana del Carmen Téllez-Anguiano

In this paper, three types of domain adaptation which are defined as image-level domain adaptation, interdomain adaptation, and intradomain adaptation are efficiently combined to construct a high efficiency framework for semantic segmentation. The proposed domain adaptation platform can achieve a high reduction of time-consuming to generate exhausted supervised data in the real world using photorealistic images. The proposed framework achieved a mean Intersection-over-Union (mIoU) of 45.0%. Furthermore, by combining the proposed method with intradomain adaptation, the improvement of 1.2% mIoU is achieved compared to previous work.

1. Introduction

Convolutional neural networks (CNNs) based approaches brought about recent development in computer vision. Semantic segmentation has attracted attention from CNN-based models with potential applications for autonomous driving technology, disease diagnosis, and image editing. Semantic segmentation is a fundamental technique that assigns class labels such as person, car, road, and tree to every pixel in an image. The segmented model needs to be trained by using a per-pixel ground truths image. However, the training process for semantic segmentation has two key issues. The first one is that accurate per-pixel annotations require long manual working hours and high costs. It is reported that the Cityscapes dataset (a dataset of driving images) needs 90 minutes per image to create per-pixel annotation [1]. The second one is that the accuracy of semantic segmentation is decreased when a domain gap between the training datasets and the test datasets is involved. For instance, the feature distribution of an image may significantly differ from that of the training images when the city, weather, or shooting conditions change. In such cases, an only supervised model cannot achieve high accurate semantic segmentation. Therefore, it is necessary to generate a trained model using the datasets optimized for various conditions.

Currently, to solve the time-consuming per-pixel annotation with all conditions, the pixel-level annotations to photorealistic images rendered from game engines are supplemented to datasets and used for the training of semantic segmentation. Consequently, the efficient domain transfer between photorealistic images and real world images is required. This means tackling problems with significantly different domain distributions. A process that can be learnt even when the domain gaps are significantly different has the potential to develop the field of learning, which is a challenge for data-driven artificial intelligence.

The different domain distributions in-game images and real driving sequences give less accurate segmentation. To solve the abovementioned issue, the technique of domain adaptation has been proposed to adjust the features across the target data and source data [2–6]. These works introduced cross-domain methodology and efficient applications on edge computing conditions. Luo et al. showed that directly aligning the high-level semantic features may lead to negative transfer and reduce the domain adaptation performance in the originally well-aligned regions [7]. To solve this issue, a local score alignment map to guide the transfer of semantic information is proposed.

In semantic segmentation, considering the interdomain gap between the game images and the real world images, the method of minimizing the entropy loss by adversarial methods has shown high accuracy [8]. Furthermore, based on minimizing the entropy loss model, a two-stage self-supervised domain adaptation approach, which minimizes large distribution gaps in the target sequence itself (intradomain gaps), has shown better performance than the previous model [9]. However, all of the previous models only consider adaptive learning in intermediate feature space and do not perform domain adaptation at the image level. Therefore, we proposed a domain adaptation framework including image-level domain adaptation.

The image-level domain adaptation has two important elements. The first is that the pixel alignment of the source domain image in the feature space is transferred to the target domain in the feature space, thus enabling the transfer of visual style. The second is that the output image is structurally matched to the input image without the need for prior per-pixel annotation. The structural match allows the ground truth to be used as it was before the transformation, thus reducing annotation time. The latest image transformation model for improving photorealism does not require annotation and is structurally consistent with input and output [10]. We also focus on the fact that various visual-style transformations, including appearance, shape, and context, enable domain adaptation at the image level with narrower domain gaps.

As our previous work, we introduced a new domain adaptation approach for semantic segmentation [11]. Based on the previous work, we focus on the accuracy improvement of the semantic segmentation performance in this paper. Because it is difficult to define a numeral photorealism of the photorealistic datasets for domain adaptation, in this work the typical photorealistic datasets which consist of urban street scenes are considered proper datasets for the evaluation of semantic segmentation.

Our approach achieved improved accuracy of semantic segmentation by using transformed photorealistic images. Our main contributions to this paper are as follows:

(i) We show the effectiveness of image-level domain adaptation on the accuracy of semantic segmentation. Moreover, we proposed a framework combining three-domain adaptation types to achieve accurate semantic segmentation.

(ii) We improve the accuracy of the semantic segmentation by a method without using real world supervised data. This suggests that the field may be able to reduce time-consuming annotation and adapt segmentation to various real world domains in the future.

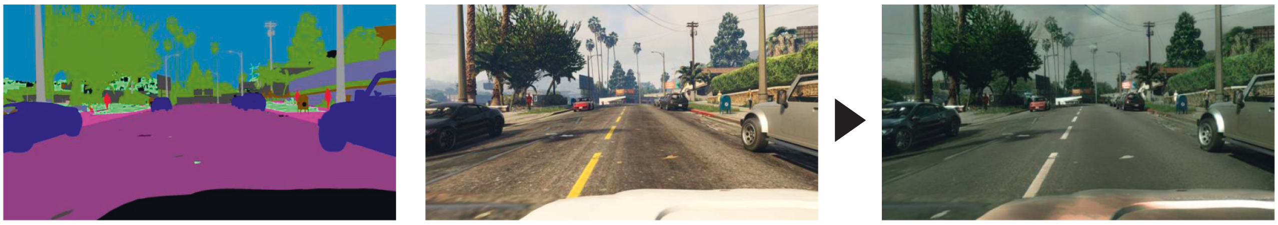
2. Related Work

Domain adaptation is considered an efficient approach to achieving a fast generation of annotation data. However, different domain adaptation algorithm makes use of different merits from different viewpoints. It can be concluded as image-level adaptation, interdomain adaptation, and intradomain adaptation. In this work, we try to find excellent adaptation algorithms from a different viewpoint and combine these algorithms into a framework to improve the adaptation performance.

In this section, three selected algorithms including image-level domain adaptation, interdomain adaptation, and intradomain adaptation will be reviewed. Firstly, a photorealism enhancement method for image-level domain adaptation, which is designed for game images, will be introduced [10]. Then, an interdomain adaptation method based on entropy minimization will be introduced [8, 12]. Finally, an intradomain adaptation method based on the ranked classification of images will be reviewed [9].

2.1. Image-Level Domain Adaptation. Image-level domain adaptation is the transfer of visual style by transferring the pixel alignment of the source domain image to the target domain in feature space. For example, CycleGAN achieves the visual transformation of a photograph into a Van Gogh painting by learning to minimize cycle-consistent loss [13]. Another method for image-level domain adaptation is to project a high-dimensional feature space onto a segmentation map, but the utilization of CycleGAN is limited because the transformable images are limited to datasets with per-pixel annotations. In addition, a method for improving the photorealism of game images has been proposed [10]. This model uses adversarial learning with strong supervision at multiple perceptual levels, which provides stability and significant photorealism improvement. The method for improving the photorealism of game images avoids the preparation of the pre-annotated labels by generating identical label maps for synthetic and real images. Figure 1 shows the results of the photorealistic enhancement generated by the model [10]. There is no change in appearance between synthetic image from GTAV (Figure 1(b)) [14] and photorealistic enhanced image generated from [10](Figure1(c)), and annotation data of ground truth can be applied to both images (Figure 1(a)). Therefore, it is confirmed from the results that it is not necessary to re-annotate the data.

2.2. Interdomain Adaptation. The main idea of unsupervised interdomain adaptation is to adjust the distributional misalignment. Domain adaptation approaches often tackle the problem by aligning the feature distribution between the source and target images [15–18]. Approaches include maximum mean discrepancies, self-learning, providing



(a) Annotation data

(b) Input image

(c) Photorealistic enhanced image

Figure 1: Ground truth and photorealistic enhanced images. (a) Annotation data. (b) Input image. (c) photorealistic enhanced image.

pseudo-labels, or adversarial learning, but here we describe a method that tackles interdomain adaptation by minimizing the distribution difference of intermediate features used in this work. Most of the approaches to minimize the distributional difference of intermediate features do not consider the feature space at the image level. This is because that domain adaptation is often plagued by the complexity of visual high-dimensional features and considers domain adaptation in the output space. The model, which proposed an efficient domain adaptation algorithm with adversarial learning in the output space, achieved improved accuracy in semantic segmentation using adversarial learning in the output space of the segmentation space [19]. The interdomain adaptation model, which applies unsupervised domain adaptation in the output space based on entropy, achieves higher accuracy improvement in semantic segmentation than the previous model. The proposed domain adaptation is applied to the entropy-based adversarial training approach targeting the entropy minimization objective and the structure adaptation from the source domain to the target domain [8, 12]. The entropy minimization method is one of the successful approaches used for semisupervised learning.

2.3. Intradomain Adaptation. In interdomain adaptation, some previous works focus on bridging the gap between domains. In contrast, model [9], which considers entropy-based intradomain adaptation, tackles intradomain adaptation by ranking the images in the target dataset and classifying them into easy or hard splits. Easy split means images with small domain gaps and easy to detect, while hard split means images with significant domain gaps and lower detection accuracy.

Intradomain adaptation is an adversarial learning based on entropy. The generator G_{inter} used for adversarial learning of intradomain adaptation takes the target image X_t as input and generates an entropy map I_t. The equation for ranking is defined as follows:

$$R\left(X_t\right) = \frac{1}{HW}\sum_{h,w} I_t^{h,w}, \tag{1}$$

where the average value of the entropy map I_t is calculated. After that, the target images are classified into easy or hard splits using the average value $R(|X_t|)$ and a simple image ratio λ as follows:

$$\lambda = \frac{\left|X_{te}\right|}{\left|X_t\right|}, \tag{2}$$

where $|X_t|$ represents the entire image and $R(|X_{te}|)$ is the set of images in the easy split. After calculating the average value of the entropy map $R(|X_t|)$, we can extract a group of images with a small domain gap from the target data by giving an arbitrary ratio λ. After the classification is done, the result of the entropy output for the images with few domain gaps is used as the supervised data, and the images with many domain gaps are used as unsupervised data to perform adversarial learning based on entropy to improve the accuracy of semantic segmentation.

3. Approach

In this paper, we focused on the domain adaptation of three types: image-level domain, interdomain, and intradomain to improve the accuracy of semantic segmentation. The implementation of each level of domain adaptation allows the utilization of transformed photorealistic images from GTAV and improves the accuracy of semantic segmentation in the real world, such as Cityscapes. Figure 2 shows an overview of the proposed framework. The proposed semantic segmentation algorithm uses image-level domain adaptation (Figure 2(a)), interdomain adaptation (Figure 2(b)), and intradomain adaptation (Figure 2(c)). Moreover, the proposed domain adaptation allows segmentation well on images without supervised data from the proposed architecture. Thereby, the proposed method reduces the time-consuming creation time of semantic labels. The details are described in the following subsections.

3.1. Image-Level Domain Adaptation. Image-level domain adaptation method for semantic segmentation is not proposed in previous work. Image-level domain adaptation suffers from diverse visual complexities, including illumination reflection, glossiness, and transparency. Our approach uses domain adaptation at the image level to improve semantic segmentation based on the method that greatly improves the realism of rendered game images [10]. This approach uses intermediate buffers produced by game images during the rendering process. These buffers provide detailed information on geometry, materials, and lighting in the scene. The previous work proposed the integration of these buffers into the photorealism enhancement flow.

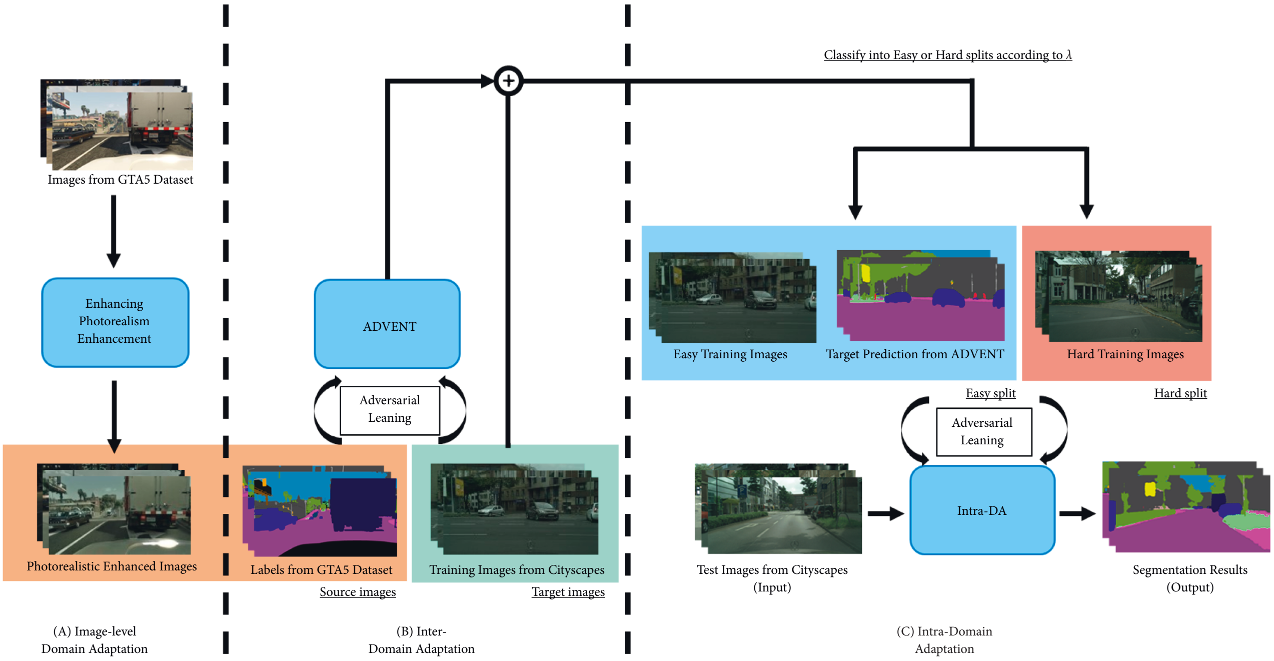

Images from GTA5 Dataset
Enhancing Photorealism Enhancement
Photorealistic Enhanced Images
(A) Image-level Domain Adaptation
ADVENT
Adversarial Leaning
Labels from GTA5 Dataset
Training Images from Cityscapes
Source images
Target images
(B) Inter-Domain Adaptation
Classify into Easy or Hard splits according to λ
Easy Training Images
Target Prediction from ADVENT
Hard Training Images
Easy split
Hard split
Adversarial Leaning
Test Images from Cityscapes (Input)
Intra-DA
Segmentation Results (Output)
(C) Intra-Domain Adaptation

Figure 2: Overview of the proposed framework for domain adaptation. The photorealistic enhanced images are generated by (a) image-level domain adaptation. In (b) interdomain adaptation, the adversarial learning represents that G_{inter} and are optimized by minimizing the segmentation loss L_{inter}^{seg} and the adversarial loss L_{inter}^{adv}. In (c) intradomain adaptation, the adversarial learning represents that G_{intra} and are optimized by using the intradomain segmentation loss L_{intra}^{seg} and the adversarial loss L_{intra}^{adv}.

Thereby, the model trained by real world datasets (Cityscapes, KITTI, and so on) can output the corresponding visual style. Moreover, since the output image is structurally consistent with the input image, this approach can be used for unsupervised domain adaptation. The following sections use images transformed into the visual style corresponding to Cityscapes by photorealism improvement. Figure 3 shows a sample frame for photorealistic enhancement. The GTAV dataset consists of temporally diverse frames that are well transformed.

3.2. Interdomain Adaptation. Interdomain adaptation aims to adjust the distributional misalignment between labeled source data and unlabeled target data. We use 19,252 images converted to photorealism and the corresponding ground truths as source images. In addition, 2,975 images from the Cityscapes dataset acquired from the real-world are used as target images.

We perform interdomain adaptation based on adversarial learning to minimize entropy loss by adversarial methods [8]. A sample X_s is defined as a source domain with its ground truth annotation Y_s. $[Y_s^{(h,w,c)}]_c$ of Y_s provides a label of a pixel (h, w) as a one-hot vector. Each C-dimensional vector $[P_s^{(h,w,c)}]_c$ at a pixel (h, w) serves as a discrete distribution over C classes which $P_s = G_{inter}(X_s)$ is defined as a segmentation map. The segmentation map is the output X_s and the interdomain generator G_{inter}. G_{inter} is optimized by minimizing the cross-entropy loss:

$$L_{\text{inter}}^{seg}(X_s, Y_s) = -\sum_{h,w}\sum_{c} Y_s^{(h,w,c)} \log\left(P_s^{(h,w,c)}\right). \qquad (3)$$

Additionally, the generator G_{inter} takes a target image X_t as an input and generates the segmentation map $P_t = G_{inter}(X_t)$. Then, the entropy map I_t is defined as follows:

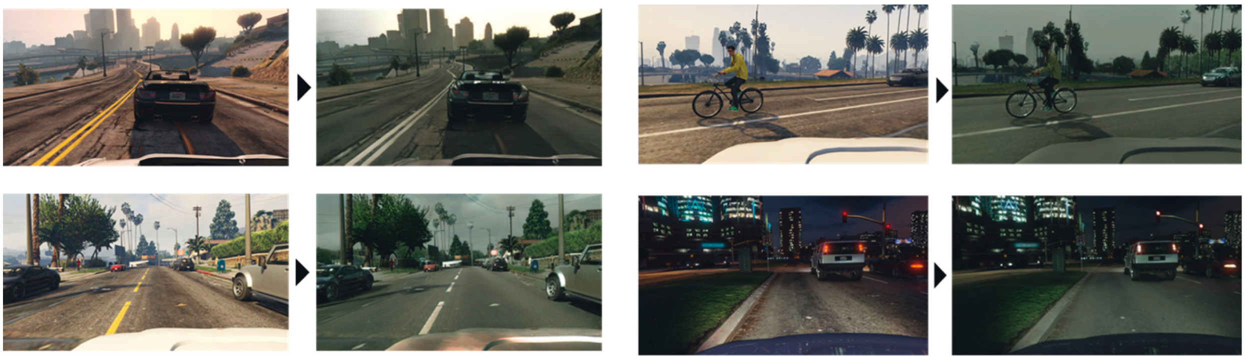
$$I_t^{(h,w)} = \sum_{c} -P_t^{(h,w,c)} \log\left(P_s^{(h,w,c)}\right). \qquad (4)$$

To align the interdomain gap, D_{inter} is trained to predict the domain labels for the entropy maps, while G_{inter} is trained to fool D_{inter}. The optimization of G_{inter} and D_{inter} achieved the following adversarial loss function:

$$\begin{aligned} L_{\text{inter}}^{a\,dv}(X_s, X_t) = -\sum_{h,w} & \log\left(1 - D_{\text{inter}}\left(I_t^{(h,w)}\right)\right) \\ & + \log\left(D_{\text{inter}}\left(I_s^{(h,w)}\right)\right), \end{aligned} \qquad (5)$$

where I_s is the entropy map of X_s. The loss functions $L_{inter}^{a\,dv}$ and L_{inter}^{seg} are optimized to align the distribution shift between the source and target data. Then, target domain and predicted entropy maps of target data are generated such that the target data can be clustered into an easy and hard split.

3.3. Intradomain Adaptation. Intradomain adaptation aims to reduce the large domain gaps in the target data. Compared to a clear image captured in a stationary state, some images in a sequence are degraded by noise. Such a situation is called the intradomain gap. Intradomain adaptation solves the problem of degraded semantic segmentation accuracy in intradomain gap sequences. To find images with intradomain gaps, we use an entropy-based ranking system (equation (1)) that classifies the target data into easy or hard



(a) Input images (b) Photorealistic enhanced images (a) Input images (b) Photorealistic enhanced images

Figure 3: Sample results of photorealistic enhanced images.

splits. The threshold for separating easy or hard images is set to 0.67, showing the best results in previous work [9].

When an image of the easy split is defined as X_{te}, the predicted segmentation map $P_{te} = G_{inter}(X_{te})$. G_{intra} is optimized by minimizing the cross-entropy loss as follows:

$$L_{\text{intra}}^{seg}(X_{te}) = -\sum_{h,w}\sum_{c} P_{te}^{(h,w,c)}\log\left(G_{\text{intra}}(X_{te})^{h,w,c}\right). \tag{6}$$

The alignment on the entropy map for both splits to bridge the intradomain gap between the easy and hard split is adopted. An image X_{th} from hard split is input to the generator G. Then, the segmentation map $P_{th} = G(X_{th})$ and the entropy map I_{th} are generated, where I_{te} is from the easy split and I_{th} is from the hard split. To close the intradomain gap, the intradomain discriminator D_{intra} is trained to predict the split labels of I_{te} and I_{th}. G is trained to fool D_{intra}. The adversarial learning loss to optimize G_{intra} and D_{intra} is calculated as follows:

$$\begin{aligned} L_{\text{intra}}^{a\,dv}(X_{te}, X_{th}) = &-\sum_{h,w}\log\left(1 - D_{\text{intra}}\left(I_{th}^{(h,w)}\right)\right) \\ &+ \log\left(D_{\text{intra}}\left(I_{te}^{(h,w)}\right)\right). \end{aligned} \tag{7}$$

Finally, all of loss function L is defined as follows:

$$L = L_{\text{inter}}^{seg} + L_{\text{inter}}^{a\,dv} + L_{\text{intra}}^{seg} + L_{\text{intra}}^{a\,dv}, \tag{8}$$

and the objective is to learn a target model G according to the following:

$$G = \arg\min_{G_{\text{intra}}}\min_{G_*}\max_{D_*} L, \tag{9}$$

where the asterisk denotes intra and inter. The domain adaptation model is two-step self-supervised approach. Firstly, G_{inter} and D_{inter} of the interdomain adaptation model are optimized. Secondly, by using a target image assigned to the easy and hard split with entropy-based ranking system, the intradomain adaptation is optimized.

4. Dataset and Evaluation Metrics

This work uses images and semantic labeling rendered from the popular game "Grand Theft Auto V," which is based on the urban landscape of Los Angeles [14]. The photorealistic datasets are commonly used for the evaluation of domain adaptation. When performing interdomain adaptation, 19,252 photorealistic enhanced GTAV images are used as training images (source images). In addition, 2,975 images from the Cityscapes dataset acquired from the real-world are used as training images (target images). We used the 500 images of the Cityscapes validation dataset to evaluate the semantic segmentation.

Semantic segmentation uses IoU as an evaluation metrics, which is commonly used in object detection challenges such as the PASCAL VOC challenge. IoU is calculated as Area of Overlap classified divided by Area of Union. The Area of Overlap is the area of overlap between the predicted area and the ground truth area, and the Area of Union is the area contained in both the predicted area and the ground truth area. By dividing the Area of Overlap by the Area of Union, we can obtain the mean Intersection-over-Union (mIoU (%)).

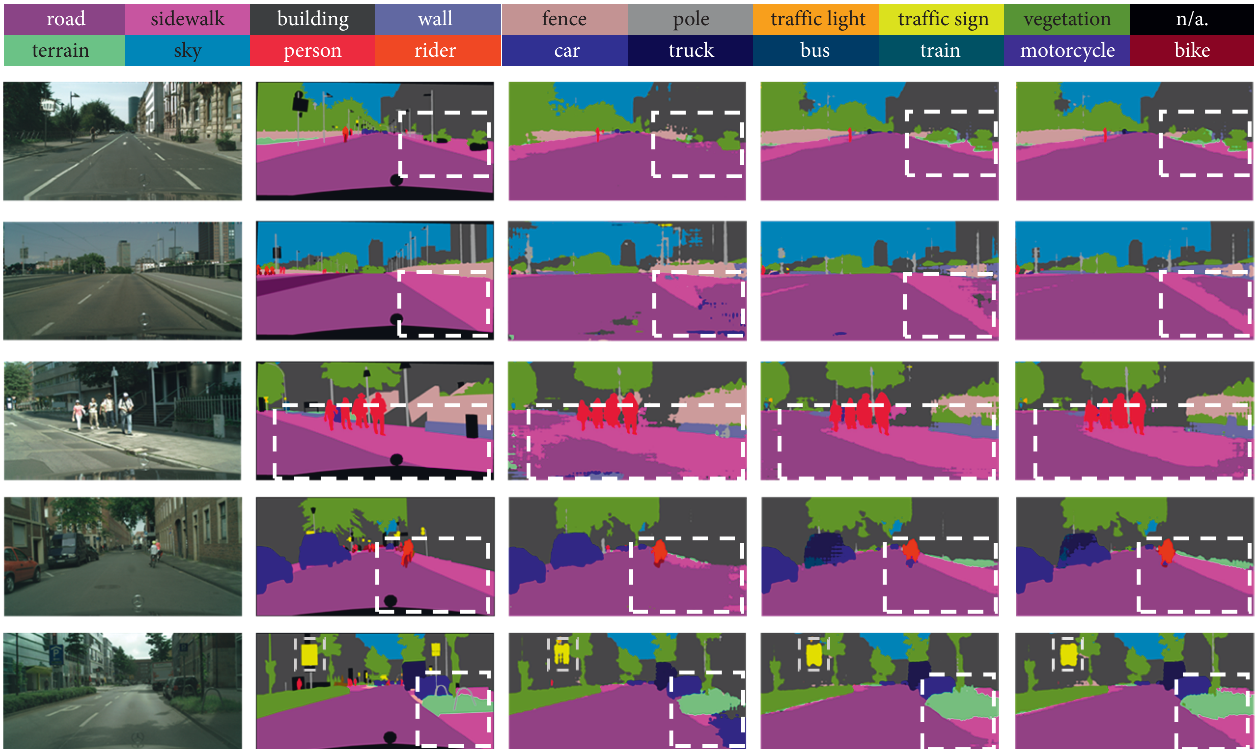
5. Simulation Results and Discussion

All the simulation results in this paper are implemented with Pytorch in a single NVIDIA TITAN RTX GPU. Building upon a good baseline model is essential to achieve high-quality segmentation results [20–22]. A typical evaluation method for semantic segmentation accuracy is used in this work which enables the comparison with various previous works. We adopt the DeepLab-v2 framework with ResNet-101 model pretrained on ImageNet as our segmentation baseline network [23, 24]. Interdomain adaptation and intradomain adaptation using the loss function of the entropy minimization is trained 120,000 times.

To evaluate the domain adaptation, we compared the results of training with GTAV and testing with Cityscapes. The adaptation results compared to various baselines are shown in Table 1. In Table 1, ours represents the result using image-level domain adaptation and interdomain adaptation, while Ours + Intra represents Ours plus intradomain adaptation. The proposed method achieved 45.0% mIoU using image-level domain adaptation and interdomain adaptation. Moreover, the proposed methods implemented with three

Table 1: Semantic segmentation results of adapting GTAV to Cityscapes.

Methods	Road	Sidewalk	Building	Wall	Fence	Pole	Light	Sign	Veg	Terrain
Baseline (ResNet)	75.8	16.8	77.2	12.5	21.0	25.5	30.1	20.1	81.3	24.6
AdvEnt [8]	89.9	36.5	81.6	29.2	25.2	28.5	32.3	22.4	83.9	34.0
AdaSegNet [19]	86.5	36.0	79.9	23.4	23.3	23.9	35.2	14.8	83.4	33.3
CLAN [7]	87.0	27.1	79.6	27.3	23.3	28.3	35.5	24.2	83.6	27.4
Ours	89.4	46.0	83.1	27.6	22.7	33.6	33.6	27.3	83.6	34.5
IntraDA [9]	90.6	36.1	82.6	29.5	21.3	27.6	31.4	23.1	85.2	39.3
Ours + Intra	91.9	49.0	84.2	29.2	24.7	33.0	34.0	34.9	84.6	39.4
Methods	Sky	Person	Rider	Car	Truck	Bus	Train	Mbike	Bike	mIoU
Baseline (ResNet)	70.3	53.8	26.4	49.9	17.2	25.9	6.5	25.3	36.0	36.6
AdvEnt [8]	77.1	57.4	27.9	83.7	29.4	39.1	1.5	28.4	23.3	43.8
AdaSegNet [19]	75.6	58.5	27.6	73.7	32.5	35.4	3.9	30.1	28.1	42.4
CLAN [7]	74.2	58.6	28.0	76.2	33.1	36.7	6.7	31.9	31.4	43.2
Ours	78.1	59.4	29.8	79.6	36.5	41.6	0.1	23.6	25.3	45.0
IntraDA [9]	80.2	59.3	29.4	86.4	33.6	53.9	0.0	32.7	37.6	46.3
Ours + intra	81.4	59.8	29.8	84.2	35.3	44.9	0.0	28.8	33.7	47.5



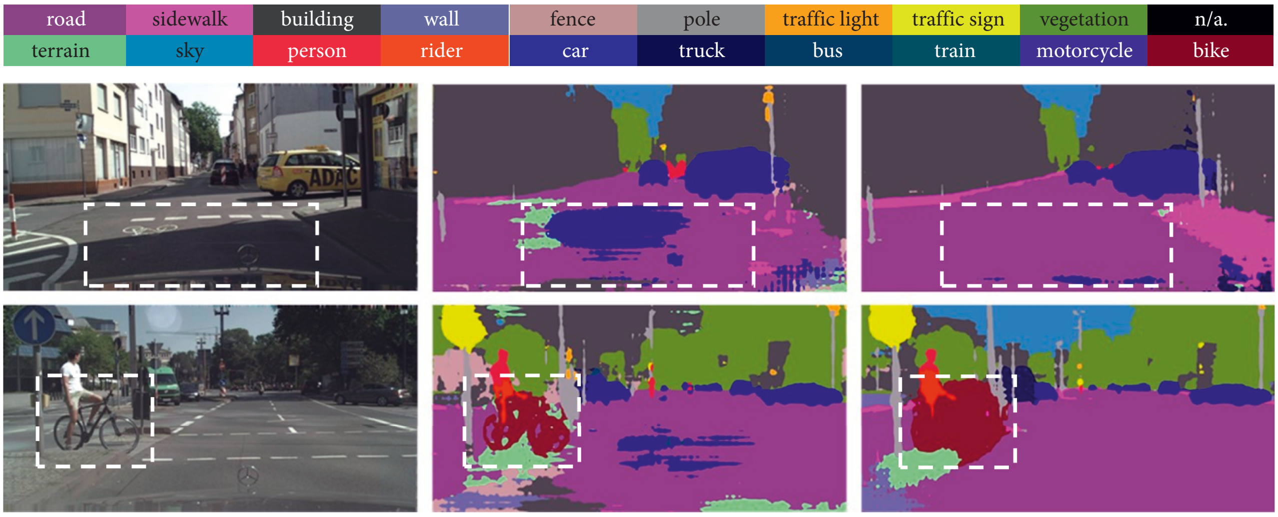
(a) Input images (b) Ground truth (c) ADVENT (d) Ours (e) Ours + Intra

Figure 4: The example results of adapted segmentation. (a, b) The images from Cityscapes validation dataset and the corresponding ground truth annotation. (c) The predicted segmentation maps of the ADVENT. (d, e) are the predicted segmentation maps from our proposed methods.

types of domain adaptation have the best of 47.5% mIoU. Our results show that the addition of image-level domain adaptation can lead to better performance.

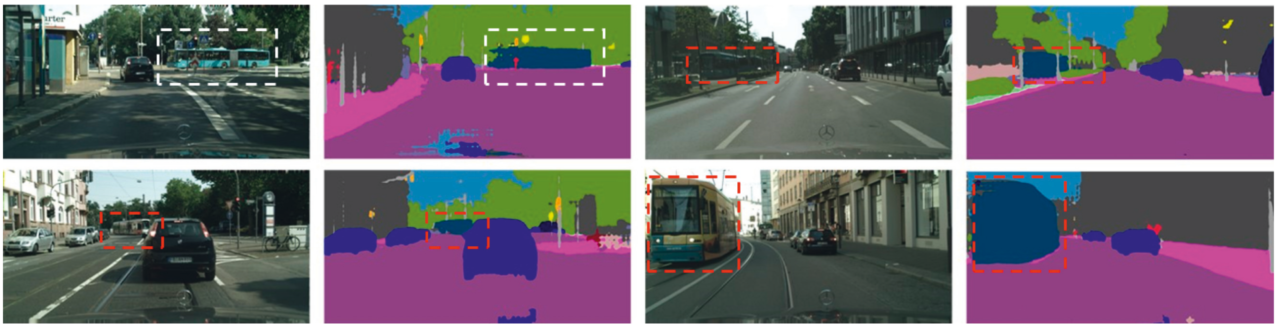
Compared with some previous works, such as AdvEnt, AdaSegNet, and CLAN, our proposed method improves the mIoU of 3.8%, 5.2%, and 4.4%. Additionally, compared with IntraDA, our method improves the mIoU by 1.2%. Interestingly, from Table 1, we can see that there is a significant improvement in accuracy for sidewalk and sign. This can be attributed to the fact that the enhanced images were able to bridge the layout gap for sidewalk and sign, where the domain distribution between game and real world images is very different. Figure 4 also shows the segmentation results. From Figure 4, we can see that the results for sidewalk and sign are close to the ground truth, which confirms that the qualitative evaluation and subjective observation are in agreement. The improved accuracy is due to the successful application of image-level domain adaptation to narrow the domain gap.

From the top line in Figure 5, our approach improves the error detection of semantic segmentation maps in the road.



(a) Input images (b) Baseline (c) Ours + Intra

Figure 5: Comparison of the Baseline and our proposed method.



(a) Input images (b) Ours + Intra (a) Input images (b) Ours + Intra

Figure 6: The semantic segmentation results of bus and train. The white dot lines represent that the ground truth label is bus and the semantic label is bus. The red dot lines represent that the ground truth label is train and the semantic label is bus.

This is because adversarial learning using the method of minimizing entropy loss is more effective. However, as shown in the bottom line of Figure 5, our approach worsens semantic segmentation for objects with a detailed structure, such as bike. The method of minimizing entropy loss also involves the disappearance of semantic segment maps when there is a small number of pixels on an object. Therefore, our future work will improve the method of minimizing entropy loss to prevent the disappearance of segment information.

Regarding the semantic segmentation map of train and bus, Figure 6 shows the example of train error detection. In this case, the error of semantic segmentation maps is caused by some reasons. The training dataset and validation dataset have a disproportionate number of train and bus. The validation dataset has a small number of trains. In contrast, the training dataset has a large number of buses. Therefore, in almost cases, the train is segmented as a bus. Additionally, the appearance and area of existence of bus and trains are similar. Therefore, the reinforcement learning algorithms of the segmentation map, including train, will be required.

6. Conclusions

In this work, we propose a domain adaptation framework, including three types. The semantic segmentation using the proposed framework achieved the best of 47.5% mIoU, and compared with IntraDA, our method improves the mIoU by 1.2%. Thereby, the effectiveness of image-level domain adaptation for improving the accuracy of semantic segmentation is confirmed. In particular, the semantic segmentation map of sidewalk and sign is significantly improved by the proposed method. However, by minimizing entropy loss, our approach worsens the semantic segmentation map for objects with a detailed structure, such as bike. Moreover, from the result in Figure 6, it is not easy to detect the semantic segmentation map of the train without reinforcement learning. Additionally, discussions concerning the numeral evaluations about the photorealism of the datasets are required as the future work. We believe that the performance can be improved by using a more robust detection architecture for semantic segmentation in future work.

Data Availability

The datasets can be acquired by contacting t.katayama@tokushima-u.ac.jp.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by JSPS KAKENHI Grant nos. 20K11790, 20K11889, and 22K1791300, "NSFC No. 61701297" NSFC No. 61701297 and in part by Tokushima University (TU) and the National Taiwan University of Science and Technology (TAIWAN TECH) Joint Research Program under Grant no. TU-NTUST-109-05.

References

[1] M. Cordts, M. Omran, S. Ramos et al., "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3213–3223, Las Vegas, NV, USA, 2016.

[2] X. Jiang, F. R. Yu, T. Song, and V. C. M. Leung, "Intelligent resource allocation for video analytics in blockchain-enabled internet of autonomous vehicles with edge computing," *IEEE Internet of Things Journal*, 2020, https://ieeexplore.ieee.org/document/9205310.

[3] X. Jiang, F. R. Yu, and T. Song, "Blockchain-enabled cross-domain object detection for autonomous driving: A model sharing approach," *IEEE Internet of Things Journal*, vol. 7, no. 5, pp. 3681–3692, 2020.

[4] H. Chen, C. Wu, B. Du, and L. Zhang, "DSDANet: deep siamese domain adaptation convolutional neural network for cross-domain change detection," 2020, https://arxiv.org/abs/2006.09225.

[5] J. Jiang, X. Wang, and M. Long, "Resource efficient domain adaptation," in *Proceedings of the 28th ACM International Conference on Multimedia (ACMMM)*, pp. 2220–2228, Seattle, WA, USA, 2020.

[6] Y.-H. Tsai, K. Sohn, S. Schulter, and M. Chandraker, "Domain adaptation for structured output via discriminative patch representations," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1456–1465, Seoul, Korea (South), 2019.

[7] Y. Luo, L. Zheng, T. Guan, J. Yu, and Y. Yang, "Taking a closer look at domain shift: category-level adversaries for semantics consistent domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2507–2516, Long Beach, CA, USA, 2019.

[8] T.-H. Vu, H. Jain, and M. Bucher, "ADVENT: adversarial entropy minimization for domain adaptation in semantic segmentation," *CVPR*, pp. 2517–2526, 2019.

[9] F. Pan, I. Shin, F. Rameau, S. Lee, and I. S. Kweon, "Unsupervised intra-domain adaptation for semantic segmentation through self-supervision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3763–3772, Seattle, WA, USA, 2020.

[10] S. R. Richter, H. A. AlHaija, and V. Koltun, "Enhancing photorealism enhancement," 2021, https://arxiv.org/abs/2105.04619.

[11] K. Nakajima, T. Katayama, T. Song, X. Jiang, and T. Shimamoto, "Domain adaptive semantic segmentation through photorealistic enhancement of video game," in *Proceedings of the IEEE International Conference on Consumer Electronics (ICCE)*, Las Vegas, NV, USA, 2022.

[12] J. T. Springenberg, "Unsupervised and semi-supervised learning with categorical generative adversarial networks," in *Proceedings of the International Conference on Learning Representation (ICLR)*, San Juan, Puerto Rico, 2016.

[13] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2242–2251, Venice, Italy, 2017.

[14] S. R. Richter, V. Vineet, and S. Koltun, "Playing for Data: ground truth from computer games," *Computer Vision - ECCV 2016*, vol. 9906, pp. 102–118, 2016.

[15] Y. Ganin, E. Ustinova, and H. Ajakan, "Domain-adversarial training of neural networks," *Journal of Machine Learning Research*, vol. 17, pp. 1–35, 2016.

[16] J. Hoffman, D. Wang, F. Yu, and T. Darrell, "FCNs in the wild: Pixel-level adversarial and constraint-based adaptation," p. 02649, 2016, https://www.vis.xyz/pub/fcns-in-the-wild/.

[17] D. Pathak, P. Krahenbuhl, and T. Darrell, "Constrained convolutional neural networks for weakly supervised segmentation," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1796–1840, Santiago, Chile, 2015.

[18] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proceedings of the International Conference on International Conference on Machine Learning (ICML)*, vol. 37, pp. 1180–1189, Lille, France, 2015.

[19] Y.-H. Tsai, W.-C. Hing, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker, "Learning to adapt structured output space for semantic segmentation," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7472–7481, Salt Lake City, UT, USA, 2018.

[20] L.-C. Chen, G. Papandreou, and I. Kokkinos, "Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, pp. 834–848, 2017.

[21] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proceedings of the International Conference on Learning Representations (ICLR)*, San Juan, Puerto Rico, 2016.

[22] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016.

[24] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: a large-scale hierarchical image database," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Miami, FL, USA, 2009.