Improving Synthetic to Realistic Semantic Segmentation with Parallel Generative Ensembles for Autonomous Urban Driving

Yi, Dewei; Fang, Hui; Hua, Yining; Su, Jinya; Quddus, Mohammed; Han, Jungong

Published in:
IEEE Transactions on Cognitive and Developmental Systems

DOI:
10.1109/TCDS.2021.3117925

Publication date:
2021

Citation for published version (APA):
Yi, D., Fang, H., Hua, Y., Su, J., Quddus, M., & Han, J. (2021). Improving Synthetic to Realistic Semantic Segmentation with Parallel Generative Ensembles for Autonomous Urban Driving. *IEEE Transactions on Cognitive and Developmental Systems*. https://doi.org/10.1109/TCDS.2021.3117925

Improving Synthetic to Realistic Semantic Segmentation with Parallel Generative Ensembles for Autonomous Urban Driving

Dewei Yi, Hui Fang, Yining Hua, Jinya Su, Mohammed Quddus, and Jungong Han

***Abstract*—Semantic segmentation is paramount for autonomous vehicles to have a deeper understanding of the surrounding traffic environment and enhance safety. Deep neural networks (DNN) have achieved remarkable performances in semantic segmentation. However, training such a DNN requires a large amount of labelled data at pixel level. In practice, it is a labour-intensive task to manually annotate dense pixel-level labels. To tackle the problem associated with a small amount of labelled data, Deep Domain Adaptation (DDA) methods have recently been developed to examine the use of synthetic driving scenes so as to significantly reduce the manual annotation cost. Despite remarkable advances, these methods unfortunately suffer from the generalisability problem that fails to provide a holistic representation of the mapping from the source image domain to the target image domain. In this paper, we therefore develop a novel ensembled DDA to train models with different upsampling strategies, discrepancy and segmentation loss functions. The models are, therefore, complementary with each other to achieve better generalisation in the target image domain. Such a design does not only improve the adapted semantic segmentation performance, but also strengthen the model reliability and robustness. Extensive experimental results demonstrate the superiorities of our approach over several state-of-the-art methods.**

***Index Terms*—Autonomous vehicles; Image processing; Deep learning; Domain adaptation; Semantic segmentation; Generative adversarial network;**

I. Introduction

A deep neural network (DNN) is powerful for extracting rich hierarchical feature representations [1, 2]. The superiority of feature extraction helps DNN based approaches to make compelling achievement on semantic segmentation. Deep learning based image segmentation model [3, 4] has been utilised to understand the surrounding traffic environment of the autonomous vehicle to enhance its driving safety. When deploying such a model, each pixel of an image is assigned to one of the semantic classes, such as car, truck, tree, or pedestrian. Since a fully convolutional network (FCN) [5] was proposed, it has significantly outperformed traditional computer vision methods. Recently, many studies, including U-Net [6] and SegNet [7], have extended the idea of FCN and achieved top-performance in semantic segmentation. However, these methods require a vast amount of labour-intensive work to label the dense image at pixel level. For instance, it takes about one and a half hours to annotate an image from Cityscapes dataset, which is unaffordable for the most of real-world applications.

Deep domain adaptation (DDA) [8, 9] is one of the most promising paradigms to achieve a generalised model without committing intensive manual labelling. The underlying idea is to minimise the discrepancy between two domains, i.e. the source and target domains. Assuming that there exists a huge amount of free annotated data in the source domain, e.g. synthetic driving scenes, while no labelled data is available in the target domain, e.g. realistic driving scenes. DDA approaches seek to find domain invariant feature representations or domain transformation functions, so that generalised models can be trained based on the data from the source domain and deployed in the target domain.

Many DDA methods [9–11] have been proposed to narrow the domain gap, which can be broadly categorised into two types [12]. The first category is known as a feature distribution alignment between the source and target domains. In these methods [13, 14], the similarity of the feature distribution from these two domains is maximised by measuring certain predefined distance metrics. The second category improves the quality of domain alignment via adversarial learning [9, 12, 15], where the generative adversarial network (GAN) is used at either pixel-level, feature-level, or output-level to ensure that the source and target domains share common characteristics across the deep learning based segmentation pipeline.

Despite the popularity of generative adversarial networks, a common failure pattern is observed while training GANs is the collapsing of large volumes of probability mass onto a few modes as highlighted in [16], where it can model one part of the data distribution well but fail to represent the entire distribution in the target domain. For a single generative network, it is difficult to guarantee the generalisability of all cases in the target domain due to insufficient learning of diversity. To deal with the problem, the concept of ensembles

Manuscript received...This work was supported by the University of Aberdeen Internal Funding to Pump-Prime Interdisciplinary Research and Impact under grant number SF10206-57. (Corresponding author: Yining Hua)

Dewei Yi is with Department of Computing Science, University of Aberdeen, Aberdeen AB24 3UE, UK. (email: dewei.yi@abdn.ac.uk).

Hui Fang is with Department of Computer Science, Loughborough University, Loughborough LE11 3TU, UK. (email: h.fang@lboro.ac.uk).

Yining Hua is with School of Arts, University of Roehampton, London SW15 5PH, UK. (email: yining.hua@roehampton.ac.uk).

Jinya Su is with School of Computer Science and Electronic Engineering, University of Essex, Colchester, CO4 3SQ, UK. (email: j.su@essex.ac.uk).

Mohammed Quddus is with School of Architecture, Building and Civil Engineering, Loughborough University, LE11 3TU Loughborough, U.K. (e-mail: m.a.quddus@lboro.ac.uk).

Jungong Han is with the Department of Computer Science, Aberystwyth University, Aberystwyth SY23 3FL, U.K. (e-mail: juh22@aber.ac.uk).

could be introduced to better represent the data distribution so that generative networks can explore the diverse alignments, between source and target domains, at the global level (adversarial learning), category level (co-training), and local level (ensemble scheme).

In this paper, we develop parallel generative ensembles of GAN (PGE-GAN) to improve the performance and reliability of traditional DDA algorithms in the semantic segmentation applications. In particular, several GANs are trained in parallel with different discrepancy loss and segmentation loss functions under different upsampling strategies. The idea behind is this concept that different discrepancy loss functions, segmentation loss functions, and upsampling strategies have their own strengths to recognise specific semantic classes in a driving scene, so that the ensembles of them is likely to provide a more holistic distribution in the target domain. In light of this, these ensembles such as ensembles for discrepancy loss, ensembles for segmentation loss, and ensembles for upsampling, are incorporated to enhance the model generalisability. The main contributions of this paper are summarised below:

- We develop a novel ensemble based DDA method by integrating multiple GAN networks. The ensemble scheme achieves remarkable performance on adapted semantic image segmentation applications when comparing to other advanced DDA methods.
- When training the parallel models in our framework, we add new optimisation targets in loss functions. These include (i) the generalised dice loss term in the segmentation loss function and (ii) the Pearson similarity in the discrepancy loss function. It is noted that the generalised dice loss can reduce overfitting problem for those classes with a small number of training data, and the Pearson similarity can alleviate the effect of scaling and shifting when the dimensions of variables are significantly different and their values may be noisy or random.
- For each GAN model, a mixture co-training framework is adopted to learn the multi-views of the same inputs via maximising the divergence of different classifiers. Both Pearson and cosine similarities are mixed with co-training framework to derive various views. Then, more diversity is introduced to generalise better in the target domain.
- Comprehensive comparison and ablation study are performed to demonstrate the superiority of our proposed method against state-of-the-art domain adaptation methods on transferring from GTA5 and SYNTHIA synthetic images to Cityscapes realistic images.

II. Related Work

A. Semantic segmentation

Semantic segmentation predicts pixel-level labels for an image to distinct objects. In previous decades, hand-crafted features were commonly used to achieve the task of semantic segmentation. These hand-crafted features are defined with the help of domain experts. An alternative way is to extract features by deep neural networks, which is able to extract efficient features automatically. Since deep learning based methods reveal their outstanding performance in feature extraction, recent work on semantic segmentation is mainly conducted through deep neural networks, such as fully connected network (FCN) [17], U-Net [6], and Seg-Net [18]. To extract efficient features, these advanced networks need to be trained by a substantial amount of dense pixel annotations while it is difficult to obtain a large number of pixel-level labels in real-word applications. To deal with this problem, annotated data can be collected from a simulator, where the pixel-level annotations can be achieved automatically. Although some advanced simulators can synthesise high-fidelity data, there still exists a gap between synthetic data and real-world data and the gap can be bridged through a technique known as *domain adaptation*.

B. Domain Adaptation

It is noted that most existing machine learning algorithms assume that the training and testing data are drawn from the same underlying distribution [19]. However, such an assumption is not always true in practice [20]. This issue often occurs when transferring knowledge from synthetic images to real images [10, 21, 22], because the domain shift exists between the training and testing data [9, 11]. Domain adaptation, a method intending to solve this issue, learns the transformation to align cross-domain data with the class regularity so as to achieve a better generalisation in the target domain [12]. Some approaches minimize the discrepancy of domain distribution through learning domain-invariant representations, where the domain distribution discrepancy can be calculated by maximum mean discrepancy or mean and covariance of feature distribution [13, 14]. Unfortunately, it is usually not sufficient to match source and target data by solely aligning mean and covariance (i.e. low order moments) of the distribution. In recent year, adversarial learning is adopted in domain adaptation as it is insensitive to the feature distribution.

C. Adversarial Learning

Adversarial learning can minimise the discrepancy of different domains by using an adversarial objective with regard to a domain discriminator [12]. GAN is one of the most popular adversarial learning methods, which consists of a generative network G and a discriminative network D. These two networks pit against each other during the training phase, where G dedicates to generating more realistic synthetic data and D aims to distinguish the synthetic data from the real data. In [23], the adaptation loss function is designed with the expected loss in the source domain, the domain divergence compared to the target domain, and the shared error of the ideal joint hypothesis on these two domains. In category-level, [15] improves semantic consistency in the target domain by aligning the distribution shift in the latent feature space. At the pixel-level, style transfer is widely adopted as it aims to make data indistinguishable across domains [24, 25]. Different from the above studies, [9, 11] consider the alignment of category-level with pixel-level simultaneously to enable the joint optimisation for the representation and the prediction.

III. PARALLEL GENERATIVE ENSEMBLES OF GAN (PGE-GAN)

Driving scene translation can be formulated as follows: given a source domain image x_s with the corresponding ground truth y_s drawn from the source set $\{X_S, Y_S\}$ and a target domain image x_t from the target set X_T without labels, the objective is to learn a generative model G for transferring knowledge from the source domain to the target domain so that G can correctly predict labels (e.g. road, building, sign, etc.) at the pixel-level in the target domain.

Our proposed PGE-GAN method is illustrated in this section, which can learn more transferable knowledge across the domains. We enhance the co-training framework by using the ensemble scheme. Each model in PGE-GAN is trained based on different types of upsampling strategies under various discrepancy loss and segmentation loss functions to obtain diverse predictions.

A. Architecture of PGE-GAN

Our network architecture of each model consists of a generative network G and a discriminator D, where G is a fully-convolutional segmentation network and D is a convolutional classification network. As illustrated in Fig. 1, G is separated into feature extractor F and two classifiers C_1 and C_2. F is used to extract features from input images and then C_1 and C_2 predict pixel-level labels by using the extracted features. To derive the divergence of co-training classifiers, we conduct the weight diversity of C_1 and C_2 through maximising the cosine or Pearson distance loss for different ensembles in the training phase. Subsequently, the distinct views of a feature can be provided by C_1 and C_2 so as to make more reliable semantic predictions. In our work, each (e-th) model has two classifiers C_1^e and C_2^e and their corresponding predictions are p_1^e and p_2^e. The final prediction map p^e for e-th model is obtained through adding up the predictions of p_1^e and p_2^e. The network G and D are alternately trained until the maximum epoch is reached. Given a source domain image $x_s \in X_S$, feature extractor F provides a feature map for classifiers C_1^e and C_2^e so as to derive the semantic prediction map p^e. The p^e is not only the input to D for computing adversarial loss, but also is compared with the ground-truth label $y_s \in Y_S$ to derive segmentation loss. Given a target domain image $x_t \in X_T$, it is the input to G and then a semantic prediction map p^e is generated from G. For the target data flow, we adopt the discrepancy between the two predictions p_1^e and p_2^e of ensemble p^e as an indicator to weight the adversarial loss.

B. Loss Function

The loss function of the proposed network consists of three losses: (i) discrepancy loss, (ii) segmentation loss, and (iii) adversarial loss.

Discrepancy Loss: As explained in [12], the co-training framework can provide different views of the same feature. To increase diversity, two similarity metrics are introduced: Cosine [26] and Pearson [27]. Two classifiers of the co-training framework need to have diverse parameters. For a deep neural network, the diversity comes from the weights of specific layers. One popular way, to obtain the difference between co-training classifiers, is to calculate their Cosine similarity as mentioned in [26]. Here, the discrepancy loss can be measured by cosine similarity as follows.

$$L_{cosine} = \frac{\vec{w_1} \cdot \vec{w_2}}{\|\vec{w_1}\| \|\vec{w_2}\|} \tag{1}$$

where $\overrightarrow{w_1}$ and $\overrightarrow{w_2}$ are flattened and concatenated weights of convolutional kernels, which belong to the two co-training classifiers.

Taking into consideration the variable weight scales and randomness, we also measure divergence of the two classifiers by the Pearson similarity, which is a scale-free metric [27]. The Pearson similarity is defined as follows.

$$L_{pearson} = \frac{(\vec{w}_1 - \overline{w}_1) \cdot (\vec{w}_2 - \overline{w}_2)}{\|\vec{w_1} - \overline{w}_1\| \|\vec{w_2} - \overline{w}_2\|} \tag{2}$$

where $\overline{w}_1$ and $\overline{w}_2$ are the means of $\vec{w}_1$ and $\vec{w}_2$, respectively.

Segmentation Loss: A source domain image x_s with the size of $H \times W$ and a label map y_s are given, where the shape of y_s is $C \times H \times W$ and C is the number of semantic classes. Here, different kinds of segmentation loss are discussed including multi-class cross-entropy loss, generalised dice loss, and their combination. The definition of multi-class cross-entropy loss can be computed by Equation (3).

$$L_{ce}(G) = \sum_{c=1}^{C} \sum_{i=1}^{H \times W} -g_{ic} \log p_{ic} \tag{3}$$

where p_{ic} is the predicted probability of class c on pixel i and g_{ic} is the ground truth of pixel i. That is, g_{ic} is assigned 1 when pixel i belongs to class c, otherwise g_{ic} is assigned 0.

The dice loss is a measure of labelled regions of images which is used to evaluate segmentation performance. The dice loss outperforms other loss functions in the case of a severe class imbalance. However, traditional dice loss can only deal with binary problems. To overcome this issue, we use the generalised dice loss to access multiple class segmentation with a single score, which is given in Equation (4).

$$L_{gdl}(G) = 1 - 2\frac{\sum_{c=1}^{C} w_c \sum_{i=1}^{H \times W} r_{ic} p_{ic}}{\sum_{c=1}^{C} w_c \sum_{i=1}^{H \times W} (r_{ic} + p_{ic})} \tag{4}$$

where w_c is the weight of invariance for semantic class c and r_{ic} is the ground truth. To address the issues of small object detection and class imbalance, their combination is introduced to derive the loss [28]. According to Equation (3) and (4), the combined loss function of segmentation is given by

$$\begin{aligned} L_{ce-gdl}(G) &= w_{ce} L_{ce} + w_{gdl} L_{gdl} \\ &= w_{ce} \left(\sum_{c=1}^{C} \sum_{i=1}^{H \times W} -g_{ic} \log p_{ic} \right) + \\ &\quad w_{gdl} \left(1 - 2\frac{\sum_{c=1}^{C} w_c \sum_{i=1}^{H \times W} r_{ir} p_{ic}}{\sum_{c=1}^{C} w_c \sum_{i=1}^{H \times W} (r_{ic} + p_{ic})} \right) \end{aligned} \tag{5}$$

where w_{ce} is the weight of cross entropy loss and w_{gdl} is the weight of generalised dice loss for multi-class segmentation.

Adversarial Loss: Adversarial learning is to train a generative model G that generates target domain samples to confuse

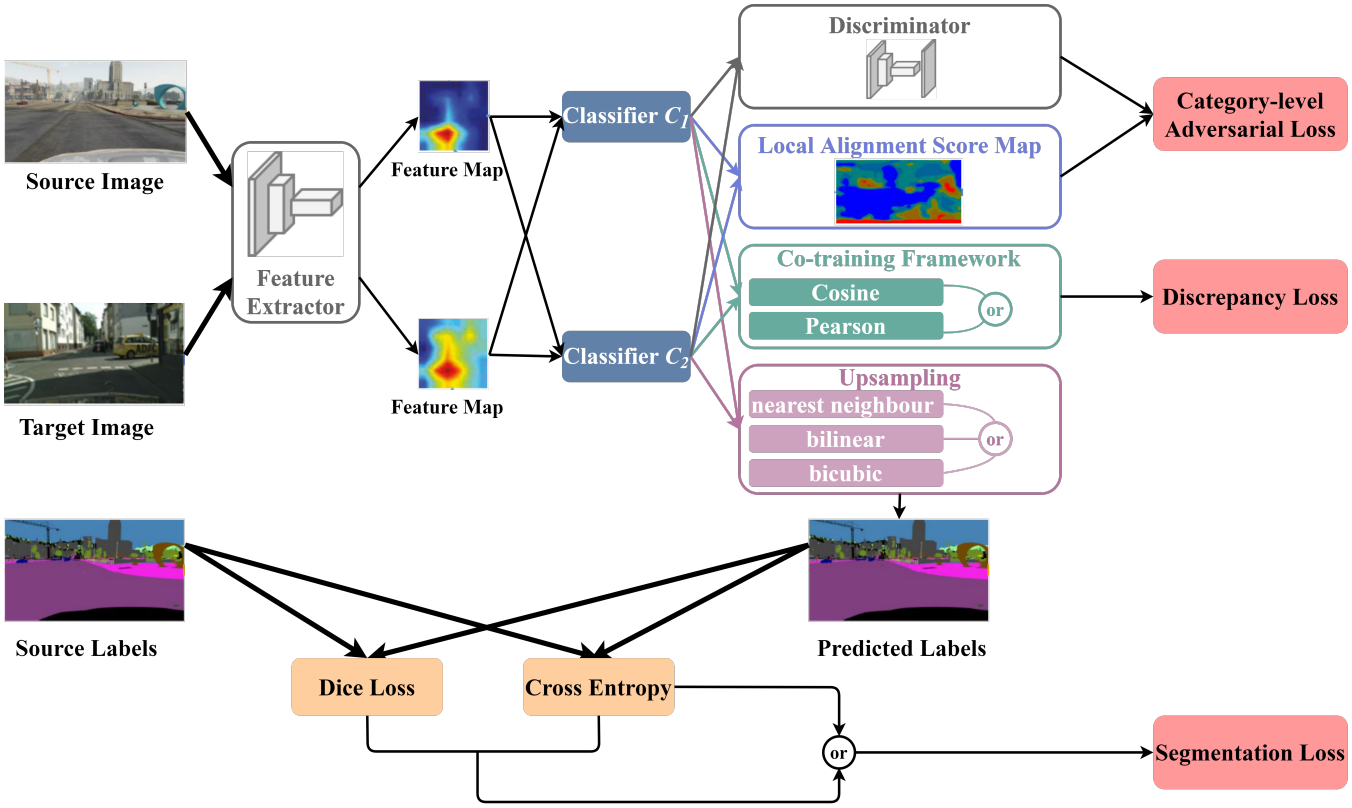

Fig. 1. Architecture of PGE-GAN. It illustrates the overall pipeline of the proposed ensembling framework, including the feature extractor, co-training classifiers, and the options of loss functions to train the ensembling GANs in our method.

the domain discriminator D, where D is able to distinguish between samples of source (synthetic images) and target (realistic images) domains [12]. In order to learn domain-invariant features, G needs to minimise the divergence between the source and target domains. D needs to maximise the classification performance. This property is achieved through minimaxing an adversarial loss as shown below:

$$\begin{aligned} L_{adv}(G,D) = &-E\left[\log(D(G(X_S)))\right] \\ &- E\left[(\lambda_{local} S(p_1^e, p_2^e) + \epsilon)\log(1 - D(G(X_T)))\right] \end{aligned} \tag{6}$$

where p_1^e and p_2^e are predictions of the co-training framework for ensemble e. $S(p_1^e, p_2^e)$ denotes the similarity of the two predictions. The similarity can be computed by cosine or Pearson, which is determined by the ensemble type. $[\lambda_{local} S(p_1^e, p_2^e) + \epsilon]$ represents the adaptive weight of adversarial loss. The impact of adversarial loss is controlled through λ_{local} in overall training objective. ϵ is used to improve the stability of training processing.

C. Description of Ensembles

Ensembles for Discrepancy Loss: A mixture co-training framework is used to formulate distinct views of features as suggested in [12]. Our proposed method trains two co-training classifiers for each individual model, and the divergence of the two classifiers is measured through discrepancy loss. Cosine and Pearson are two similarity metrics as the loss function terms. Their main difference is that one is scaling-related and the other is scaling-free. To guarantee the diversity of models in our PGE-GAN, both cosine and Pearson similarity metrics are drawn to compute the discrepancy loss of co-training classifiers. Therefore, two ensembles are derived by considering diverse discrepancy loss. One of the corresponding training objectives is given in Equation (7) with bicubic upsampling and the other one is provided in Equation (8).

$$\begin{aligned} G^*, D^* = \arg\min_G \max_D [&L_{ce}(G) + \lambda_{dl} L_{cosine}(G) \\ &+ \lambda_{adv} L_{adv}(G,D)] \end{aligned} \tag{7}$$

$$\begin{aligned} G^*, D^* = \arg\min_G \max_D [&L_{ce}(G) + \lambda_{dl} L_{Pearson}(G) \\ &+ \lambda_{adv} L_{adv}(G,D)] \end{aligned} \tag{8}$$

where $L_{ce}(G)$ is a multi-class cross entropy loss. $L_{cosine}(G)$ and $L_{Pearson}(G)$ are the two types of discrepancy loss. $L_{adv}(G,D)$ is the adversarial loss. Moreover, λ_{dl} and λ_{adv} are the weights of discrepancy loss and adversarial loss respectively, which control the relative importance of different kinds of losses.

Ensembles for Segmentation Loss: Cross entropy is commonly used to calculate the segmentation loss, which evaluates the class predictions for each pixel individually and then averages over all pixels. However, cross entropy underperforms in handling imbalanced data. To address this issue, the combination of cross entropy and generalised dice loss is proposed to achieve better results for those imbalanced semantic classes. To keep diversity in the ensembles for both balanced and imbalanced semantic classes, two segmentation loss functions are designed in our method. One uses cross

entropy and the other one uses the combination of cross entropy and dice loss. Therefore, Equation (9) and Equation (10) are the overall training objectives of the two ensembles.

$$G^*, D^* = \arg\min_G \max_D [L_{ce-gdl}(G) + \lambda_{dl} L_{Pearson}(G) + \lambda_{adv} L_{adv}(G, D)] \tag{9}$$

$$G^*, D^* = \arg\min_G \max_D [L_{ce-gdl}(G) + \lambda_{dl} L_{cosine}(G) + \lambda_{adv} L_{adv}(G, D)] \tag{10}$$

where $L_{ce-gdl}(G)$ is the segmentation loss measured by the combination of cross entropy and generalised dice loss.

Ensembles for upsampling: In the previous work, only one upsampling strategy is utilised to derive the original size of images. There are three popular upsampling strategies including nearest neighbour upsampling, bilinear upsampling, and bicubic upsampling. However, it is difficult to determine which upsampling strategies should be used. To determine an interpolated pixel, it mainly relies on the nearest pixel or several surrounding pixels, even on more surrounding pixels. To this end, three ensembles are trained corresponding to the nearest neighbour, bilinear, and bicubic upsampling strategies. Such ensemble decision scheme improves the performance and reliability of upsampling.

The nearest neighbour upsampling strategy is to interpolate a new pixel according to the nearest existing pixel. The bilinear and bicubic upsampling strategies mentioned in [29] are presented in Equation (11) and (12).

$$G = G_{bil}(x, y) = (1-t)\left[(1-s)g_{i,j} + sg_i + sg_{i+1,j}\right] + t\left[(1-s)g_{i,j+1} + sg_{i+1,j+1}\right] \tag{11}$$

where G is a two-dimensional digital image. $G(x, y)$ is a upsampling function to pixel at the position of (x, y) in the upsampled image. g is an raw output without upsampling. $G_{bil}(\bullet, \bullet)$ is the bilinear upsampling function, where (x, y) is a interpolated pixel and $s = x - x_i$, $t = y - y_i$.

$$G = G_{bic}(x, y) = \sum_{n=-1}^{2} \sum_{m=-1}^{2} g_{i+m,j+n} P_{m+1}(s) P_{n+1}(t) \tag{12}$$

where $G_{bic}(\bullet, \bullet)$ is the bicubic upsampling function. $P_0(z) = (-z^3 + 2z^2 - s)/2$, $p_1(z) = (3z^3 - 5z^2 + 2)/2$, and $p_3(z) = (z^3 - z^2)/2$.

D. Weights of Ensembles

Different upsampling strategies rely on different number of existing pixels to interpolate a new one. For instance, it requires different number of existing pixels for nearest neighbour, bilinear, and bicubic upsampling strategies to interpolate a new pixel. As suggested by [29], it results in a better upsampling when using more existing pixels. To this end, we determine the weight of each ensemble with regard to its upsampling strategy and the final prediction of a pixel is derived from a set of weighted ensembles as below.

$$P_f = \arg\max_C \left(\sum_{i=1}^{n} \sum_{e=1}^{E} w^e p_i^e \right) \tag{13}$$

TABLE I
THE CONFIGURATION OF NETWORKS

Network Type	Generative network G	Discriminative network D
Optimiser	SGD	Adam
Initial learning rate	2.5×10^{-4}	5×10^{-5}
Momentum	0.9	–
β_1	-	0.9
β_2	-	0.99
Decay	5×10^{-4}	
λ_{local}	40	
λ_{dl}	0.01	
λ_{adv}	1×10^{-3}	
ϵ	0.4	
Max Epoch	1×10^5	

where P_f is the final pixel-level prediction of semantic class, n is the number of co-training classifiers for each ensemble and E is the number of ensembles. w^e is the weight of e-th ensemble and p_i^e is a vector of i-th co-training classifier for e-th ensemble. Such vector corresponds to all semantic classes and each element of the vector is the prediction value of its corresponding semantic class. In our model, the weights are chosen based on the upsampling methods, so different upsampling methods lead to different weights. Specifically, in our proposed method, the setting of the GAN networking combines 1 nearest neighbouring, 1 bilinear and 4 bicubic upsampling method together, which rely on 1, 4, and 16 pixels, respectively. Thus, one interpolated pixel in our proposed method is determined by 69 existing pixels. The weight of prediction model with bicubic upsampling method is 4 times that of the bilinear one, and 16 times that of the nearest neighbour one.

IV. EXPERIMENTAL EVALUATION

In this section, experimental evaluation is conducted on the proposed method for synthetic-to-realistic translation. The experiments are evaluated by synthetic and realistic datasets, which are described in Section IV-A. The implementation details are discussed in Section IV-B, where the settings of generative and discriminative networks are clarified along with the explanation of platform. To quantitatively evaluate the performance of synthetic-to-realistic translation, our proposed method is compared against other methods in Section IV-C.

A. Datasets

In the experiments, three benchmark datasets are used. The GTA5 synthetic dataset [30] and SYNTHIA [31] are chosen as the source domains and the Cityscapes realistic dataset [32] is chosen as the target domain. In particular, GTA5 contains 24,966 high-resolution vehicle-egocentric images, where these synthetic images are produced by using a photorealistic open-world computer game called "Grand Theft Auto V". In addition, SYNTHIA is also a synthetic collection of imagery and annotations, a large-scale collection of photo-realistic frames rendered from some virtual cities. In SYNTHIA dataset, there are two image datasets and seven video sequences with a resolution of 1280x760. Specifically, it contains 9,400 images with 13-class categories. Some examples of GTA5 and SYNTHIA datasets are present in Fig. 2. In contrast, Cityscapes

is a real-life dataset including 5,000 images of street scenes in Germany and neighbouring countries. Some examples of Cityscapes are presented in Fig. 2. GTA5, SYNTHIA and Cityscapes datasets all provide dense pixel-level labels and their annotations are compatible with each other. Following the settings in [9, 12, 30], all 24,966 images from the GTA5 dataset and 9,400 synthetic images from SYNTHIA dataset are used for training generative networks, respectively. The validation set of Cityscapes dataset are used for assessing the performance.

B. Implementation Details

The proposed algorithm is implemented under PyTorch framework. In particular, the pre-trained ResNet-101 [33] is chosen as the backbone for source-only generative network G. For each GAN model, the last classification module is duplicated for co-training. The discriminative network structure in [12] is adopted to build up the discriminative network D. The discriminative network D includes 5 convolution layers with the kernel size of 4-by-4, where the channel number is $n_c \in \{1, 64, 128, 256, 512\}$ and stride step is 2. The activation function parametric ReLu is used to concatenate the convolutional layer, where the parameter is $\alpha = 0.2$. After the last layer, an upsampling layer is added to make the size of an output image match the size of a local alignment score map. Stochastic gradient descent (SGD) and Adam are used to optimise generative network G and discriminative network D, respectively. More details of network configuration are provided in Table I. In the training phase, an original input image is resized to the resolution of 512×1024. In the evaluation phase, a prediction map is up-sampled by a factor of 2 for assessing the performance of mean intersection of union (mIoU). All experiments are conducted on a PC with the following configuration: Intel 2.20 GHz Xeon(R) E5-2630 CPU, GeForce RTX 2080ti, 16 GB of RAM.

C. Performance of Semantic Segmentation

This section provides the adapted semantic segmentation results of driving scenes for various domain adaptation methods. All experiments are evaluated on synthetic datasets: GTA5 and SYNTHIA, and realistic dataset: Cityscapes. To quantitatively evaluate the results of semantic segmentation, intersection-over-union (IoU) is used to assess the performance for each semantic class as it is not affected by the class imbalances. The definition of IoU is given by

$$IoU(Y, \hat{Y}) = \frac{Y \cap \hat{Y}}{Y \cup \hat{Y}} = \frac{t_p}{t_p + f_n + f_p} \tag{14}$$

where Y are the ground truth labels of pixels and $\hat{Y}$ are the predictions of pixels. Moreover, t_p, f_n, f_p represent the true positives, false negatives, and false positives, respectively. In addition, the overall performance of different methods is measured through mean IoU which is calculated by averaging the IoU of various semantic classes. Here, we first show the performance of individual GAN models in the PGE-GAN. Subsequently, we compare our proposed method against a number of recently developed methods.

TABLE II
THE SETTINGS OF ENSEMBLES

Ensemble	Upsampling	Discrepancy Loss	Segmentation Loss
E_1	nearest neighbour	Cosine	Cross Entropy
E_2	bilinear	Cosine	Cross Entropy
E_3	bicubic	Cosine	Cross Entropy
E_4	bicubic	Pearson	Cross Entropy
E_5	bicubic	Cosine	CE-GDL
E_6	bicubic	Pearson	CE-GDL

1) Individual Model Performance in PGE-GAN: The settings of the GAN networks in our PGE-GAN are provided in Table. II. They are formed by using different combinations of upsampling strategies, discrepancy loss functions and segmentation loss functions. Here, the combination of cross entropy and generalised dice loss in the segmentation loss functions is represented by CE-GDL. It is found that each GAN network outperforms others on some of specific semantic classes. Thus, diverse ensembles can help avoid overfitting and improve the generalisation. The quantitative results of individual semantic classes and mIoU (overall) on GTA5 and Cityscapes datasets are summarised in Table III. In addition, the qualitative comparison of different ensembles are provided in Fig. 3. To identify the strength of different ensembles clearer, the results are obtained before image transferring. It illustrates that there exist variations across individual models, which indicates that the emsembling scheme can provide a mechanism to increase the reliability of the final outputs.

TABLE III
ADAPTED RESULTS FROM GTA5 TO CITYSCAPES FOR ENSEMBLES

	Ensembles						
Semantic Class	E_1	E_2	E_3	E_4	E_5	E_6	Combined
road	86.7	87.0	**88.4**	87.4	88.1	87.9	**88.9**
sidewalk	21.7	27.1	**32.9**	27.6	24.7	28.5	28.8
building	79.3	79.6	80.8	79.9	80.0	**81.1**	**81.5**
wall	28.4	27.3	28.5	25.5	27.3	**30.4**	**32.6**
fence	18.9	23.3	24.8	**24.9**	22.7	20.6	**25.4**
pole	19.6	28.3	30.0	28.5	29.7	**32.0**	30.9
light	28.5	**35.5**	34.9	31.6	34.4	31.7	35.0
sign	18.2	24.2	**26.1**	25.2	23.2	20.5	23.5
vegetation	82.1	83.6	83.2	82.9	**83.9**	83.3	**84.4**
terrain	32.9	27.4	34.3	32.6	**37.4**	37.3	**39.6**
sky	**76.0**	74.2	75.8	75.5	73.5	73.8	**76.3**
person	53.7	**58.6**	58.5	57.6	57.3	57.6	58.3
rider	25.4	28.0	**28.7**	27.0	27.8	26.0	**28.9**
car	81.7	76.2	83.2	80.6	**83.5**	84.2	**85.3**
truck	30.9	33.1	32.1	32.6	**33.9**	29.2	**35.2**
bus	40.7	36.7	44.7	**46.2**	45.4	43.5	**48.1**
train	0.29	**6.7**	1.42	3.9	0.34	0.9	0.8
motorcycle	23.6	**31.9**	24.9	27.7	25.3	20.8	28.1
bicycle	34.4	31.4	27.1	18.1	29.4	**35.2**	30.1
mIoU	41.2	43.2	**44.2**	42.9	43.6	43.4	**45.4**

2) Comparative Results: We compare the adapted semantic segmentation results between our proposed method with other state-of-the-art methods on the task of domain adaptation, which transfers the source domain data (GTA5 and SYNTHIA) to the target domain data (Cityscapes). The qualitative results of adapted segmentation are illustrated in Fig. 4 and the quantitative comparison is summarised in Table IV and V with regard to the IoU of various semantic classes and mIoU of all

Fig. 2. The examples of source domain are the synthetic images which are obtained from GTA5 and SYNTHIA datasets as shown in the first and second columns. These data are used to train the model. The examples of target domain are realistic images which are obtained from Cityscapes dataset. These data are used to evaluate model performance.

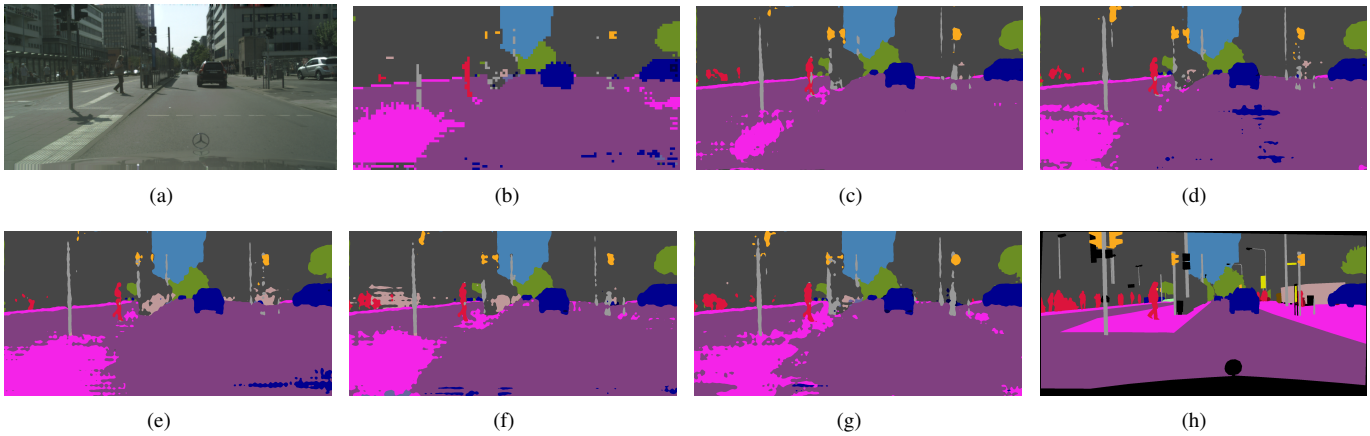

Fig. 3. Semantic segmentation results of different ensembles: (a) target image, (b)-(g) the adaption results of E_1-E_6, (h) ground truth annotations.

semantic classes (Top 3 performances are highlighted in blue). The following observations can be drawn:

(i) By introducing generative ensembles in parallel, our developed method can significantly improve the generalisation ability from the source domain to the target domain. As a result, our method outperforms other advanced methods in term of mIoU, which can reach 46.9% and 47.1% on GTA5 to Cityscapes and SYNTHIA to Cityscapes adaptation respectively. More specifically, compared to the results obtained by FCN wild [15] and AdvSemiSeg [38], our method can significantly improve segmentation results from 27.1% to 46.9% on GTA5 to Cityscapes adaptation and 34.9% to 47.1% on SYNTHIA to Cityscapes adaptation.

(ii) For various semantic classes on GTA5 to Cityscapes adaptation, when considering Top 3 performance of each segmentation class, there are 17 classes out of total 19 classes achieving Top 3 performance in our method except for the segmentation classes of fence and train. For the segmentation class of fence, the results of our method is very close to Top 3 performance, where our method can achieve 23.1% and Top 3 performance is

TABLE IV
QUANTITATIVE COMPARISON RESULTS FROM GTA5 TO CITYSCAPES (TOP 3 IN BLUE, TOP 1 IN **BLUE BOLD**)

Method	road	sidewalk	building	wall	fence	pole	light	sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorcycle	bicycle	mIoU
FCN wild [15]	70.4	32.4	62.1	14.9	5.4	10.9	14.2	2.7	79.2	21.3	64.6	44.1	4.2	70.4	8.0	7.3	0.0	3.5	0.0	27.1
MCD [10]	86.4	8.5	76.1	18.6	9.7	14.9	7.8	0.6	82.8	32.7	71.4	25.2	1.1	76.3	16.1	17.1	1.4	0.2	0.0	28.8
CDA [21]	74.9	22.0	71.7	6.0	11.9	8.4	16.3	11.1	75.7	13.3	66.5	38.0	9.3	55.2	18.8	18.9	0.0	16.8	14.6	28.9
CyCADA [11]	85.2	37.2	76.5	21.8	15.0	23.8	22.9	21.5	80.5	31.3	60.7	50.5	9.0	76.9	17.1	28.2	4.5	9.8	0.0	35.4
CBST [20]	**90.4**	**50.8**	72.0	18.3	9.5	27.2	28.6	14.1	82.4	25.1	70.8	42.6	14.5	76.9	5.9	12.5	1.2	14.0	28.6	36.1
DCAN [22]	82.3	26.7	77.4	23.7	20.5	20.4	30.3	15.9	80.9	25.4	69.5	52.6	11.1	79.6	24.9	21.2	1.3	17.0	6.7	36.2
IBAN [30]	88.2	33.7	80.1	23.4	21.8	27.7	27.9	16.3	83.2	38.3	76.2	57.5	20.3	81.1	25.9	33.4	1.9	22.4	20.7	40.7
AdaSegNet [9]	86.5	36.0	79.9	23.4	23.3	23.9	35.2	14.8	83.4	33.3	75.6	58.5	27.6	73.7	32.5	35.4	3.9	30.1	28.1	42.4
CLAN [12]	87.0	27.1	79.6	27.3	23.3	28.3	**35.5**	24.2	**83.6**	27.4	74.2	58.6	**28.0**	76.2	33.1	36.7	**6.7**	**31.9**	31.4	43.2
DS [34]	89.4	46.4	78.7	**34.0**	**26.9**	15.6	11.8	8.5	81.8	**40.5**	78.6	36.4	7.3	77.9	31.9	33.9	0.0	8.4	2.4	37.4
SIBAN [30]	88.5	35.4	79.5	26.3	24.3	28.5	32.5	18.3	81.2	40.0	76.5	58.1	25.8	82.6	30.3	34.4	3.4	21.6	21.5	42.6
CDA (full) [32]	72.9	30.0	74.9	12.1	13.2	15.3	16.8	14.1	79.3	14.5	75.5	35.7	10.0	62.1	20.6	19.0	0.0	19.3	12.0	31.4
ALST [35]	81.0	19.6	65.8	20.7	12.9	20.9	6.6	0.2	82.4	33.0	68.2	54.9	6.2	80.3	28.1	41.6	2.4	8.5	0.0	33.3
PGE-GAN (ours)	89.7	42.5	**82.7**	33.9	23.1	**29.3**	34.1	**30.4**	83.2	38.4	**80.9**	**59.2**	26.2	**83.7**	**42.3**	**48.7**	0.1	28.5	**33.6**	**46.9**
Oracle [36]	96.7	76.5	88.2	45.2	42.7	42.7	46.8	60.5	88.5	55.9	88.4	69.3	51.2	91.5	73.3	70.6	45.4	52.2	65.1	65.8

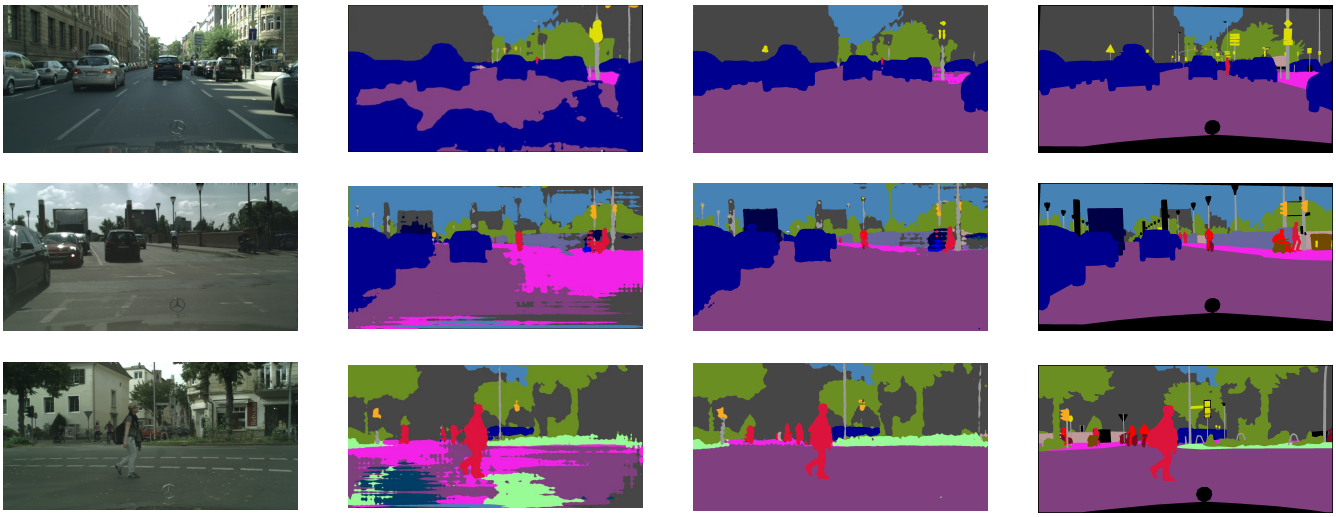

Fig. 4. Qualitative semantic segmentation results on adaptation from GTA5 to Cityscapes. From top to bottom: target image, non-adapted results (source only), adapted results with PGE-GAN, and the ground truth label map, respectively.

TABLE V
QUANTITATIVE COMPARISON RESULTS FROM SYNTHIA TO CITYSCAPES (TOP 3 IN BLUE, TOP 1 IN **BLUE BOLD**)

Method	road	sidewalk	building	light	sign	vegetation	sky	person	rider	car	bus	motorcycle	bicycle	mIoU
SemanticDA [37]	78.4	0.1	73.2	0.0	0.2	84.3	78.8	46.0	0.3	74.9	30.8	0.0	0.1	35.7
AdvSemiSeg [38]	72.5	0.0	63.8	0.0	0.5	**84.7**	76.9	45.3	1.5	77.6	31.3	0.0	0.1	34.9
SUIT [39]	75.1	**31.4**	77.4	11.7	15.0	79.2	77.4	54.2	18.1	78.1	27.4	9.4	30.2	45.0
IBAN [30]	78.2	19.7	**80.5**	9.4	8.9	77.4	**82.0**	**56.3**	9.6	76.3	22.8	17.5	23.3	43.2
AdaSegNet (feat. only) [9]	62.4	21.9	76.3	11.7	11.4	75.3	80.9	53.7	18.5	59.7	13.7	**20.6**	24.0	40.8
CLAN†[36]	78.0	34.1	78.1	8.8	13.4	78.1	81.5	55.3	21.1	66.4	22.3	12.4	31.5	44.7
CLAN (AMP) [12]	74.2	30.0	76.4	11.9	15.1	77.2	79.4	51.3	19.6	68.7	25.0	11.3	30.3	43.9
ALST [35]	**80.7**	0.3	75.0	0.0	0.4	84.0	79.4	46.6	0.8	80.8	32.8	0.5	0.5	37.0
PGE-GAN (ours, AMP)	73.9	29.9	77.0	**16.9**	**17.4**	78.4	81.9	48.9	**23.6**	73.3	**32.8**	17.5	**40.3**	**47.1**

23.3%. The difference is only 0.2%. For segmentation class of train, it is challenging to recognising train in urban scenes. We can see some existing methods (e.g. FCN wild [15], CDA [32], DS [34], and CDA [32]) even cannot correctly recognise the objects of trains at all, where the IoU is zero for these methods, and our method is still sightly better than these methods.

(iii) For various semantic classes on SYNTHIA to Cityscapes adaptation, when considering Top 3 performance of each segmentation class, there are 7 classes out of total 13 classes achieving Top 3 performance in our method. Our method has a superior performance on the semantic classes of light, sign, rider, and bicycle. For example, our method achieves 40.3% on recognising bicycle providing the best performance. For recognising bicycle, the second best result is provided by CLAN [12] which is 30.3%. For recognising the rest semantic classes, our method can provide the results close to Top 3 performance. Please note that CLAN [12] is implemented by automatic mixed precision (AMP) to save graphic memory. To make it comparable, our method is also implemented by AMP.

3) Ablation Study: In the ablation study, we analyse the contributions of various components of our method. Extensive experiments are conducted to figure out their roles in our proposed method. The improvement of mIoU by considering one more components at each stage is presented in Table VI, where Adversarial adaptation is denoted as AA, the co-training framework is denoted as CT, parallel generative ensembles are denoted as PGE, and image translation is denoted as IT. We can see that a poor performance of mIoU will be obtained with only 38.6% if simply training on GTA5 dataset and then evaluating on Cityscapes dataset. If we introduce AA, CT, PGE, and IT into domain adaptation, they will bring the performance gain as 3.0%, 1.6%, 2.2%, and1.5%, respectively. When all of them are adopted in our method, we can achieve the best performance of 46.9%.

TABLE VI
ABLATION STUDY ON THE ADAPTATION BASED RESIDUAL NETWORK BACKBONE

Method	FCN(baseline)	AA	CT	PGE	IT	mIoU
source only	✓					36.6
+AA	✓	✓				41.6
+CT	✓	✓	✓			43.2
+PGE	✓	✓	✓	✓		45.4
+IT	✓	✓	✓	✓	✓	46.9

V. CONCLUSIONS

This paper develops a parallel generative ensembles method to improve the generalisation of semantic segmentation, where a perception model trained on the data generated by a simulator can generalise in real-world scenarios reliably. Due to the high cost of collecting and annotating of real-world data relating to traffic scenes, this study would facilitate the development of autonomous on-road vehicles by creating synthetic data related to traffic scenes. The developed method can translate synthetic data into realistic data so as to bridge the gap between the two domains.

In the proposed method, multiple GAN models are trained on various discrepancy loss and segmentation loss functions under different upsampling strategies for obtaining diverse predictions. Ensemble scheme is utilised on semantic segmentation in unsupervised domain adaptation. The developed method holds the advantages of generative adversarial learning to learn domain-invariant features by minimax game and overcoming the drawback of low segmentation accuracy brought by the imbalanced data. Such a design makes the parallel models complement with each other to achieve a synergy effect on enhancing the performance of segmentation in the target domain. Moreover, the final prediction of each pixel is determined through fusing the predictions from different ensembles, where the ensemble scheme determines the weights of each ensemble based on its upsampling strategy.

The developed method is evaluated by transfer learning tasks on synthetic datasets, GTA5 and SYNTHIA, and realistic dataset Cityscapes. As a result, our proposed method outperforms other competing methods on semantic segmentation in the target domain with regard to overall performance and individual segmentation class performance. We also conduct an ablation study to investigate the contributions of various components in our method. In addition, a comparison of different ensembles is enforced to identify the strength of different ensembles on various semantic classes. To make full use of the advantages of different ensembles, the developed method combines different ensembles to derive final predictions.

Furthermore, within autonomous urban driving, although there have been several studies attempting to apply the emerging deep learning model such as convolutional neural network etc., more studies on optimising the deep learning framework are required for adapting to the real-world applications better. In this study, we have made pioneering efforts on introducing an ensemble scheme to handle the problem of imbalanced data and improve the reliability of unsupervised domain adaptation on semantic segmentation. Since the deep neural networks are "black-box" models, these models are lack of interpretability. In the future, the interpretability of the developed model should be further explored for providing human-understandable explanation on how domain-variant information can be learnt from a simulator so as to generalise in real-world driving scenes reliably.

REFERENCES

[1] W. Yuan, M. Yang, C. Wang, and B. Wang, "Vrdriving: A virtual-to-real autonomous driving framework based on adversarial learning," *IEEE Transactions on Cognitive and Developmental Systems*, 2020.

[2] Y. Liu, X. Zhang, Y. Lin, and H. Wang, "Facial expression recognition via deep action units graph network based on psychological mechanism," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 12, no. 2, pp. 311–322, 2019.

[3] Y. Lu, Y. Chen, D. Zhao, B. Liu, Z. Lai, and J. Chen, "Cnn-g: convolutional neural network combined with graph for image segmentation with theoretical analysis," *IEEE Transactions on Cognitive and Developmental Systems*, 2020.

[4] D. Yi, J. Su, L. Hu, C. Liu, Q. Mohammed, D. Mehrdad, and W.-H. Chen, "Implicit personalization in driving assistance: State-of-the-art and open issues," *IEEE Trans. Intell. Vehicles*, 2020.

[5] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the*

IEEE conference on computer vision and pattern recognition, pp. 3431–3440, 2015.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015.

[7] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[8] S. Motiian, M. Piccirilli, D. A. Adjeroh, and G. Doretto, "Unified deep supervised domain adaptation and generalization," in *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 5715–5725, 2017.

[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 Conf. Comput. Vis. Pattern Recognit.*, pp. 7472–7481, 2018.

[10] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, "Maximum classifier discrepancy for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 3723–3732, 2018.

[11] 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 *International conference on machine learning*, pp. 1989–1998, PMLR, 2018.

[12] 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*, pp. 2507–2516, 2019.

[13] M. Mancini, L. Porzi, S. Rota Bulò, B. Caputo, and E. Ricci, "Boosting domain adaptation by discovering latent domains," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 3771–3780, 2018.

[14] H. Yan, Y. Ding, P. Li, Q. Wang, Y. Xu, and W. Zuo, "Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 2272–2281, 2017.

[15] J. Hoffman, D. Wang, F. Yu, and T. Darrell, "Fcns in the wild: Pixel-level adversarial and constraint-based adaptation," *arXiv preprint arXiv:1612.02649*, 2016.

[16] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International conference on machine learning*, pp. 214–223, 2017.

[17] X. Zhang, Z. Chen, Q. J. Wu, L. Cai, D. Lu, and X. Li, "Fast semantic segmentation for scene perception," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 2, pp. 1183–1192, 2018.

[18] J. Su, D. Yi, B. Su, Z. Mi, C. Liu, X. Hu, X. Xu, L. Guo, and W.-H. Chen, "Aerial visual perception in smart farming: Field study of wheat yellow rust monitoring," *IEEE Transactions on Industrial Informatics*, 2020.

[19] D. Yi, J. Su, C. Liu, and W.-H. Chen, "Trajectory clustering aided personalized driver intention prediction for intelligent vehicles," *IEEE Transactions on Industrial Informatics*, 2019.

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

[21] Y. Zhang, P. David, and B. Gong, "Curriculum domain adaptation for semantic segmentation of urban scenes," in *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 2020–2030, 2017.

[22] Z. Wu, X. Han, Y.-L. Lin, M. Gokhan Uzunbas, T. Goldstein, S. Nam Lim, and L. S. Davis, "Dcan: Dual channel-wise alignment networks for unsupervised scene adaptation," in *Proc. Eur. Conf. Comput. Vis.*, pp. 518–534, 2018.

[23] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine learning*, vol. 79, no. 1-2, pp. 151–175, 2010.

[24] 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.*, pp. 2223–2232, 2017.

[25] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 8789–8797, 2018.

[26] B. Oreshkin, P. R. López, and A. Lacoste, "Tadam: Task dependent adaptive metric for improved few-shot learning," in *Advances in Neural Information Processing Systems*, pp. 721–731, 2018.

[27] M. Hojat and G. Xu, "A visitor's guide to effect sizes–statistical significance versus practical (clinical) importance of research findings," *Advances in health sciences education*, vol. 9, no. 3, pp. 241–249, 2004.

[28] W. Zhu, Y. Huang, H. Tang, Z. Qian, N. Du, W. Fan, and X. Xie, "Anatomynet: Deep 3d squeeze-and-excitation u-nets for fast and fully automated whole-volume anatomical segmentation," *bioRxiv*, p. 392969, 2018.

[29] J. W. Hwang and H. S. Lee, "Adaptive image interpolation based on local gradient features," *IEEE signal processing letters*, vol. 11, no. 3, pp. 359–362, 2004.

[30] Y. Luo, P. Liu, T. Guan, J. Yu, and Y. Yang, "Significance-aware information bottleneck for domain adaptive semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 6778–6787, 2019.

[31] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3234–3243, 2016.

[32] Y. Zhang, P. David, H. Foroosh, and B. Gong, "A curriculum domain adaptation approach to the semantic segmentation of urban scenes," *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[33] 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 transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[34] Q. Wang, J. Gao, and X. Li, "Weakly supervised adversarial domain adaptation for semantic segmentation in urban scenes," *IEEE Transactions on Image Processing*, vol. 28, no. 9, pp. 4376–4386, 2019.

[35] U. Michieli, M. Biasetton, G. Agresti, and P. Zanuttigh, "Adversarial learning and self-teaching techniques for domain adaptation in semantic segmentation," *IEEE Transactions on Intelligent Vehicles*, 2020.

[36] H. Tang, X. Zhu, K. Chen, K. Jia, and C. P. Chen, "Towards uncovering the intrinsic data structures for unsupervised domain adaptation using structurally regularized deep clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[37] M. Biasetton, U. Michieli, G. Agresti, and P. Zanuttigh, "Unsupervised domain adaptation for semantic segmentation of urban scenes," in *Proceedings of the IEEE/CVF Conference on*

Computer Vision and Pattern Recognition Workshops, pp. 0–0, 2019.

[38] W. C. Hung, Y. H. Tsai, Y. T. Liou, Y.-Y. Lin, and M. H. Yang, "Adversarial learning for semi-supervised semantic segmentation," in *29th British Machine Vision Conference, BMVC 2018*, 2019.

[39] R. Li, W. Cao, Q. Jiao, S. Wu, and H.-S. Wong, "Simplified unsupervised image translation for semantic segmentation adaptation," *Pattern Recognition*, vol. 105, p. 107343, 2020.

Dewei Yi (M'18) received the B.Eng. degree in software engineering from the Zhejiang University of Technology, Zhejiang, China, in 2014, the M.Sc. degree from the Department of Computer Science, Loughborough University, Loughborough, U.K., in 2015, and the Ph.D. degree from the Department of Aeronautical and Automotive Engineering, Loughborough University, in 2018. He was a Research Fellow with the Warwick Manufacturing Group (WMG), University of Warwick, U.K., in 2019. He is currently a Lecturer with the School of Natural and Computing Sciences, University of Aberdeen. His current research interests include trustworthy autonomous systems, personalized driving assistance, autonomous vehicles, and vehicular networks.

Hui Fang received the B.S. degree from the University of Science and Technology, Beijing, China, in 2000 and the Ph.D. degree from the University of Bradford, U.K., in 2006. He is currently with the Computer Science Department at Loughborough University. Before, he has carried out research at several world-leading universities, such as University of Oxford and Swansea University. His research interests include computer vision, image/video processing, pattern recognition, machine learning, data mining, scientific visualisation, visual analytics, and artificial intelligence. Recently, he was awarded several grants as PI and co-PI, including Innovate UK funded "An agent-based modelling solution for reliable decision making in crisis and market turmoil in consumer retail", EPSRC funded "RAMP VIS: Making Visual Analytics an Integral Part of the Technological Infrastructure for Combating COVID-19", and NIHR funded "Computer vision to automatically monitor urine output". During his career, he has published more than 70 journal and conference papers.

Yining Hua received the B.Eng. degree in information security from Northeastern University, Shenyang, China, in 2016, the Ph.D. degree from the Department of Computer Science, Loughborough University, Loughborough, U.K., in 2020. She was a Postdoctoral Research Assistant with School of Computer Science, University of Glasgow, U.K., in 2021. She is currently a lecturer in computer science with School of Arts, University of Roehampton, U.K. Her research interests include autonomous systems, computer vision, Internet-of-Things, edge/fog computing and next-generation networks.

Jinya Su (M'16) received his B.Sc. degree in Mathematics from Shandong University, China in 2011 and a Ph.D degree in the Department of Aeronautical and Automotive Engineering, Loughborough University, U.K. in 2016.From 2015, he was a research associate in the same institute. He joined the School of Computer Science and Electronic Engineering, University of Essex, as a lecturer in Computer Science and AI in 2019. His research interests include autonomous systems and applied machine learning, and their real-world applications such as intelligent vehicle and precision agriculture.

Mohammed Quddus is a Professor of Intelligent Transport Systems in the School of Architecture, Building and Civil Engineering at Loughborough University. His main expertise and interests are in the areas of safety, intelligent transport systems (ITS), traffic microsimulation, Artificial Intelligence (AI) and connected and autonomous vehicles (CAVs). Over the last 20 years, his cutting-edge research has led to innovative, influential and transformative outcomes in the forms of new algorithms for in-car navigation and path planning, new safety theories and the development of a new integrated CAV simulation platform. Prof Quddus has published more than 200 papers, 110 in peer-reviewed high-quality journals. Prof Quddus is a Fellow of the UK Higher Education Academy and has been teaching in the School of Architecture, Building and Civil Engineering at Loughborough University since 2006. He has been an Associate Editor of Transportation Research Part C: Emerging Technologies since 2012.

Jungong Han is a Chair Professor and the Director of Research of Computer Science at Aberystwyth University, UK. He also holds an Honorary Professorship at the University of Warwick, UK. His research interests include Computer Vision, Artificial Intelligence and Machine Learning. He has published over 180 articles, including more than 40 IEEE TRANSACTIONS and more than 40 A* conference papers.