Received: 27 November 2022 | Revised: 5 April 2023 | Accepted: 23 May 2023

CAAI Transactions on Intelligence Technology

DOI: 10.1049/cit2.12257

ORIGINAL RESEARCH

Domain-adapted driving scene understanding with uncertainty-aware and diversified generative adversarial networks

Yining Hua[1] | Jie Sui[2] | Hui Fang[3] | Chuan Hu[4] | Dewei Yi[1]

[1]Department of Computing Science, University of Aberdeen, Aberdeen, UK

[2]School of Psychology, University of Aberdeen, Aberdeen, UK

[3]Department of Computer Science, Loughborough University, Loughborough, UK

[4]School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai, China

Correspondence

Dewei Yi.
Email: dewei.yi@abdn.ac.uk

Funding information

Fisheries Innovation & Sustainability; U.K. Department for Environment, Food & Rural Affairs, Grant/Award Numbers: FIS039, FIS045A

Abstract

Autonomous vehicles are required to operate in an uncertain environment. Recent advances in computational intelligence techniques make it possible to understand driving scenes in various environments by using a semantic segmentation neural network, which assigns a class label to each pixel. It requires massive pixel-level labelled data to optimise the network. However, it is challenging to collect sufficient data and labels in the real world. An alternative solution is to obtain synthetic dense pixel-level labelled data from a driving simulator. Although the use of synthetic data is a promising way to alleviate the labelling problem, models trained with virtual data cannot generalise well to realistic data due to the domain shift. To fill this gap, the authors propose a novel uncertainty-aware generative ensemble method. In particular, ensembles are obtained from different optimisation objectives, training iterations, and network initialisation so that they are complementary to each other to produce reliable predictions. Moreover, an uncertainty-aware ensemble scheme is developed to derive fused prediction by considering the uncertainty from ensembles. Such a design can make better use of the strengths of ensembles to enhance adapted segmentation performance. Experimental results demonstrate the effectiveness of our method on three large-scale datasets.

KEYWORDS

adaptive intelligent systems, autonomous vehicles, computer vision, measurement uncertainty, neural network, object segmentation

1 | INTRODUCTION

Uncertainty is unavoidably involved in performing the perceptions of Autonomous vehicles (AVs) [1]. Semantic segmentation (SS) is one of the key environmental perception technologies to help AVs understand driving scenes, which can locate traffic objects accurately [2]. By providing high semantic level understandings of images captured from visual sensors, SS can facilitate the autonomous navigation of vehicles. In recent years, deep learning (DL) based methods [3] have achieved remarkable performance on SS in driving scenes. However, building such models requires a large amount of annotated data. As it is expensive and time-consuming to relabel image data and retrain the model, this issue has become a severe barrier for adapting a model in an unseen scenario.

Domain adaptation [2, 4] provides an intuitive solution to solve the annotation problem. Its goal is to learn a generalised model by exploring the extraction of invariant feature representations [5] or mapping the features [6] between two related domains, where one domain has abundant labelled data (named source domain) and the other domain has no labelled data (named target domain). In the context of driving scene understanding scenario, the source domain contains free annotated synthetic images generated from virtual driving simulations while the target domain has realistic driving scene data with no annotations. Many Deep Domain Adaptation methods [2, 7] are proposed to narrow the data distribution gap between the domains to achieve a generalised model that is trained by using the data from simulators but deployed in the real world. Among these methods, adversarial learning based unsupervised

domain adaptation (UDA) algorithms [7] have demonstrated their potentials as the adversarial learning can provide rich hierarchical representations from feature-level, pixel-level, and output-level to achieve the domain shift.

Despite the success of the UDA algorithms, it is observed that a single UDA model is difficult to align the data distribution of two domains well due to the complex data distribution in the target domain. In particular, various conditions in the driving scenes, such as arbitrary object geometries, illumination variations from different weather conditions and object occlusions, make the single pair of generative and discriminative networks difficult to converge to a well defined holistic mapping function for the segmentation. Naturally, ensemble mechanism is a promising solution to tackle the model generalisation problem. For example, xgboost, an ensemble tree based algorithm, has become one of the most popular machine learning algorithms used in many Kaggle competitions [8]. Recently, ensemble algorithms are also introduced into DL classifiers [9] and generative adversarial networks (GANs) [10] to improve model performance.

In this paper, we propose a novel uncertainty-aware ensemble-based GANs framework, named uncertainty-aware enhanced ensembles of diversified generative adversarial networks (UE^2D-GAN) to train a generalised and reliable model that can be adapted from virtual data to realistic data in the context of driving scene image segmentation. Unlike the traditional ensemble methods, our focus is to increase the diversity of mapping models in our ensemble framework by designing suitable model combinations and uncertainty-aware scheme. In specific, the ensemble models are composed of diversified GAN models trained on different optimisation objectives, training iterations, and network initialisations so as to improve the performance and reliability of domain adaptation. Such a design enables the ensemble model to represent a more holistic mapping by diversifying and combining the set of individual models. Moreover, we consider the uncertainty of predictions made by diversified models. The final SS prediction is determined through the Bayesian information fusion for taking the uncertainty of diversified models into account.

The main contributions of our work can be summarised in threefold. First, we attempt to introduce diversified ensembles of GANs into virtual-to-realistic driving scene understanding. These diversified ensembles are obtained from different optimisation objectives, training iterations, and network initialisations for extracting diverse features. Second, we propose a novel uncertainty-aware ensemble scheme to fuse all diversified ensembles together based on the uncertainty of their SS predictions. Third, various ensemble schemes and fusion strategies are investigated to find out the most promising combination so as to obtain a better synergy effect of ensembles. Furthermore, multi-view perceptual loss is integrated into ensemble scheme so as to minimise the perceptual discrepancy. In addition, a comprehensive comparison is performed to demonstrate the superiority of our proposed method on transferring knowledge from synthetic images to realistic images. A list of adapted segmentation methods are compared on three large-scale datasets, that is, GTA5, SYNTHIA, and Cityscapes.

The rest of this paper is organised as follows. Section 2 introduces the related work. Section 3 provides detailed explanations of our proposed uncertainty-aware ensemble-based GAN method. Section 4 presents experimental results on evaluating open datasets. Section 5 concludes this paper along with future work.

2 | RELATED WORK

SS is a key vehicular technology to help AVs understand driving scenes, which targets on assigning pixel-level labels to an image. In the recent decade, the number of new collected SS datasets increases dramatically due to the reduction of cost for sensors, such as MS COCO Challenge and Pascal VOC2012 challenge which include approximate 200,000 and 10,000 annotated images. In addition, for urban driving scenes understanding, some urban scene segmentation datasets are already publicly available, such as GTA5, SYNTHIA, and Cityscapes [2].

Although the advent of deep convolutional neural networks has shown exceptional performances on SS due to their rich hierarchical features and end-to-end framework [11], to achieve state-of-the-art performance, a segmentation network needs to be trained with an enormous number of dense pixel-level labelled images for extracting efficient features. As pointed in ref. [12], it takes about 90 min to manually annotate all pixels for each image in Cityscapes dataset. Thus, it is not easy to obtain such a large number of labelled data. A widely-accepted alternative solution is to collect annotated images by a simulation, where pixel-wise annotation of an image can be carried out in an automatic manner.

However, due to domain shift of the images from simulator and real world, the model trained on virtual data cannot generalise well in realistic data. Therefore, traditional SS approaches is challenging to address virtual-to-realistic driving scene understanding problem. To address the domain shift issue, domain adaptation is introduced to minimise such a shift. As inspired by the achievement of domain adaptation for image classification, many domain adaptation methods are proposed to tackle SS problem for reducing the human interference. These methods attempt to utilise the virtual data from a simulator, where the annotations can be generated automatically, and then transfer the learnt in-variant knowledge to real-world data. Motivated by the recent achievement of UDA, adversarial learning is treated as a promising solution to align synthetic data and realistic data. Adversarial learning realises domain invariant via a minmax game, where a generative network G is trained to generate target domain style images to fool a discriminative network D and D gives its best effort to distinguish synthetic images and realistic images [2]. Since the domain adaption is introduced into SS problem by [4, 13], where synthetic images are adapted to realistic images through global feature alignment, numerous domain adaptation methods are proposed to learn an adapted segmentation model for generalising better in across domains. These work can be divided into two categories. First one focuses on removing global-level mismatch between source

and target domain so as to achieve output space and spatial-aware adaptations [2]. Second one is to bridge the gap across domains by synthesising target domain image in a pixel-level [4].

Despite GANs have achieved compelling performance on adapted segmentation, it is still challenging to model complex data distribution by a single generative network. As a consequence, a generative network could cause a model collapse problem. Parts of the data distribution can be well-modelled while it leads to failure on describing the whole distribution of the target domain [14]. More specifically, the generative network cannot provide various synthetic data sufficiently, due to the convergence of only one or a few modes of the data distribution which leads that the generative network is not able to provide synthetic data in certain regions of the space [15]. To deal with the problem, co-training framework is adopted due to its robustness and reliability through training alternately on different views with confident labels from the unlabelled data, such as [7]. Inspired by ref. [16–19], we attempt to minimise domain shift by taking the uncertainty of neural networks into account. Different from the previous work in ref. [20], the proposed method fuses different GANs with considering the uncertainty of predictions. Instead of training many GANs from scratch, the proposed method can reduce the number of GANs that need to be trained since models trained from different training iterations can be considered as diversified GANs. In addition, focal loss is also used for optimising the weights of networks so as to increase the diversity of GANs. The proposed method achieves a holistic representation of mapping across domains by introducing uncertainty-aware ensemble scheme meanwhile perceptual discrepancy is minimised by integrating our mathematically formulated multi-view perceptual loss into full objective when training ensembles. We attempt to propose a solution on top of GAN and therefore our proposed method can be transferred to various GANs.

3 | UNCERTAINTY-AWARE ENHANCED ENSEMBLES OF DIVERSIFIED GANs (UE^{2}GAN)

Virtual-to-realistic domain adaptation could be formulated as follows. Let x_s denote as an image from the source domain image set X_S. y_s denote as its corresponding ground truth from source domain annotation set Y_S. A target image is denoted as x_t from target domain image set X_T. Since our objective is to learn a segmentation model from source domain that is adapted in target domain, a generative model G is trained to obtain invariant features so that G could correctly predict pixel-level labels in the target domain. However, it is challenging for a single generative model G to generalise well on all semantic classes across domains.

To address this issue, we design the UE^2D-GAN to bridge the domain gap by learning a holistic representation across domains. There are three distincitve features in our work compared to many other ensemble based models. The first feature is that we enrich the model diversity in our ensemble framework via engaging multi-view perceptual loss and extracting diversified information from different training iterations under various network initialisations. The second feature is that a unified multi-class segmentation loss function is introduced into the ensemble scheme, which can alleviate the problems of imbalanced data by easily tuning two parameters: weighting factor and focusing factor, so that the performance of recognising imbalanced classes and small objects is enhanced during training process. The third feature is our uncertainty-aware fusion strategy for achieving an optimal complement among diversified ensembles.

In this section, we provide detailed explanation of the proposed UE^2D-GAN. In Subsection 3.1, we present the overall architecture of the proposed UE^2D-GAN. In Subsection 3.2, we discuss the components of loss function for ensembles. These components include discrepancy loss, segmentation loss, and adversarial loss. Following this, diversified GANs are discussed in Subsection 3.3. Finally, in Subsection 3.4, the uncertainty-aware ensemble scheme is described and compared to other ensemble schemes to demonstrate its superiority.

3.1 | Architecture of UE^2D-GAN

As illustrated in Figure 1, various ensembles are fused by the uncertainty-aware scheme to provide dense pixel-level predictions of an image. Under the proposed framework, a pair of generator G and a discriminator D is contained in each ensemble. G is a fully-convolutional segmentation network and can be further divided into a feature extractor F and a classifier module H. F extracts features from input images. With the use of the extracted features from F, H predicts labels of each pixel in an image. To make semantic predictions more reliable, multi-view perceptual loss is minimised when training ensembles and then we diversify ensembles by considering GAN models from different optimisation objectives, training iterations, and network initialisation as described in Subsection 3.3. Finally, the final segmentation result is predicted by our proposed uncertainty-aware fusion module.

3.2 | Training objective function

The design of loss function is the core of a DL based segmentation algoirthm. Given a source domain image x_s, its annotation y_s, and a target domain image x_t, feature maps are produced by feature extractor F. Multi-view learners I_1^w and I_2^w utilise the feature maps of images in source domain to generate a semantic prediction map p. The adversarial loss could be calculated by inputting p to the discriminator D and the segmentation loss derived by comparing the pixel-wise prediction p with its corresponding annotation y_s. In addition, x_t is also passed to G for generating p while p is used to obtain the loss of perceptual discrepancy loss, where the difference of multi-view learners and image-level transferring is calculated. Then, it is combined with the adversarial loss for tuning the weight of

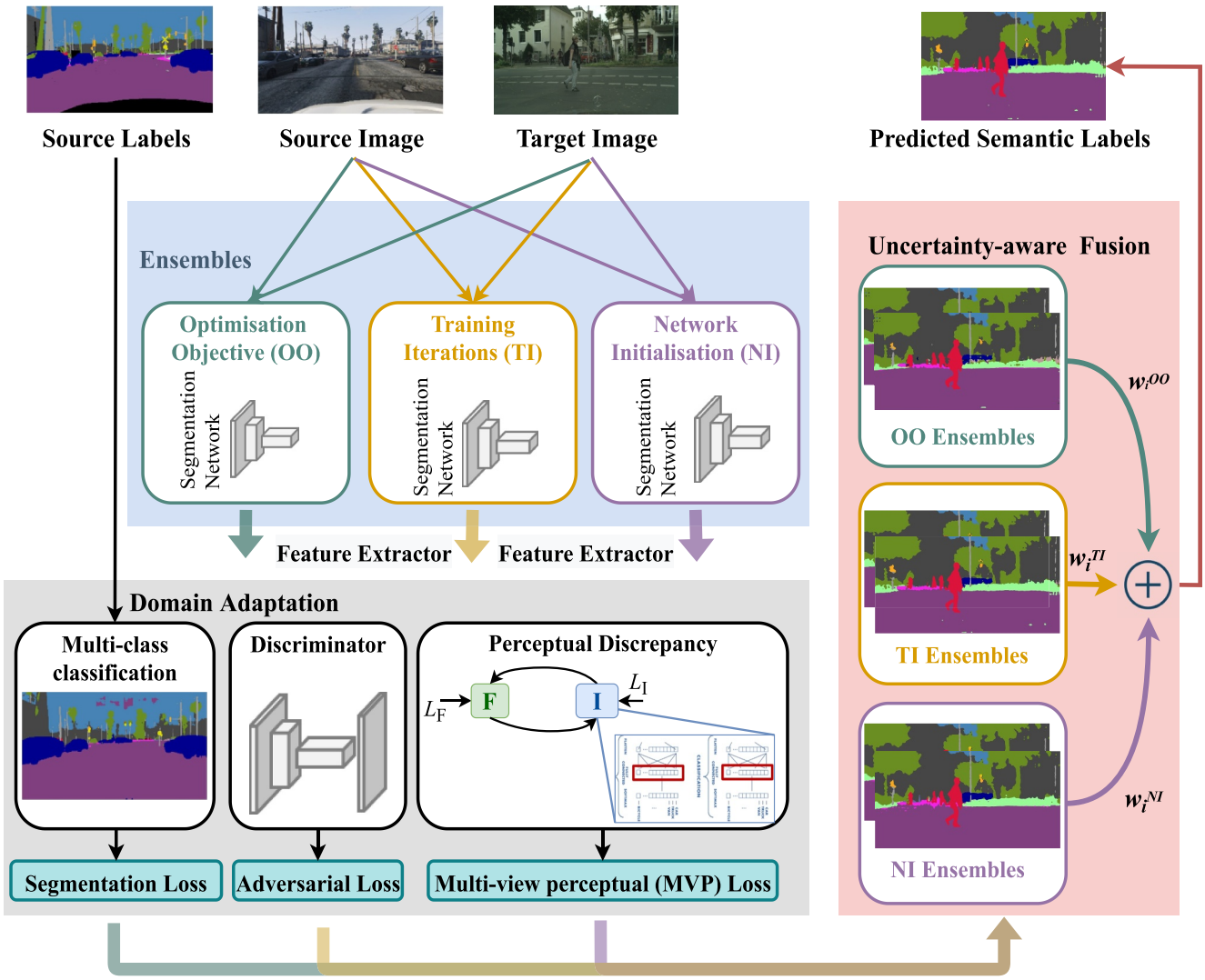

FIGURE 1 The overview of UE^{2}D-GAN. It presents the whole architecture of our uncertainty-aware ensemble scheme, including the multi-classification module (segmentation loss), discriminator module (adversarial loss), perceptual discrepancy module (multi-view perceptual loss), and uncertainty-aware fusion module (where w_i^{OO}, w_i^{TI}, and w_i^{NI} are the classification scores of ith ensemble of different optimisation object, training iterations, or network initialisation), and the options of training objectives for obtaining the ensembling GANs in our method. GANs, generative adversarial network.

each pixel on the segmentation map. Three kinds of losses are integrated into the optimisation objective of our network, which are segmentation loss, adversarial loss, and multi-view perceptual loss.

1) Segmentation loss: Given the height H, width W, and a label map y_s of a source domain image x_s, the shape of original image x_s, and label map are (H, W) and (H, W, C), respectively. C is the number of semantic classes. A unified form is utilised to define multi-class cross-entropy loss and focal loss. Thus, the segmentation loss is computed below.

$$L_{seg} = \sum_{i=1}^{H}\sum_{j=1}^{W}\sum_{c=1}^{C} - g_{ijc}\alpha\left(1-p_{ijc}\right)^{\gamma}\log\left(p_{ijc}\right) \qquad (1)$$

where p_{ijc} is the predicted probability of class c on the pixel with height index i and width index j. g_{ijc} is the ground truth of pixel with height index i and width index j. If a pixel, with height index i and width index j, belongs to class c, then g_{ijc} is given 1. If not, then g_{ijc} is given 0. As mentioned in ref. [21], the multi-class cross entropy is overwhelmed when there is a large imbalance existing. To handle the problem, focal loss is proposed, where another two parameters are α and β. α is a weighting factor ($\alpha \in [0, 1]$), which balances the importance between positives and negatives. γ is a focusing factor and the larger value of γ means the negatives are paid more attention on ($\gamma \in [0, 5]$).

2) Adversarial loss: Adversarial loss term is set to learn domain-invariant features so that the segmentation outputs generated from the source and target domains are not distinguishable between each other so that the domain gap could be narrowed. This adversarial loss is defined as follows:

$$\begin{aligned} L_{adv}(G,D) = &-\lambda_{adv}E[\log(D(G(X_S)))] \\ &-\lambda_{adv}E[\log(1-D(G(X_T)))] \end{aligned} \qquad (2)$$

where X_S and X_T are the images of source and target domains. G and D are the generative network and discriminative network. Moreover, λ_{adv} is the weight factor to control the relative importance of adversarial losses.

3) Multi-view perceptual loss: The multi-view perceptual loss aims to find out the perceptual discrepancy from multi-views, which can be divided into two parts perceptual loss and discrepancy from multi-view learning. The perceptual loss helps align source domain to the target domain, which measures the difference between source data and image translated source data or target data and image translated target data. The loss is used to guide the training process for obtaining an ideal segmentation adaptation model. Multi-view learning is introduced into segmentation network. The multi-view perceptual loss is given in Equation (3).

$$L_{mvp} = \lambda_{pl} E_{X_S} \| I(X_S) - I(G(X_S)) \|_1 + \lambda_{pl}^{inv} E_{X_S} \| I(F(X_S))) - I(X_S) \|_1 + \lambda_{mv} L_{mv}\left(I_1^w, I_2^w\right) \tag{3}$$

where I is segmentation network. G is image-to-image translation network from $X_S \rightarrow X_T$. F is feature extraction network from $X_T \rightarrow X_S$. λ_{pl} and λ_{pl}^{inv} are the weighted factors for constructing and reconstructing paths. Due to the symmetry, the L_{ppl} of X_T and $F(X_T)$ is similar as shown above. To obtain multi-view, I_1^w and I_2^w are two predictors of segmentation network and the discrepancy of multi-views is measured by cosine distance of their weights.

4) Full objective: The full objective is to train a pair of generative network G and discriminative network D by a minimax game until the loss function converges. Therefore, the full objective is given in Equation (4).

$$G^*, D^* = \arg \min_G \max_D \left[L_{seg} + L_{adv}(G, D) + L_{mvp} \right] \tag{4}$$

where L_{seg} is a unified multi-class segmentation loss as defined in Equation (1). L_{mvp} is multi-view perceptual loss. $L_{adv}(G, D)$ is the adversarial loss.

3.3 | Ensembles of diversified GANs

There are various ensembles of GANs in the literature, such as e-GAN, collective e-GAN, loss ce-GAN, se-GAN, and mixture ce-GAN [10]. e-GAN uses ensembles of GAN models in a simple way, where a set of GANs is trained with random initialisations from scratch. Then, one of the GANs is randomly selected to predict target domain labels. Collective e-GAN combines models with different network initialisations to make the prediction. Loss ce-GAN attempts to explore the information from different objective (loss) functions, where GANs are trained with the same network initialisation while under different objective functions. In contrast to e-GAN and collective e-GAN which train a set of GANs from scratch, se-GAN obtains ensemble models from different training iterations. Mixture GAN focuses on exploring information from different objective functions and network initialisation, where the GANs are trained from scratch with a random initialisation of parameters with different objective functions, respectively.

With considering that the diversity of the models is crucial for the success of an adapted ensemble scheme, our proposed ED-GAN inherits advantages from collective e-GAN, se-GAN, ce-GAN, loss ce-GAN, and mixture ce-GAN, which makes full use of the diverse features from various objective functions, training iterations, and network initialisations. To further improve the performance of SS, an uncertainty-aware version of ED-GAN is also proposed, where a pool of models are fused together based on the uncertainty of their outputs so as to generate reliable predictions. The Figure 1 describes overall architecture of the proposed UE^2D-GAN.

3.4 | Uncertainty-aware ensemble scheme

In our ensemble scheme, as each model G generates predictions for all semantic classes, it requires a decision consensus algorithm to fuse predictions from multiple models to reach the final prediction. Such fusion could be conducted in either decision level, for example, majority voting, or output level, for example, weighted average fusion. Majority voting is a decision level fusion strategy, where ensembles predict the semantic class respectively and then final decision is made according to majority selection. Weighted averaging is a common strategy to achieve output level fusion. However, weighted averaging fusion gives the same weight for all ensembles and semantic classes. Such a design neglects the uncertainty of each ensemble for predicting semantic classes.

An uncertainty-aware fusion strategy is proposed and it could fuse the classification scores of different ensembles according to their uncertainty. In this work, we use the outputs of a generative network to derive uncertainty of each semantic class. If the output of a specific semantic class is much higher than the other semantic classes, the classification score of the semantic class makes more contributions to the final prediction. According to the Bayesian theory, given the scores of classifiers x_i, $i = 1, \ldots, C$ and for the SS, the pixel should be predicted as class z_l by maximising the posterior probability as below

$$P(z_l|x_1, \ldots, x_C) = \max_{k=1}^{L} P(z_k|x_1, \ldots, x_C) \tag{5}$$

The Bayesian decision rule Equation (5) states that it is essential to compute the probabilities of the various hypotheses by considering all the classifier scores. This is because such a manner can utilise all the available information correctly to reach a decision. According to the Bayes theorem, we can rewrite the a posteriori probability as follows

$$P(z_k|x_1, \ldots, x_C) = \frac{P(x_1, \ldots, x_C|z_k)p(z_k)}{P(x_1, \ldots, x_C)} \tag{6}$$

where $p(x_1, \ldots, x_C)$ is the joint probability density of unconditional classifier scores. The conditional joint probability distribution of the classifier score is represented by $p(x_1, \ldots, x_C|z_k)$. Following ref. [18], suppose that the classifier scores are conditionally independent with each other by given z_k. As a consequence, we derive the Equation (7)

$$\begin{aligned} P(x_1, \ldots, x_C|z_k) &= \prod_{i=1}^{C} p(x_i|z_k) \\ P(x_1, \ldots, x_C) &= \sum_{k=1}^{L} p(z_k) \prod_{i=1}^{C} p(x_i|z_k) \end{aligned} \tag{7}$$

where $p(x_i|z_k)$ is the model of the ith classifier. Substituting from Equation (7) into Equation (6), we obtain

$$P(z_k|x_1, \ldots, x_C) = \frac{p(z_k) \prod_{i=1}^{C} p(x_i|z_k)}{\sum_{k=1}^{L} p(z_k) \prod_{i=1}^{C} p(x_i|z_k)} \tag{8}$$

Moreover, we can find the following decision rule by using Equation (8) in Equation (5), which quantifies the likelihood of a hypothesis by combining the a posteriori probabilities generated by the individual classifiers via a product rule.

$$p(z_l) \prod_{i=1}^{C} p(x_i|z_l) = \max_{k=1}^{L} p(z_k) \prod_{i=1}^{C} p(x_i|z_k) \tag{9}$$

According to the Bayes theorem, the Equation (9) can be expressed as follows

$$\begin{aligned} p^{-(C-1)}(z_l) \prod_{i=1}^{C} p(z_l|x_i) = \\ \max_{k=1}^{L} p^{-(C-1)}(z_k) \prod_{i=1}^{C} p(z_k|x_i) \end{aligned} \tag{10}$$

Following ref. [22], we assume that a posteriori probability of each classifier does not deviate dramatically from the prior probability, where posteriori probabilities can be given by

$$p(z_k|x_i) = p(z_k)(1 + \epsilon_{ki}) \tag{11}$$

where ϵ_{ki} is far <1. The following equation can be obtained by substituting Equation (11) for the posteriori probabilities in Equation (10)

$$\begin{aligned} p^{-(C-1)}(z_k) \prod_{i=1}^{C} p(z_k|x_i) &= p(z_k) \prod_{i=1}^{C} (1 + \epsilon_{ki}) \\ &= p(z_k) + p(z_k) \sum_{i=1}^{C} \epsilon_{ki} \end{aligned} \tag{12}$$

In our work, we make full use of both probability and raw information obtained from convolutional neural networks, which are received with and without using normalisation, respectively. Finally, we can obtain a sum decision rule given by

$$\begin{aligned} F(z_l) &= \max_{k=1}^{L} p^{-(C-1)}(z_k) \prod_{i=1}^{C} p(z_k|x_i) O_k \\ &= \max_{k=1}^{L} p^{-(C-1)}(z_k) \prod_{i=1}^{C} p(z_k)(1 + \epsilon_{ki}) O_k \\ &= \max_{k=1}^{L} \left[p(z_k) O_k + \sum_{i=1}^{C} p(z_k) \epsilon_{ki} O_k \right] \\ &= \max_{k=1}^{L} \left[(1 - C) p(z_k) O_k + \sum_{i=1}^{C} p(z_k|x_i) O_k \right] \end{aligned} \tag{13}$$

where O_k is the raw output value of the network for ith label and $F(z_l)$ is the final prediction which means that the given pixel belongs to semantic class z_l.

4 | OPEN DATASET EVALUATION

In this section, we evaluate the performance of our proposed method on virtual-to-realistic driving scene understanding. The used synthetic and realistic datasets are described in Section 4.1. Section 4.2 presents the network configuration generator and discriminator along with description of platform. Section 4.3 discusses the metrics used for quantitatively evaluating the performances among different methods. Section 4.4 conducts an ablation study to identify the contributions of different components of our method. Section 4.5 discusses the quantitative and qualitative results on synthetic-to-realistic driving scene understanding, where our proposed method is compared against other advanced methods.

4.1 | Datasets

Three open datasets are evaluated including two synthetic datasets (GTA5 [2], SYNTHIA [2]) and one realistic dataset (Cityscapes [2]). GTA5 and SYNTHIA are the source domain datasets. There are 24,966 vehicle-egocentric images in GTA5 with the resolution 1914 × 1024. SYNTHIA contains 9400 images of the resolution 1290 × 760. These high-resolution synthetic images are produced through a photorealistic open-world computer game called 'Grand Theft Auto V' and annotated with 19 semantic classes for evaluation. Cityscapes is the target domain dataset, which is a real-world dataset with 5000 images of urban scenes in Germany and neighbouring countries. These images are annotated with 19 semantic classes for evaluation as well. The resolution of images within Cityscapes dataset is 1280 × 720 pixels for an image. Both GTA5 and Cityscapes datasets use the same 19 semantic labels in pixel-level so their annotations are compatible with each other. For SYNTHIA, images of 13-class categories are used to assess the performance. For achieving fair comparison, we follow the settings in ref. [2, 7, 23], where generative

networks are trained on all 24,966 images from GTA5 dataset and 9000 images from SYNTHIA dataset, respectively and then their performances are evaluated by using Cityscapes dataset.

4.2 | Implementation details

We implement our method by using PyTorch toolbox and the networks are trained and tested on a single GeForce RTX 2080ti with 11 GB graphic memory. To extract more efficient features, the backbone of source-only generative network G uses the pre-trained ResNet-101 [24], where more efficient features are extracted by residual blocks. For discriminator D, it consists of five convolution layers. Each layer is filtered by a 4-by-4 kernel with the stride size of 2. The channel number from front to back layers are 512, 256, 128, 64, and 1, respectively. It follows the setting of ref. [2, 7, 25] and the parametric ReLU is activation function to concatenate a sequence of convolutional layers. The definition of parametric ReLU activation function is given as follows.

$$f(x) = \begin{cases} x & if \quad x \geq 0 \\ \frac{x}{\alpha} & if \quad x < 0 \end{cases}$$

where the parameter α is a fix positive value. According to ref. [7], it is set as 0.2. The output of last layer is upsampled to the original size of an input image. Stochastic gradient decent is an optimiser used in generative network G. Adam is an optimiser used in discriminative network D, respectively. For SGD, the initial learning rate is 2.5×10^4 with momentum set as 0.9. For Adam, the initial learning rate is 5×10^4 with β_1 and β_2 set as 0.9 and 0.99. The original input image is resized to the resolution of 512×1024 during training for saving the computation load. A prediction map is utilised to derive original size of the image during evaluation for the convenience of assessing the performance of mean intersection of union (mIoU). As clarified in ref. [26], the interpolated surface of bicubic upsampling is smoother with fewer interpolation artefacts and thus bicubic upsampling is used to produce the original size of images in our work. For the ensembles of multi-class cross entropy, α. The decay and maximum epoch are set as 5×10^4 and 1×10^5.

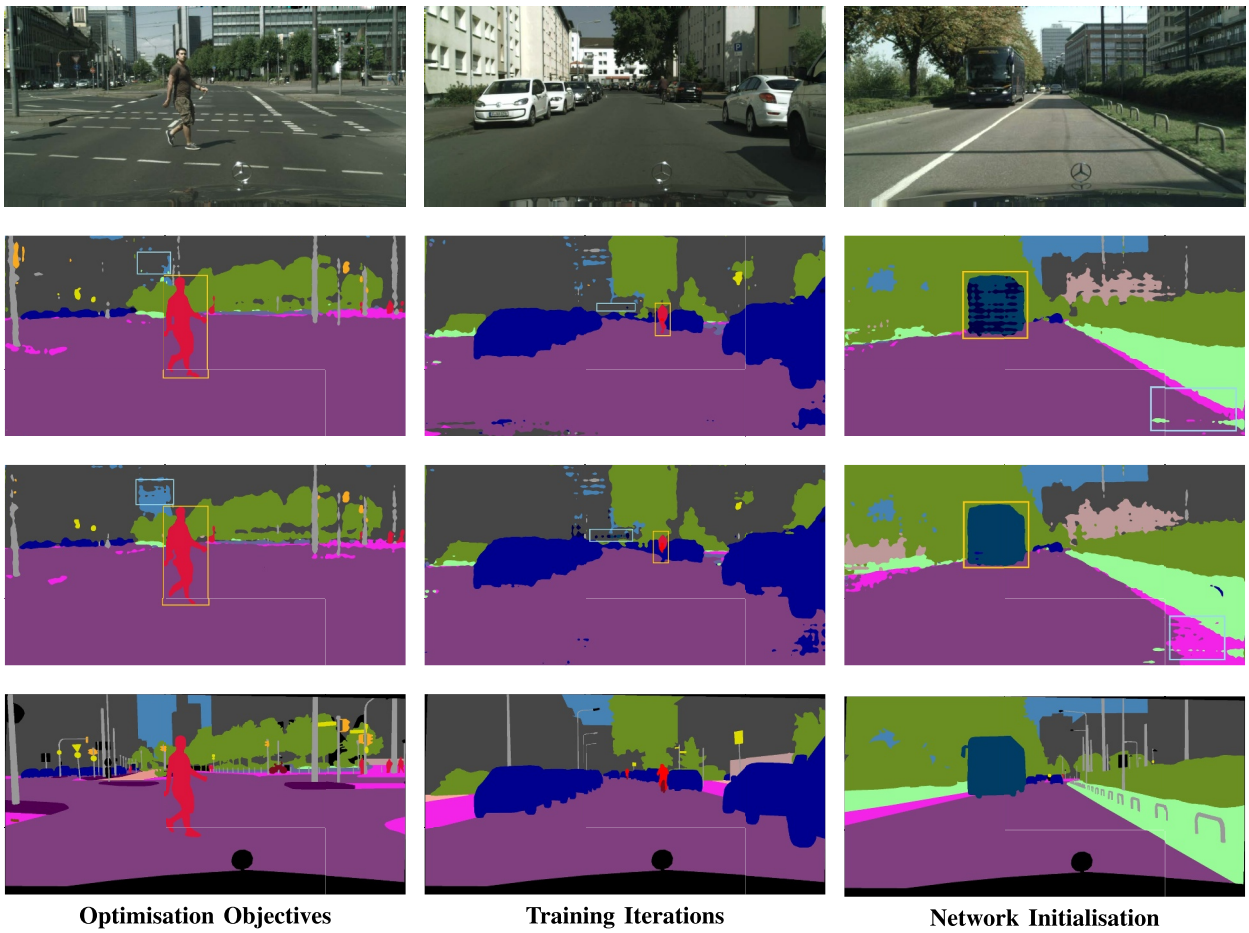

FIGURE 2 Semantic segmentation results of various optimisation objectives, training iterations, and network initialisation: the first row is target images, the second and third rows are prediction results from different ensembles (the first column represents the results of different optimisation objectives; the second column represents the results from different training iterations; the third column represents the results from different network initialisation), and the last row is ground truth annotations. The predictions of different optimisation objectives, training iterations, and network initialisation are highlighted in yellow and blue rectangle boxes as shown each column respectively.

4.3 | Evaluation metrics

Various domain adaptation methods are compared to assess the performance of virtual-to-realistic driving scene understanding. In the experiments, all comparative methods are evaluated on GTA5, SYNTHIA synthetic datasets and Cityscapes realistic dataset. The quantitative results of these methods are evaluated with regards to intersection-over-union (IoU) and mean of IoU, where the former is to assess the performance of each class so as to avoid the effect of class imbalances and the latter is to assess the overall performance of all classes. The IoU is defined as follows.

$$IoU = \frac{TP}{TP + FN + FP}$$

where *TP*, *FN*, and *FP* denote the true positives false negatives, and false positives, respectively.

4.4 | Ablation study

In this section, ablation study is conducted on comparing different ensemble schemes and fusion strategies.

4.4.1 | Comparison of ensemble schemes and configurations

Various optimisation objectives, training iterations, and network initialisation can provide diverse information for SS. Models trained on different optimisation objectives lead to diverse performance on different semantic classes. In our proposed method, we use a unified form to present multi-class cross entropy loss and focal loss to handle problems of imbalanced data and small object detection meanwhile keep the good performance on major classes. The first column of Figure 2 presents the predicted semantic map by using different optimisation objectives. We can see multi-class cross entropy performs better in recognising a person and the focal loss can predict sky better, where the corresponding predictions are highlighted in yellow and blue rectangle boxes respectively. During the training of adversarial learning, the model from a specific training iteration can be efficient to recognise some semantic classes. The prediction results of different training iterations are not same with each other. As shown in the second column of Figure 2, models of different training iterations are more efficient to recognise bus and road respectively, which is highlighted in yellow and blue rectangle boxes. Similar to different training iterations, the diverse information can be also

TABLE 1 Comparison of ensemble schemes (unit %).

Semantic class	eGAN (1 ensemble)	ceGAN (2 ensembles)	Self-GAN (2 ensembles)	Loss ce-GANs (2 ensembles)	Mixture ce-GAN (4 ensembles)	ED-GAN
Road	87.4	88.9	89.0	89.0	89.4	**89.4**
Sidewalk	30.8	30.8	**34.1**	31.3	32.8	32.1
Building	80.7	81.2	80.8	**81.5**	81.3	81.3
Wall	29.5	30.1	29.7	31.0	**33.0**	32.2
Fence	23.9	25.0	**26.1**	26.0	22.0	25.2
Pole	30.3	29.2	29.8	30.1	**30.4**	29.5
Light	34.8	35.1	33.5	**35.0**	34.6	34.7
Sign	23.1	22.7	24.5	23.9	23.6	**24.9**
Vegetation	82.9	83.4	83.7	84.0	83.8	**84.0**
Terrain	31.6	34.8	38.2	37.2	38.7	**39.2**
Sky	75.1	75.6	76.2	76.7	**78.0**	77.7
Person	58.4	60.0	59.2	60.1	59.8	**60.4**
Rider	24.9	28.9	29.0	28.7	28.3	**28.4**
Car	83.7	84.8	84.8	84.5	84.8	**84.7**
Truck	31.3	33.0	34.9	33.9	37.2	**39.7**
Bus	42.7	43.6	43.3	**45.0**	44.3	44.5
Train	**2.1**	1.2	0.9	0.3	0.3	0.0
Motorcycle	26.8	29.3	28.4	27.4	30.2	**30.8**
Bicycle	**29.4**	26.8	26.3	24.2	25.4	25.3
mIoU	43.7	44.4	44.8	45.0	45.1	**45.5**

Note: The bold values represent the highest-performing results achieved by various methods within each semantic class.

obtained from different network initialisation. In the third column of Figure 2, the network initialisation of the second row model is more powerful to classify a road and the network initialisation of the third row model is more powerful to classify a bus when comparing them. With this consideration, a pool of generative models are combined together to make the final predictions. In this section, we compare our proposed GAN-based ensemble scheme with other GAN-based ensemble schemes, including e-GAN, ce-GAN, self-GAN, and mixture ce-GAN with different configurations. In order to guarantee the fairness of comparison, we fuse ensembles in different ensemble schemes by using weighted averaging strategy. To carry out comprehensive comparison, mixture ce-GAN is implemented under different configurations. To make the difference of various ensemble schemes clearer, we obtained the results before considering perceptual loss. The corresponding experimental results of different ensemble schemes and configurations are summarised in Table 1. We can draw the following observations:

- Our proposed ED-GAN outperforms other ensemble schemes by exploring diverse information from different optimisation objectives, training iterations, and network initialisation. More specifically, our proposed ensemble scheme achieves best performance with regard to mIoU, which is 45.5%.
- Compared to different ensemble schemes, our proposed method achieves the best performance of SS, where there are 9 out of 19 semantic classes obtaining the best results. For the rest of semantic classes, the performance of our proposed method is close to most of the best ones.
- If only averaging the different predictions of ED-GAN, the performance of recognising train lead to a decreasing. To tackle the problem, we propose a uncertainty-aware fusion strategy for combining the predictions of different ensembles, so that the performance of recognising train can be significantly improved. The detailed discussion and corresponding results can be found in Section 4-4.4.2.

4.4.2 | Comparison of fusion strategies

In our virtual-to-realistic adaptation problem, we do not have labels from a realistic dataset. This leads supervised based methods are not suitable for our work. Instead of, several unsupervised methods are used to fuse ensembles to derive final predictions. Here, our proposed uncertainty-aware fusion strategy is compared with other fusion strategies, such as majority voting and weighted averaging. To make the comparison fair, we test different fusion strategies. The results of different fusion strategies are presented in Table 2. According to Table 2, two observations can be obtained as follows.

- Taking uncertainty into account for fusing different ensembles, our proposed uncertainty-aware fusion strategy outperforms other fusion strategies, which provides the best performance, 47.9% of mIoU, when perceptual loss is considered along with the discrepancy of multi-views. More specifically, there are 15 out of 19 semantic classes providing the best performance with using uncertainty-aware fusion compared to the other two fusion strategies.
- In general, majority voting fusion underperforms the other two fusion strategies for predicting semantic classes, except for terrain and motorcycle. For majority voting, it provides a sightly better performance when recognising terrain, car, and motorcycle.

TABLE 2 Comparison of fusion strategies (unit %).

Semantic class	Majority voting	Weighted averaging (ED-GAN)	Uncertainty-aware (UE^2D-GAN)
Road	89.1	89.4	**90.9**
Sidewalk	29.8	32.1	**47.1**
Building	80.8	81.3	**84.0**
Wall	31.0	32.2	**32.8**
Fence	23.1	25.2	**25.6**
Pole	28.9	29.5	**32.2**
Light	34.4	34.7	**37.5**
Sign	21.8	24.9	**33.2**
Vegetation	84.0	84.0	**84.2**
Terrain	**41.2**	39.2	38.4
Sky	78.1	77.7	**83.7**
Person	59.8	60.4	**60.4**
Rider	28.5	28.4	**28.6**
Car	**85.0**	84.7	84.1
Truck	39.1	**39.7**	36.4
Bus	43.9	44.5	**46.7**
Train	0.0	0.0	**0.4**
Motorcycle	**31.0**	30.8	25.8
Bicycle	23.6	25.3	**37.6**
mIoU	44.9	45.5	**47.9**

Note: The bold values represent the highest-performing results achieved by various methods within each semantic class.

4.5 | Comparative performance of SS

For assessing the performance of adapted segmentation, our proposed method is compared with other advanced methods. The qualitative results and quantitative comparison of adapted segmentation are presented in Figure 3 and in Tables 3 and 4, respectively. The performance of individual semantic class is assessed by the IoU and the overall performance of all semantic classes is assessed by mIoU various semantic classes, which are provided in Tables 3 and 4. Here, we can draw the three observations:

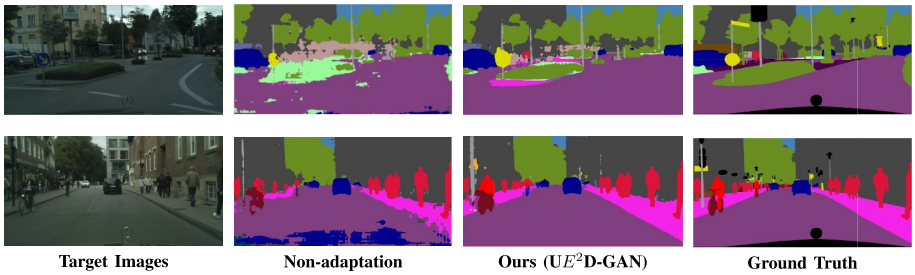

FIGURE 3 Qualitative semantic segmentation results on adaptation from GTA5 to Cityscapes. From left to right: target image, non-adapted results (source only), adapted results with UE²D-GAN, and the ground truth annotations, respectively.

TABLE 3 Quantitative comparison results from GTA5 to cityscapes (unit %).

Semantic class	FCNs wild [4]	MCD [27]	CDA [28]	CyCADA [29]	CBST [30]	DCAN [31]	AdaSegNet [2]	CLAN [32]	Ours
Road	70.4	86.4	74.9	85.2	90.4	82.3	86.5	87.0	**90.9**
Sidewalk	32.4	8.5	22.0	37.2	**50.8**	26.7	36.0	27.1	47.1
Building	62.1	76.1	71.7	76.5	72.0	77.4	79.9	79.6	**84.0**
Wall	14.9	18.6	6.0	21.8	18.3	23.7	23.4	27.3	**32.8**
Fence	5.4	9.7	11.9	15.0	9.5	20.5	23.3	23.3	**25.6**
Pole	10.9	14.9	8.4	23.8	27.2	20.4	23.9	28.3	**32.2**
Light	14.2	7.8	16.3	22.9	28.6	30.3	35.2	35.5	**37.5**
Sign	2.7	0.6	11.1	21.5	14.1	15.9	14.8	24.2	**33.2**
Vegetation	79.2	82.8	75.7	80.5	82.4	80.9	83.4	83.6	**84.2**
Terrain	21.3	32.7	13.3	31.3	25.1	25.4	33.3	27.4	**38.4**
Sky	64.6	71.4	66.5	60.7	70.8	69.5	75.6	74.2	**83.7**
Person	44.1	25.2	38.0	50.5	42.6	52.6	58.5	58.6	**60.4**
Rider	4.2	1.1	9.3	9.0	14.5	11.1	27.6	28.0	**28.6**
Car	70.4	76.3	55.2	76.9	76.9	79.6	73.7	76.2	**84.1**
Truck	8.0	16.1	18.8	17.1	5.9	24.9	32.5	33.1	**36.4**
Bus	7.3	17.1	18.9	28.2	12.5	21.2	35.4	36.7	**46.7**
Train	0.0	1.4	0.0	4.5	1.2	1.3	3.9	**6.7**	0.4
Motorcycle	3.5	0.2	16.8	9.8	14.0	17.0	30.1	**31.9**	25.8
Bicycle	0.0	0.0	14.6	0.0	28.6	6.7	28.1	31.4	**37.6**
mIoU	27.1	28.8	28.9	35.4	36.1	36.2	42.4	43.2	47.9

Note: The bold values represent the highest-performing results achieved by various methods within each semantic class. The value highlighted in blue represents the highest overall performance (mIoU) across all semantic classes.

- With the help of introducing uncertainty-aware fusion into ensemble scheme, our proposed UE²D-GAN method does not only enhance the segmentation results, but also improve the generalisation. Consequently, our proposed method outperforms other advanced methods with regard to mIoU, which are able to achieve 47.9% of mIoU on GTA5 to Cityscapes (Table 3) and 47.3% of mIoU on SYNTHIA to Cityscapes adaptation (Table 4), respectively.
- Compared to different ensemble schemes, our proposed method achieves the best performance of SS for GTA5 to Cityscapes adaptation, where there are 16 out of 19 semantic classes obtaining the best results. For the rest of semantic classes, the performance of our proposed method is very close to the best ones. Our proposed method also achieves the best performance for SYNTHIA to Cityscapes adaption, especially for the classes of sidewalk, light, sign, and bicycle.
- The proposed uncertainty-aware fusion strategy can efficiently capture uncertain information to achieve better segmentation predictions. For instance, our proposed fusion strategy can significantly outperform the other fusion strategies when recognising the semantic class of train as shown in Table 2.

TABLE 4 Quantitative comparison results from SYNTHIA to cityscapes (unit %).

Semantic class	SemanticDA [33]	AdvSemiSeg [34]	SUIT [35]	IBAN [23]	AdaSegNet (feat. Only) [2]	ALST [13]	AdvEnt [25]	CLAN [32]	Ours
Road	78.4	72.5	75.1	78.2	62.4	**80.7**	76.6	78.0	73.9
Sidewalk	0.1	0.0	31.4	19.7	21.9	0.3	28.3	34.1	**30.2**
Building	73.2	63.8	77.4	**80.5**	76.3	0.3	79.1	78.1	76.6
Light	0.0	0.0	11.7	9.4	11.7	0.0	5.8	8.8	**23.8**
Sign	0.2	0.5	15.0	8.9	11.4	0.4	9.6	13.4	**24.0**
Vegetation	84.3	**84.7**	79.2	77.4	75.3	84.0	78.9	78.1	78.1
Sky	78.8	76.9	77.4	82.0	80.9	79.4	**84.3**	81.5	78.6
Person	46.0	45.3	54.2	**56.3**	53.7	46.6	52.8	55.3	38.8
Rider	0.3	1.5	18.1	9.6	18.5	0.8	**27.6**	20.6	21.0
Car	74.9	77.6	78.1	76.3	59.7	**80.3**	60.3	66.4	73.4
Bus	30.8	31.3	27.4	22.8	13.7	**32.8**	24.1	22.3	30.0
Motorcycle	0.0	0.0	9.4	17.5	**20.6**	0.5	13.4	12.4	17.3
Bicycle	0.1	0.1	30.2	23.3	24.0	0.5	30.2	31.5	**49.2**
mIoU	35.7	34.9	45.0	43.2	40.8	37.0	43.4	44.7	47.3

Note: The bold values represent the highest-performing results achieved by various methods within each semantic class. The value highlighted in blue represents the highest overall performance (mIoU) across all semantic classes.

5 | CONCLUSIONS

This paper proposes an uncertainty-aware diversified ensemble method to narrow the gap between synthetic data and realistic data for solving the problem of virtual-to-realistic driving scene understanding. In the proposed method, we explore the strengths of different optimisation objectives, training iterations, and network initialisation and their strengths are made fully use through an ensemble scheme. Moreover, an uncertainty-aware fusion strategy is developed to integrated diversified ensembles together based on the uncertainty of predictions. Subsequently, the final pixel-level prediction can be provided a SS map for each image. Such a design fully takes advantage of generative ensembles so as to improve the performance of SS in the target domain. To evaluate our method, SS models are trained on synthetic driving dataset and test on realistic driving data. Experimental results demonstrate that our method can learn in-variant features more efficiently so that the knowledge can be transferred from synthetic datasets, GTA5 and SYNTHIA, to realistic dataset, Cityscapes. Therefore, our proposed method outperforms other advanced methods on adapted segmentation results of the target domain. For now, most of GANs need to be trained for a large number iterations to provide promising results. In future, we will attempt to reduce the training workload of GANs by simplifying the architecture of deep neural networks.

ACKNOWLEDGEMENTS

This work was supported by Fisheries Innovation & Sustainability (FIS) and the U.K. Department for Environment, Food & Rural Affairs (DEFRA) under grant number FIS039 and FIS045A.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in the repositories and URLs below. The GTA5 Dataset at https://download.visinf.tu-darmstadt.de/data/from_games/. The SYNTHIA Dataset at http://synthia-dataset.net/downloads/. The Cityscapes Dataset at https://www.cityscapes-dataset.com/.

ORCID

Hui Fang https://orcid.org/0000-0001-9365-7420
Dewei Yi https://orcid.org/0000-0003-1702-9136

REFERENCES

1. Ma, Y., et al.: Artificial intelligence applications in the development of autonomous vehicles: a survey. IEEE/CAA J. Autom. Sin. 7(2), 315–329 (2020). https://doi.org/10.1109/jas.2020.1003021
2. Tsai, Y.-H., et al.: Learning to adapt structured output space for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7472–7481 (2018)
3. Kamal, U., et al.: Automatic traffic sign detection and recognition using SegU-Net and a modified Tversky loss function with L1-constraint. IEEE Trans. Intell. Transport. Syst. 21(4), 1467–1479 (2019). https://doi.org/10.1109/tits.2019.2911727
4. Hoffman, J., et al.: FCNs in the wild: pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649* (2016)
5. Bousmalis, K., et al.: Unsupervised pixel-level domain adaptation with generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3722–3731 (2017)
6. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4401–4410 (2019)

7. Luo, Y., et al.: 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)
8. Chen, T., Guestrin, C.: Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016)
9. Strodthoff, N., et al.: Deep learning for ECG analysis: benchmarks and insights from PTB-XL. IEEE J. Biomed. Health Inf. 25(5), 1519–1528 (2020). https://doi.org/10.1109/jbhi.2020.3022989
10. Wang, Y., Zhang, L., van de Weijer, J.: Ensembles of generative adversarial networks. *arXiv preprint arXiv:1612.00991* (2016)
11. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
12. Zhang, Y., et al.: A Curriculum Domain Adaptation Approach to the Semantic Segmentation of Urban Scenes. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020)
13. Michieli, U., et al.: Adversarial learning and self-teaching techniques for domain adaptation in semantic segmentation. IEEE Trans. Intell. Veh. 5(3), 508–518 (2020). https://doi.org/10.1109/tiv.2020.2980671
14. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: International Conference on Machine Learning, pp. 214–223 (2017)
15. Tolstikhin, I.O., et al.: AdaGAN: boosting generative models. In: Advances in Neural Information Processing Systems, pp. 5424–5433 (2017)
16. Kendall, A., Gal, Y.: What uncertainties do we need in Bayesian deep learning for computer vision? Adv. Neural Inf. Process. Syst. 30 (2017)
17. Wang, J., et al.: Synchronization of generally uncertain Markovian inertial neural networks with random connection weight strengths and image encryption application. IEEE Transact. Neural Networks Learn. Syst., 1–15 (2021). https://doi.org/10.1109/tnnls.2021.3131512
18. Kittler, J., et al.: On combining classifiers. IEEE Trans. Pattern Anal. Mach. Intell. 20(3), 226–239 (1998). https://doi.org/10.1109/34.667881
19. Wang, J., et al.: Synchronization criteria of delayed inertial neural networks with generally Markovian jumping. Neural Network. 139, 64–76 (2021). https://doi.org/10.1016/j.neunet.2021.02.004
20. Yi, D., et al.: Improving Synthetic to Realistic Semantic Segmentation with Parallel Generative Ensembles for Autonomous Urban Driving. IEEE Transactions on Cognitive and Developmental Systems (2021)
21. Lin, T.-Y., et al.: Focal loss for dense object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2980–2988 (2017)
22. Ma, A.J., Yuen, P.C., Lai, J.-H.: Linear dependency modeling for classifier fusion and feature combination. IEEE Trans. Pattern Anal. Mach. Intell. 35(5), 1135–1148 (2012). https://doi.org/10.1109/tpami.2012.198
23. Luo, Y., et al.: Significance-aware information bottleneck for domain adaptive semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6778–6787 (2019)
24. Luan, S., et al.: Gabor convolutional networks. IEEE Trans. Image Process. 27(9), 4357–4366 (2018). https://doi.org/10.1109/tip.2018.2835143
25. Ganin, Y., et al.: Domain-adversarial training of neural networks. J. Mach. Learn. Res. 17(1), 2096–2030 (2016)
26. Dodgson, N.A.: Quadratic interpolation for image resampling. IEEE Trans. Image Process. 6(9), 1322–1326 (1997). https://doi.org/10.1109/83.623195
27. Saito, K., et al.: Maximum classifier discrepancy for unsupervised domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3723–3732 (2018)
28. Zhang, Y., David, P., Gong, B.: Curriculum domain adaptation for semantic segmentation of urban scenes. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2020–2030 (2017)
29. Cycada: Cycle consistent adversarial domain adaptation. In: International Conference on Machine Learning (ICML) (2018)
30. Zou, Y., et al.: Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In: Proceedings of the European Conference on Computer Vision, pp. 289–305 (2018)
31. Wu, Z., et al.: DCAN: dual channel-wise alignment networks for unsupervised scene adaptation. In: Proceedings of the European Conference on Computer Vision, pp. 518–534 (2018)
32. Tang, H., et al.: Towards Uncovering the Intrinsic Data Structures for Unsupervised Domain Adaptation Using Structurally Regularized Deep Clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)
33. Biasetton, M., et al.: Unsupervised domain adaptation for semantic segmentation of urban scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 0 (2019)
34. Hung, W.-C., et al.: Adversarial learning for semi-supervised semantic segmentation. *arXiv preprint arXiv:1802.07934*. (2018)
35. Li, R., et al.: Simplified unsupervised image translation for semantic segmentation adaptation. Pattern Recogn. 105, 107343 (2020). https://doi.org/10.1016/j.patcog.2020.107343

How to cite this article: Hua, Y., et al.: Domain-adapted driving scene understanding with uncertainty-aware and diversified generative adversarial networks. CAAI Trans. Intell. Technol. 1–12 (2023). https://doi.org/10.1049/cit2.12257