PcGAN: A Noise Robust Conditional Generative Adversarial Network for One Shot Learning

Lizhen Deng, *Member, IEEE,* Chunming He, Guoxia Xu, *Member, IEEE,* Hu Zhu, *Member, IEEE,* and Hao Wang *Member, IEEE*

***Abstract*—Traffic sign classification plays a vital role in autonomous vehicles for its powerful capability in information representation. However, the low-quality data of traffic signs captured by in-vehicle cameras often inevitably bring inherent challenges to the one-shot classification task. Apart from the problem of data degradation, learning-based classification techniques of real traffic signs also come across the challenges of intra-class and inter-class data imbalance from the training data. To overcome the aforementioned problems, we propose an end-to-end degradation robust deep model, termed PcGAN, to classify traffic signs in a manner of few-shot learning. The proposed PcGAN models the joint distribution between the degraded traffic signal data and the corresponding prototypes from both degradation removal and generation perspectives by two alternating optimized modules, which ensures the generalization of the learned embedding of latent space for novel tasks. A multi-task loss function is designed to improve the robustness of PcGAN. Numerous experiments comprehensively demonstrate that the accuracy of our proposed PcGAN is improved by 5% compared with other state-of-the-art (SOTA) approaches in few-shot classification.**

***Index Terms*—One-shot Learning, Generative Adversarial Network, Prototypical Data, Traffic Signs, Intelligence Transportation Systems.**

I. Introduction

Traffic signs are the road facilities that convey guidance, restriction, warning, or instruction information in the form of words or symbols, which are significant for traffic driving. Compared to the accessible word form, the symbol form tends to be isolated from any specific language and can only be mastered by those familiar with the prior conventions, e.g., shape similarity. This is a tremendous challenge for drivers in the case of the considerable number of symbol-based traffic signs [1]. Fortunately, with the rapid development of intelligent transportation systems (ITS) [2], traffic signal techniques have been widely applied in autonomous vehicles [3]–[5], which can both provide accurate judgment for various traffic signs to human drivers and adaptively correct their autonomous driving behavior.

This work is supported by the National Natural Science Foundation of China under Grant 62072256. *(Lizhen Deng and Chunming He contributed equally. Corresponding author: Hu Zhu, E-mail: peter.hu.zhu@gmail.com)*

Lizhen Deng is with National Engineering Research Center of Communication and Network Technology, Nanjing University of Posts and Telecommunications, Nanjing, 210003, China. Chunming He is with Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China. Guoxia Xu and Hao Wang are with Department of Computer Science, Norwegian University of Science and Technology, 2815 Gjovik, Norway. Hu Zhu is with Jiangsu Province Key Lab on Image Processing and Image Communication, Nanjing University of Posts and Telecommunications, Nanjing 210003, China.

In order to achieve the optimal environment sensing quality, [6] presented a novel framework combining fractal constraint with group sparsity. Different from the multi-modality-based data fusion strategies, traffic sign techniques can only be implemented by a single sensor, i.e., a visible camera, for its original design intention being friendly to the human visual system [7], which is no wonder sensitive to the surrounding environments, variable illumination conditions, complex weathers [8], [9]. Furthermore, the camera imaging system can bring some degradation, such as the hardware-induced noises and the in-camera pre-processing distortion [10].To overcome the aforementioned challenges, many noise-robust approaches are proposed both in traditional and learning-based methods. Traditional methods mainly rely on the handcrafted feature operators to suppress degradation, e.g., local entropy [11], local gradient constraint [12]. [13] proposed a novel target-aware method based on a non-local low-rank model and saliency filter regularization to suppress the noise from the background and enable joint target saliency learning in a lower dimensional discriminative manifold. Although these methods suppress the degradation to a certain extent, the problem is that the design of the feature operator must not only ensure that it is easy to implement but also have a certain degree of adaptability to the degradation, which is not easy. In deep learning-based methods, Tian et al. [14] adopt a recurrent attention mechanism to attenuate the effect of background noise in traffic signs. However, the existing target noise can also influence the accuracy of subsequent processing, and the attention map can be interfered with by the data degradation. A deep learning-based framework for robust traffic sign detection (DFR-TSD) [15] exploited a challenge classifier to classify the degradation condition of the input data, e.g., lens blur, snow, haze, etc., which effectively delivered targeted reconstruction techniques for different types of degradation. Nevertheless, this work is time-consuming and is highly dependent on the precision of the challenge classification, which consistently fails to present a high-quality reconstruction in case of misclassification.

Apart from the image degradation from real captured traffic signs, the recovered high-quality data of traffic signs can have a significant visual gap with the prototypical data. Because the prototypical data is the most standard data in the corresponding class and is not explicitly designed for the input, which can cause visual domain discrepancy between the input data and the prototypical data. For this challenge, some works like Temel et al. [16] were dedicated to publishing a novel dataset, which can alleviate the domain discrepancy to some extent. [17] proposed a bilateral weighted regression ranking model

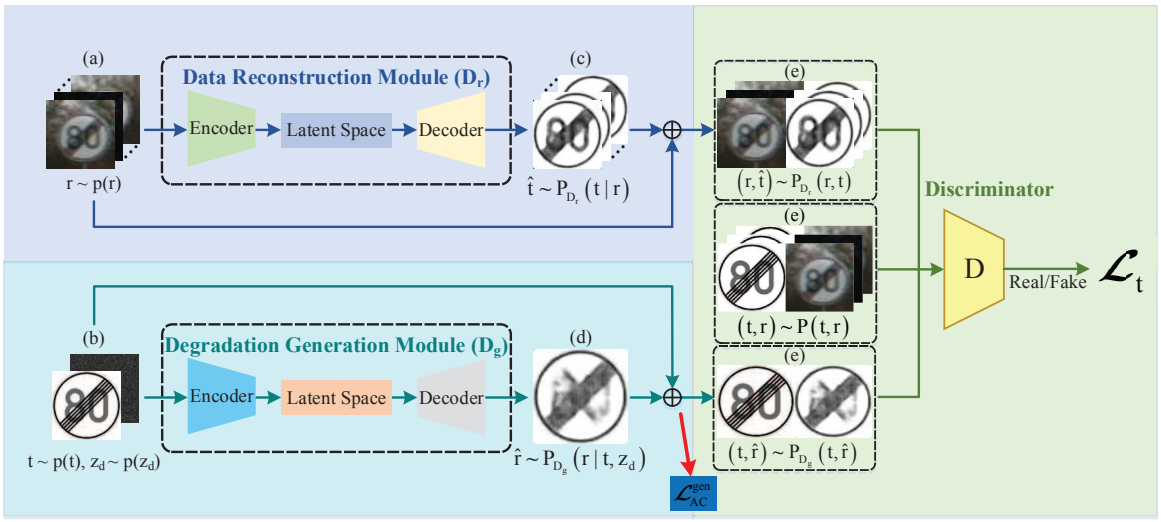


Fig. 1. The framework of the proposed PcGAN, where (a) is the real data of traffic sign in the same class, (b) is the corresponding prototype and the initial degradation operator, (c) is the reconstructed data processed by data reconstruction module, (d) is degraded by degradation generation module, and (e) is the concat for ease of the approximation of joint distribution $p(t, r)$.

termed BWRR to solve the loss from data fidelity term. The recently published public datasets [16], [18], [19] inevitably have problems with intra-class and inter-class data imbalance. Therefore, the pure data-driven deep network often comes across issues caused by data imbalance.

To overcome the challenges mentioned above, we propose a degradation robust conditional generative adversarial network with prototypical data, termed PcGAN, to classify traffic signs by few-shot learning. As shown in Fig. 1, the proposed PcGAN is an end-to-end model to simulate the degradation process, which formulas the implicit joint distribution between real traffic sign data and its corresponding prototypes instead of forcing the reconstructed traffic signs to be similar to their prototypes by two alternating update modules, i.e., data reconstruction module and degradation generation module. To further improve the robustness of PcGAN, a multi-task loss function is proposed to joint constrain the degradation removal and generation. The contributions of PcGAN can be summarized as follows:

- To the best of our knowledge, it is the first time to introduce the dual adversarial strategy to approximate the implicitly joint distribution in the field of one-shot learning of traffic sign classification, which can simultaneously simulate the process of degradation removal and generation. Therefore, the proposed PcGAN can comprehensively learn the latent relationship between real traffic signs and their prototypical data.
- To achieve a more accurate classification performance, we propose a multi-task loss function to joint constrains the network training, including the reconstruction of real data of traffic signs and the degradation of the corresponding prototypes.
- Numerous experiments comprehensively illustrate the superiority of our PcGAN in comparison with other state-of-the-art (SOTA) techniques in both qualitative and quantitative analysis.

The main structure of the paper is as follows. Section I is the Introduction of this paper. Section II introduces the related works. Section III introduces the framework of PcGAN. Then Section IV introduces the related experimental results. Finally, Section V provides a brief conclusion of our work.

II. Related Work

Few-shot learning, which aims to recognize new classes by adapting the learned knowledge with extremely limited few-shot (support) examples, remains a significant open problem in computer vision. It learns patterns with a set of data (base classes) and adapts to a disjoint set (new classes) with limited training data. Data imbalance can be effectively solved by few-shot learning for its powerful generalization capability with a few prototypical data of novel classes, by which the generic prior knowledge of data can be learned in the latent space. In [20], the existing few-shot learning methods were divided into two categories which are data-augmentation-based methods and prior-knowledge-based techniques according to the principles of whether the number of available labeled samples for the target classes increases is increased. The former uses transformation operations, simulation, or deep generative models to generate samples without actually collecting new data, which can improve the generalization ability of the model and suppress the risk of overfitting. While the latter mainly focuses on learning with limited labeled data, which means making full use of prior knowledge and experience to guide the learning progress of new tasks. The pioneering working of few-shot learning, i.e., Li et al. [21], assumed that prior knowledge facilitates people with more efficient learning. Li et al. [21] explored the latent and generic prior information with a bayesian strategy. The results demonstrate that the learned prior can be easily adjusted to other problems with small data. This solves the problem of data imbalance to a particular extent and shows a certain generalization ability. In [22], a low-rank representation of samples in the feature space is exploited by the label distribution learning for the classification task. Xing et al. [23]–[25] had proposed to combine the auto-

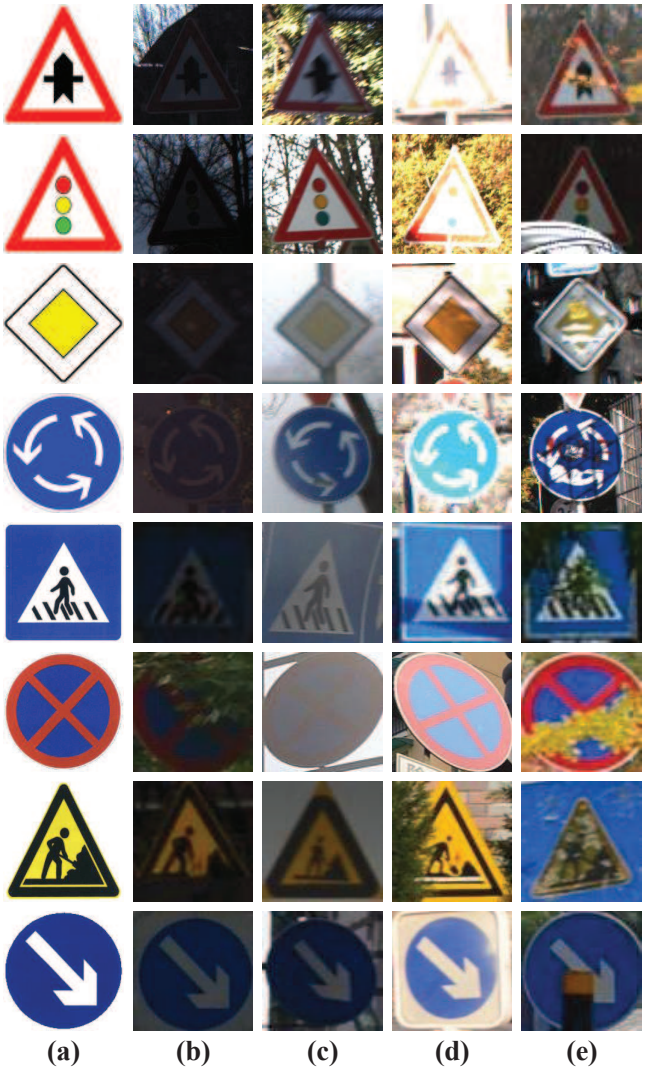


Fig. 2. The examples of prototypes and their real traffic signal data from GTSRB and TT100K datasets, where (a)-(e) correspond to prototypes, over-darkness, blur, overexposure, occlusion, respectively.

encoder with the generative adversarial network for the zero-shot cross-modal retrieval task, which leveraged the shared latent space learning, knowledge transfer, and feature synthesis within the distribution alignment.

In addition, few-shot image classification is the one with the most focus and research. To tackle this problem, two actions can be taken as follows. One is optimization-based methods [26], [27], which firstly train a network with base class data, then fine-tune the classifier or the whole network with support data from unseen classes. On the other hand, few-shot learning is solved by applying an existing or learned metric to the extracted features of images in the metric-based method. MatchingNet [28] adopts a memory module to merge the information in each task and cosine distance as the metric to classify unseen data. ProtoNet [29] proposes the prototype as a simple representation of each category and adopts euclidean distance as the metric. In [30], a dual stream neural network is proposed to reconstruct the original infrared spectroscopy, which effectively strengthens the capability to represent the feature of the infrared spectrum.

Moreover, some of the recent approaches, e.g., Lake et al. [31] paid attention to the procedure of the generation, which also explored a few examples with hierarchical Bayesian. Under this strategy, the extracted procedures can also be generalized to new tasks, even though the number of examples drops to one. It is worth noting that few-shot learning is a continuously challenging task regardless of the rapid development of deep learning-based algorithms, which tend to be lower than some handcrafted methods. The reasons mainly lie in that the quite limited data can result in the problem of over-fitting. In this case, recent learning-based algorithms mainly focus on embedding learning and meta-learning strategies. The variational prototyping-encoder(VPE) [32] is a one-shot learning-based approach with the former strategy, which classified traffic signs by nearest neighbor classification in a variational auto-encoder (VAE) structure and achieved an advanced performance in both data classification and image retrieval. However, VPE forced the generated output of real traffic signs to be similar to their prototypes, which are the most standard traffic signs. It seems unreasonable to learn common space directly from a mapping relationship under several challenging cases, which are shown in Fig. 2. Moreover, this method seems to be very challenging to implement.

III. PcGAN Framework

In this section, we introduce the proposed PcGAN model, which is analogous to metric-based learning solutions with the purpose of learning a generalizable embedding. Unlike traditional generative adversarial network (GAN), our model is applied in a few-shot learning task which focuses on the generalization capability of the model and the embedding versatility of the latent space with only one support data. Therefore, the training phase of PcGAN aims to construct a generic embedding space with a large amount of training data, and the learned embedding space will be utilized for nearest neighbors classification between the tests with new classes and their prototypical data in the testing phase. Furthermore, compared with the metric-based learning approaches that construct embedding space with a selected metric, the embedding construction process of PcGAN is assisted with a meta task, i.e., learning a mapping from real data to prototypical data, which means that our PcGAN can achieve more valid prior than the manually selected metric. In this paper, the proposed PcGAN is applied to handle the few-shot classification problem by single prototypical data, whose framework is presented in Fig. 1 with the following modules, including data reconstruction module D_r, degradation generation module D_g, and discriminative module D.

In the process of PcGAN network model design, first of all, we respectively carry out relevant mathematical derivations on the joint distribution formulas of the Data Reconstruction Module and Degradation Generation Module in the network structure. Then, based on the above derivation, we conduct relevant analysis on the design of the loss function for

network optimization. In addition, we introduce the network architecture of the data reconstruction module and degradation generation module in detail. Finally, we train and test our PcGAN network.

A. Problem Formulation and our proposed PcGAN

We make a definition of the one-shot classification problem as a conditional GAN with two generators, which corresponds to the encoder-decoder structure in the data reconstruction module and noise generation module. Given an aligned pair of real degradation image r captured by a vehicle camera and its prototypical image t, the problem of our task is formulated by solving the joint distribution $p(r,t)$ instead of forcing the generator to learn a mapping from r to t [32]. In the following, we describe the principles of the aforementioned modules and the discriminative module in detail.

Data Reconstruction Module. The data reconstruction module D_r focuses on reconstruct a recovered image in the case where the degradation image r is known, i.e., learning a implicit distribution $p_{D_r}(t \mid r)$ to approximate $p(t \mid r)$, where the approximated recover data $\hat{t} = D_r(r)$. In this case, the recovered joint distribution is defined as follows:

$$p_{D_r}(t,r) = p_{D_r}(t \mid r)p(r), \tag{1}$$

where $p(r)$ is the distribution of the captured degradation image r, which is a fixed value. From the equation mentioned above, it is evident that the performance of the data reconstruction module $D_(r)$ is proportional to the degree of approximation between $p_{D_r}(t,r)$ and $p(t,r)$.

Degradation Generation Module. Traffic signs can only be captured by camera sensor in autonomous technique, which can bring a few hardware-induced noises and some in-camera pre-processing distortion [10]. In this case, we import a latent vector z_d to denote the aforementioned degradable condition, where the distribution of the real degradation process from prototypical data t to real distorted data r can be represented by $p(r \mid t, z_d)$ and that generated by degradation generation module D_g can be depicted by $p_{D_g}(r \mid t, z_d)$. Therefore, the degraded data $\hat{r}$ follows:

$$\hat{r} = D_g\left(t, z_d\right) \sim p_{D_g}(r \mid x, z). \tag{2}$$

Then the degraded joint distribution can be achieved by:

$$p_{D_g}(t,r) = \int_{z_d} p_{D_g}\left(r \mid t, z_d\right) p(t) p\left(z_d\right) dz_d, \tag{3}$$

the above equation can be simplified by [33]:

$$p_{D_g}(t,r) \approx \frac{1}{K} \sum_{k}^{K} p_{D_g}\left(r \mid t, z_{d_k}\right) p(t), \tag{4}$$

where $p(t)$ is the distribution of the prototypical data t and K is the number of samples in latent vector z_d. The K is the number of samples. Similar with data reconstruction module, a better D_g can contribute to more accurate approximation in $p_{D_g}(t,r)$.

B. Loss Function

For the recovered joint distribution $p_{D_r}(t,r)$ and the degraded joint distribution $p_{D_g}(t,r)$, we further describe how to approximate the two fake distributions to the real distribution $p(t,r)$, i.e., how to train data reconstruction module D_r and degradation generation module D_g productively. To gradually and smoothly update $p_{D_r}(t,r)$ and $p_{D_g}(t,r)$ towards the ground truth $p(t,r)$, we train the framework adversarially for the tractability of equations (1) and (4). Inspired by [34], the basic loss function of our PcGAN is formulated as follows:

$$\begin{aligned} L_{GAN} = & \min_{D_r, D_g} \max_{D} L\left(D_r, D_g, D\right) = E_{(t,r)}[D(t,r)] \\ & - \left\{ \lambda E_{(\hat{t},r)}[D(\hat{t},r)] + (1-\lambda) E_{(t,\hat{r})}[D(t,\hat{r})] \right\}, \end{aligned} \tag{5}$$

where D denotes the discriminator of PcGAN, whose aim is to distinguish the real data pair against the generated ones $(\hat{t}, r)$ and $(t, \hat{r})$. λ is a trade-off parameter that keeps a balance between data construction and degradation generation. To portray the discrepancies more clearly, Wasserstein-1 distance [35] was applied to characterize the above distribution differences.

As discussed in [36], the commonly used loss functions can ensure the stability of the adversarial training. In this case, referring to [37], the loss function of the data reconstruction task and degradation generation task is formulated by $L_2 - norm$, i.e., mean square error (MSE). However, considering the randomness of the latent vector z_d in the degradation generation module, we pay more attention to the statistical information of degraded data $\hat{r}$. Therefore, the total loss function of PcGAN is presented as follows:

$$\begin{aligned} L_t &= L_{GAN} + L_{\text{tradition}} \\ &= \min_{D_r, D_g} \max_{D} L\left(D_r, D_g, D\right) \\ &\quad + \alpha \,\mathrm{MSE}(\hat{t}, t) + \beta \,\mathrm{MSE}(f(\hat{r} - t), f(r - t)), \end{aligned} \tag{6}$$

where $f(\cdot)$ denotes the extracted statistic information, e.g., Gaussian filter, Prewitt operator, etc. α and β are the trade-off parameter.

Moreover, the degradation generation module is the focus of PcGAN and is more difficult to generate than the data reconstruction module in the traffic sign classification task for the perturbation robust of the prototypical data. In this case, the structure of the data reconstruction module in PcGAN is different from the traditional generation model in GAN and is more similar to the variational autoencoder (VAE) structure despite the presence of prototypical data. To ensure the accuracy of encoder-decoder architecture and perturbation robust of the embedding latent vectors, we propose an adaptive consistency loss, which only constrains the internal parameters of the degradation generation module, as shown in Fig. 1:

$$\begin{aligned} L_{AC}^{gene} = L_{\varphi,\vartheta}(r,t) = & \frac{1}{M} \sum_{m=1}^{M} -\log p_{\varphi}\left(t \mid z_r^{(m)}\right) \\ & + D_{KL}\left[q_{\vartheta}\left(z_r \mid r\right) \| p_{\varphi}\left(z_r\right)\right], \end{aligned} \tag{7}$$

where $q_{\vartheta}\left(z_r \mid r\right)$ and $p_{\varphi}\left(t \mid z_r\right)$ corresponds to the probability encoder and decoder, which are both modeled by a network structure, and latent vectors $\left\{z_r^{(m)}\right\}_{m=1}^{M}$ are sampled from $q_{\vartheta}\left(z_r \mid r\right)$ with a reparameterization trick [33].

C. Training and Testing Strategy

Training Phase. There are three modules to be updated in PcGAN, i.e., data reconstruction module D_r, degradation generation module D_g, and discriminative module D. We follow the training strategy in [35], where the three modules are co-trained and alternatively updated. Furthermore, to overcome the turbulent situation in the training of GAN, the Lipschitz constraint is applied to D with the penalty strategy of gradient [38].

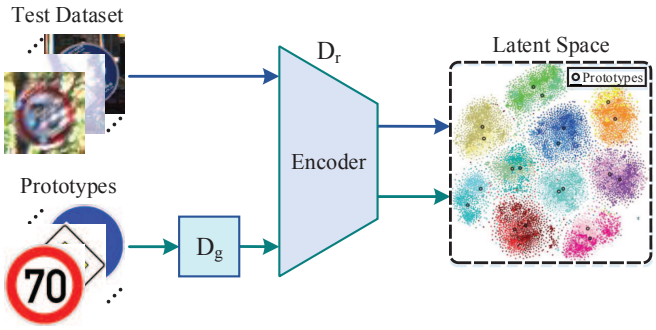


Fig. 3. The illustration of the test phase, where the encoder is a part of the data reconstruction module. The prototypes are degraded by D_g, embedded with the encoder of D_r, and classified with the nearest neighbor method between the high-level features encoded from the test dataset by Euclidean distance.

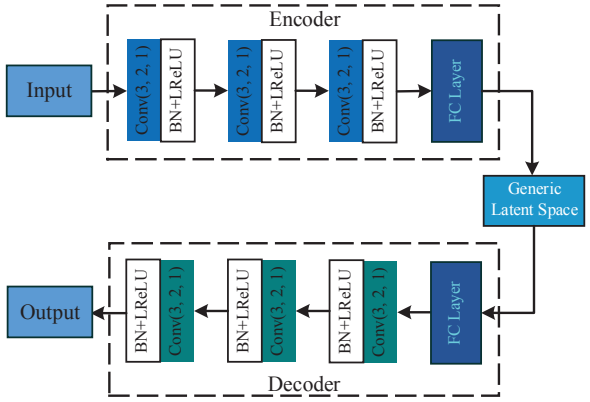


Fig. 4. The network architecture of data reconstruction module D_r and degradation generation module D_g, where Conv(j,k,l) denotes the kernel size $j \times j$, stride k, and padding l, respectively, and LReLU and FC layer are short for LeakyReLU and fully connected layer.

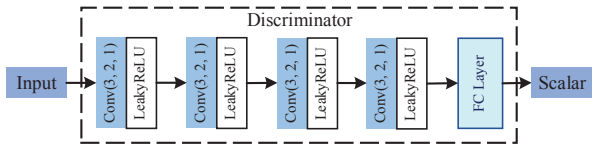


Fig. 5. The network architecture of discriminator module D, where Conv(j,k,l) denotes the kernel size $j \times j$, stride k, and padding l, respectively, and FC layer is short for fully connected layer.

Testing Phase. Considering the complexity of the driving condition, the data captured by the camera sensor can be disturbed in various conditions, thus resulting in serious degradation. Therefore, compared to the manner that calculates the similarity between processed data, it is desirable to complete the classification task from a clustering perspective. As shown in Fig. 3, we initially input the novel classes with the prototypical data and extract the features by the encoder from degradation generation module D_g. In subsequent, when the real data are input, the corresponding features encoded by data reconstruction module D_r will be retrieved by the nearest neighbor technique with Euclidean distance. Finally, the input data is categorized as the class whose prototypical data is nearest to the input in Euclidean distance.

D. Network Architecture

Deep networks are adept in the three aforementioned modules, i.e., data reconstruction module D_r, degradation generation module D_g, and discriminator module D. Note that D_r and D_g share the same backbones with a encoder-decoder structure, as shown in Fig. 4. The encoder of D_r and D_g are constructed with three convolution layers followed by batch normalization and LeakyReLU, whose kernel size, stride, and padding are 3×3, 2, and 1, and a fully connected layer, which is used to offer a feature map into the latent space. The structure of the decoder is symmetric to that of the encoder concerning the latent space z_r. As presented in Fig. 5, t The discriminator module D consists of four convolution layers followed by LeakyReLU, which share the same kernel size, stride, and padding with that in D_r, and one fully connected layer.

IV. Experiment

In this section, the performance of our proposed PcGAN is in comparison with the other three SOTA methods on the two commonly used datasets. In the following content, we will describe the experiment settings, datasets, and comparison methods in detail. In addition, the experiments of the one-shot classification task and data retrieval test are implemented on the German Traffic Sign Recognition Benchmark (GTSRB) [39] and Tsinghua-Tecent 100K (TT100K) [19].

A. Datasets and Comparison Methods

Datasets. Two commonly used datasets are selected for the experiments, which are GTSRB and TT100K. In traffic sign recognition, GTSRB is the most popular dataset with an enormous scale, consisting of three categories, including forbidden, dangerous, and compulsory, with forty-three specific classes. In these datasets, the captured real traffic signal data are degraded by surrounding environments, terrible illumination, complex driving condition, and variable weather, as shown in Fig. 6. Furthermore, the aforementioned problem of intra-class and inter-class data imbalance is occurring in GTSRB, although the scale of the training set and testing set is up to over thirty-nine thousand and twelve thousand images, respectively. Different from GTSRB, TT100K was initially proposed for traffic signal detection tasks with over two hundred classes. Therefore, we exclude the data without

Fig. 6. The reconstruction performance of the challenging data from GTSRB dataset, where (a) represents the prototypes and every two lines of (b)-(e) represent the challenging data and the corresponding recovered data.

clear annotations. Referring to [32], thirty-six classes with more than twenty thousand images are picked out for the traffic signal classification task, which has four overlapping classes with GTSRB.

Comparison Methods. Apart from our proposed PcGAN and VPE [32], QuadNet [40], MatchNet [28], and VPE++ [41] are also added into comparison for one-shot classification task and data retrieval test, where the configuration is in accordance with the publicly available code online without any manual modification. QuadNet [40] is composed of two Siamese networks, the weights shared within each pair. One part embeds features from template images and the other part for real images. Quadruple images from each domain are fed into the corresponding Siamese networks. While MatchNet [28] is not a single network, but about 3 to 4 in the specific implementation of a sequential collection of multiple networks. MatchNet is a fast learning network that combines the more popular attention structure and memory network. In order to quickly learn and train the network, it trains by showing only a few examples of each class, just like how to test when providing some examples of new tasks, switching the task from minibatch to minibatch. For our proposed method, the corresponding features encoded by the data reconstruction module D_r will be retrieved by the nearest neighbor technique with Euclidean distance for the testing phase. Finally, the input data is categorized as the class whose prototypical data is nearest to the input in Euclidean distance.

B. Experimental Settings

As discussed above, GTSRB was initially conceived for the traffic signal recognition task, while TT100K requires some specific filters. Therefore, GTSRB is divided into two partitions for seen classes and unseen classes in cross-dataset evaluation to further estimate the generalization ability of the learned latent space. In specific, twenty-two classes are selected as seen classes and the rest are unseen classes, i.e., our PcGAN is trained with the training set with twenty-two known classes and is evaluated by testing set with all of the classes. Therefore, the twenty-two unseen classes can constitute a validation set for model optimization.

TABLE I
THE ACCURACY (%) OF ONE-SHOT CLASSIFICATION, I.E., ONE NEAREST NEIGHBOR, ON GTSRB AND TT100K DATASETS, WHERE THE BEST VALUES ARE MARKED WITH BOLD.

	GTSRB	TT100K	
	21 unseen classes	All classes	32 unseen classes
QuadNet [40]	45.2754	42.3428	N/A
MatchNet [28]	26.0332	53.1671	49.5355
VPE [32]	56.9861	55.5882	53.0437
VPE++ [41]	70.2461	65.1567	65.7863
PcGAN	**75.7749**	**68.2207**	**66.1562**

In the training phase, the initial weights of data reconstruction module D_r and degradation generation module D_g are set in accordance with the strategy of [42]. Besides, the initial weights of discriminator D follow a normal distribution. The average value and standard deviation are 0 and 0.02, respectively. We apply ADAM optimizer [43] to train the three modules with momentum terms (0.5, 0.9) for D_r and D_g, and (0.9, 0.999) for D. In addition, the learning rates of the three aforementioned modules are 2×10^{-5}, 1×10^{-4}, and 2×10^{-4}, respectively. The mini-batch size is set as 128 with the size 64×64 for inputs. λ, α, and β are set to be 0.5, 1000, and 10 across the entire experimental section. As for the gradient penalty strategy, the coefficients are set as default, followed by [38]. All the experiments in this paper are conducted on a laptop computer with an Intel Core i5-6300 CPU and 16GB RAM using PyTorch on an Nvidia 1080 GPU.

C. One-shot Classification

In this subsection, we offer quantitative analysis between our PcGAN and the aforementioned methods on three conditions from GTSRB and TT100K datasets with the metric of accuracy, which is the twenty-one unseen classes in GTSRB, all classes in TT100K, and thirty-two unseen classes in TT100K. For all unseen classes, we randomly select one sample as a representative of the novel class. As mentioned in Testing Phase, our PcGAN completes the classification task in a clustering manner. Therefore, our network is only trained on the base class, and the samples from novel classes are encoded by PcGAN and stored in the support set.

As shown in Table I, the classification accuracy of the proposed PcGAN is much higher than other methods in the unseen classes in GTSRB, which fully demonstrates the generalization of the implicit embedding space of our PcGAN. Furthermore, the metric values in cross-dataset evaluation, i.e., TT100K,

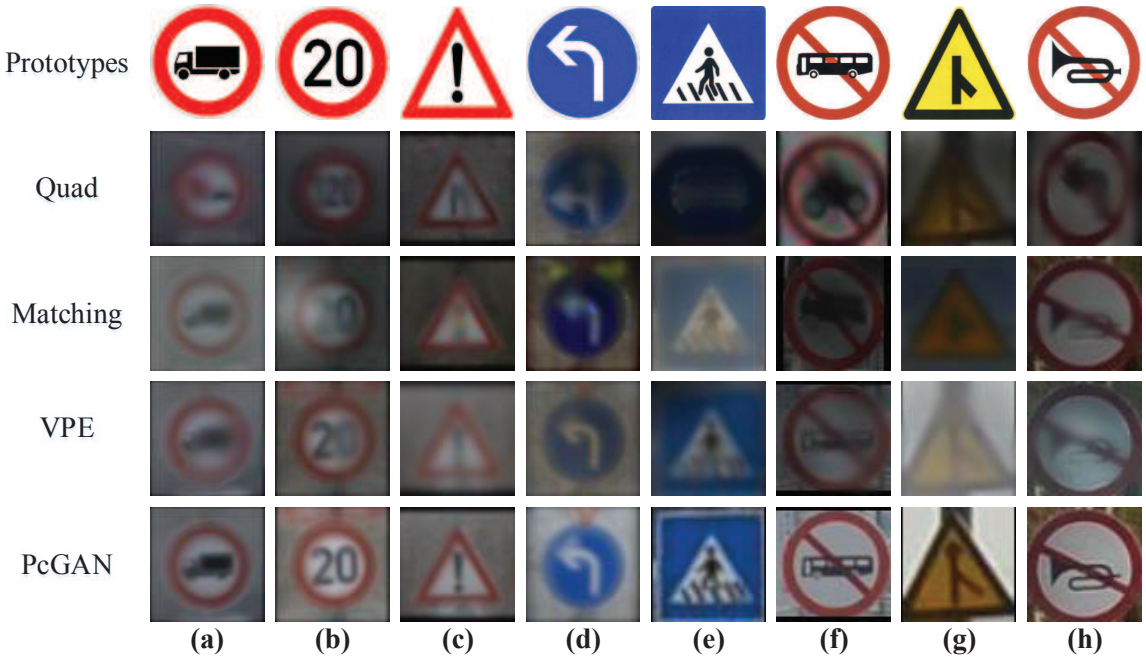


Fig. 7. The average of the top 50 data retrieved by the nearest neighbor, where the prototypical data are presented in the first line, and the average of the retrieved data is presented from the second line to the fifth line. (a)-(d) are the unseen classes from GTSRB dataset, and (e)-(h) are those from TT100K dataset.

also outperform the existing SOTAs both in all classes and thirty-two unseen classes, which can be attributed to the degradation robust structure, including the two alternating optimized modules of degradation generation and removal, to resist the subtle variations of the input data. In Table I, the poor classification performance of MatchNet [28] mainly lies in that the attention-based kernel is sensitive to the seen classes and fails to generate the proper relationship between the real inputs and the unseen prototypes.

TABLE II
THE AUC OF DATA RETRIEVAL TEST, WHERE THE BEST VALUES ARE MARKED WITH BOLD.

	QuadNet	MatchNet	VPE	VPE++	PcGAN
GTSRB	N/A	54.29	64.32	65.27	**68.47**
TT100K	N/A	41.32	41.31	46.23	**47.25**

D. Data Retrieval Test

To vividly illustrate the embedding performance among the compared methods, we further provide the data retrieval test both qualitatively and quantitatively. We retrieve the data by querying prototyping data in the embedding space with the encoded real inputs under the 50-nearest neighbor (NN) algorithm, the clustering metrics of which are referring to the compared techniques [28], [32], [40], e.g., we select Euclidean distance as the NN metric in our PcGAN, and the final outputs of the test are the average of the 50 nearest retrieved data of compared methods. As shown in Fig. 7, it is evident that our retrieval results outperform the compared methods for the clearer outputs, which is due to the degradation robust module to resist the perturbation of the inputs and classify them to the correct classes. The quantitative experiment is presented in Table II with AUC as the evaluation metric, which further demonstrates the superiority of the proposed PcGAN. The mediocre retrieval performance of VPE [32] mainly lies in that VPE forces the real inputs to learn the implicit distribution of their prototypes without any anti-degradation strategy, which can result in the far distance between the embedding of inputs and their prototypical data in the learned latent space.

E. Real Data Reconstruction Task

As mentioned in Fig. 2, the real captured traffic data can be challenging to post-processing for the variable surrounding environments, complex weather, etc. Therefore, Fig. 6 is presented to demonstrate the degradation robustness of our proposed PcGAN. It is worth noting that we only test the reconstruction performance in GTSRB under both seen and unseen classes to demonstrate the anti-degradation and generalization of PcGAN without the cross-dataset evaluation for the intra-class and inter-class data imbalance [32]. Considering that most of the compared techniques are metric-based learning strategies, which only focus on the high-level features of the input data rather than the reconstruction task, we only compare them in a quantitative manner, i.e., Table III, and the quantitative analysis, i.e., Fig. 8, contains the rest thirty-nine classes apart from the four classes mentioned in Fig. 6. As shown in Table III, both PSNR and SSIM of our PcGAN are much higher than the compared method in the test set of GTSRB for over twelve thousand real data, which comprehensive demonstrate the degradation robustness and texture fidelity of our proposed PcGAN for the specific designed alternating update structure and the multi-task loss

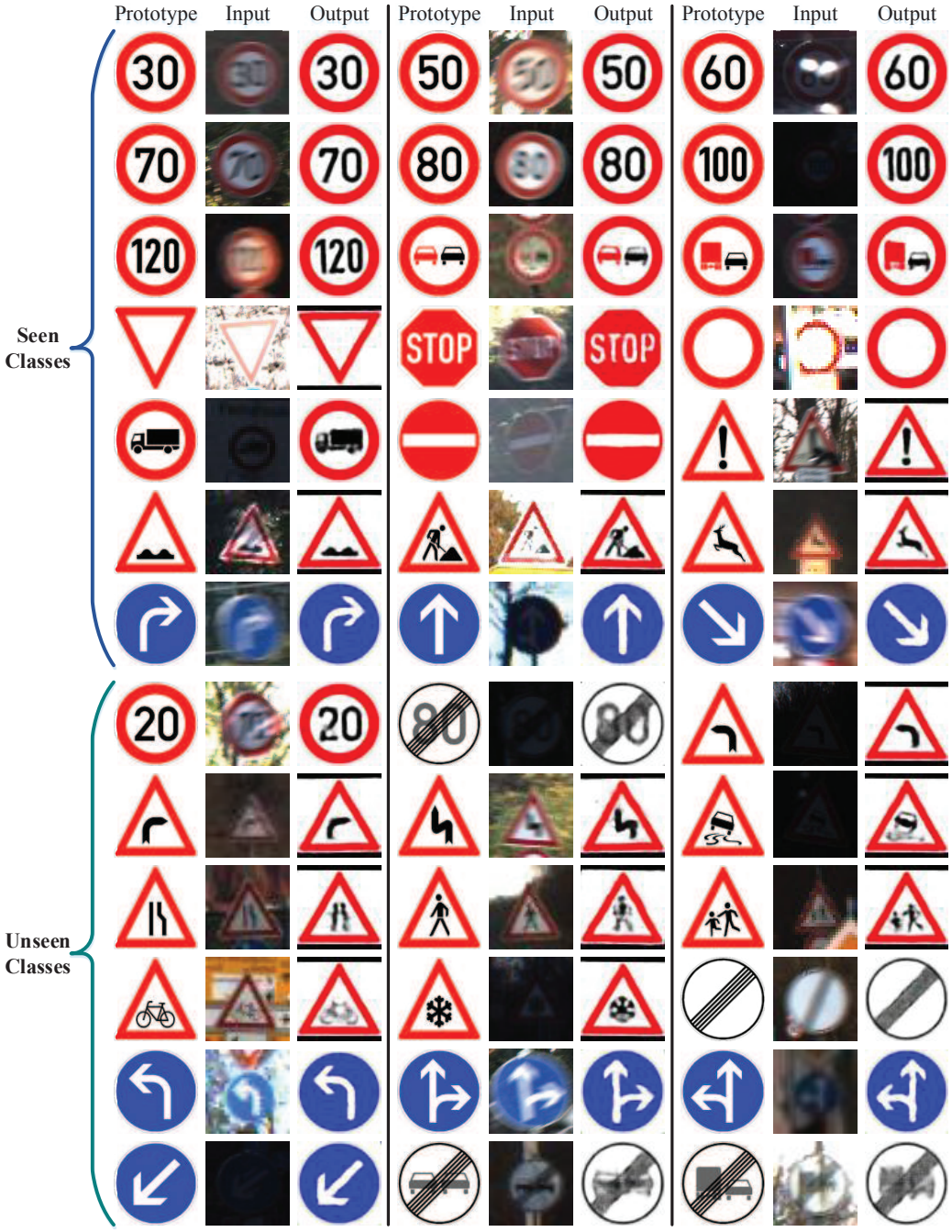


Fig. 8. The output of PcGAN in GTSRB dataset apart from the classes mentioned in Fig. 6, where the seen classes are presented from the first line to the seventh line and the unseen classes are exhibited from the eighth line to the thirteenth line.

function. Figs. 6 and 8 further vividly illustrate the aforementioned arguments. In Fig. 8, there are remaining twenty-one seen classes and eighteen unseen classes showing with their real challenging data from the test set and the corresponding reconstruction images. It is evident that although the inputs are degraded from one to even all of the aforementioned challenging conditions, i.e., over-darkness, blur, overexposure, occlusion, the corresponding outputs of seen classes are quite high-quality and very similar to the prototypes and those of unseen classes can also explicitly convey most of the semantic information of the input traffic sign, which can be attributed to the degradation robustness module and the multi-task loss function.

TABLE III
THE AVERAGE VALUES OF TWO METRICS ON THE GTSRB DATASETS.

	QuadNet	MatchNet	VPE	VPE++	PcGAN
PSNR	6.2763	7.6095	8.7687	10.7443	**14.2146**
SSIM	0.2087	0.3192	0.5263	0.6885	**0.7421**

TABLE IV
ABLATION STUDIES OF DIFFERENT COMBINATIONS OF LOSS FUNCTIONS IN THREE DISTINCT EXPERIMENTS, WHERE GTSRB(21), TT100K(ALL), AND TT100K(32) CORRESPOND TO 21 UNSEEN CLASSES IN GTSRB, ALL CLASSES IN TT100K, AND 32 UNSEEN CLASSES IN TT100K.

L_{GAN}	$L_{\text{tradition}}$	L_{AC}^{gene}	One-shot Classification			Retrieval Experiment		Reconstruction Test	
			GTSRB(21)	TT100K(all)	TT100K(32)	GTSRB	TT100K	PSNR	SSIM
✓	✗	✗	70.27	62.83	61.75	64.63	42.87	12.26	0.668
✓	✓	✗	73.12	67.35	65.24	66.71	46.13	13.54	0.716
✓	✓	✓	**75.77**	**68.22**	**66.16**	**68.47**	**47.25**	**14.21**	**0.742**

F. Ablation Study

In Table IV, we present the ablation studies of different combinations of loss functions in three distinct experiments, i.e., one-shot classification, retrieval experiment, and reconstruction test. As shown in Table IV, the proposed PcGAN with the complete combination of loss functions achieves the best performance in all the experiments, which fully demonstrates the superiority of the multi-task loss function.

V. CONCLUSIONS

In this paper, we propose a novel traffic sign classification method based on the conditional generative adversarial network for intelligent transportation systems. The proposed PcGAN is an end-to-end network, which has the alternative update modules, i.e., the data reconstruction module and the degradation generation module, and introduces the dual adversarial strategy to approximate the implicitly joint distribution in the field of one-shot learning of traffic sign classification, which can simultaneously simulate the process of degradation removal and generation. This approach solves inevitable inherent challenges to one-time classification tasks due to the low quality of traffic sign data captured by in-vehicle cameras and the challenge of imbalanced data from training data within and between classes. In the network framework, we propose a multi-task loss function to jointly constrain the network training, including a basic GAN loss, a task-based loss, and an adaptive consistency loss, to implement the reconstruction of real data of traffic signs and the degradation of the corresponding prototypes so as to achieve a more accurate classification performance. However, our work has a small visual gap with the prototypical data in recovering high-quality data of traffic signs because the prototypical data is the most standard data in the corresponding class and is not explicitly designed for the input. Sufficient experiments on publicly available databases with other three states-of-the-art fusion algorithms (QuadNet, MatchNet, and VPE) comprehensively illustrate the superiority of our proposed method both in a one-shot classification task and data retrieval task. In the future, others in the field can utilize this method to overcome the aforementioned problems and achieve more accurate classification performance.

REFERENCES

[1] F. Almutairy, T. Alshaabi, J. Nelson, and S. Wshah, "Arts: Automotive repository of traffic signs for the united states," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 1, pp. 457–465, 2019.

[2] Y. Li, S. Yang, Y. Zheng, and H. Lu, "Improved point-voxel region convolutional neural network: 3d object detectors for autonomous driving," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–7, 2021, doi=10.1109/TITS.2021.3071790.

[3] C. G. Serna and Y. Ruichek, "Traffic signs detection and classification for european urban environments," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 10, pp. 4388–4399, 2019.

[4] Z. Wang, J. Wang, Y. Li, and S. Wang, "Traffic sign recognition with lightweight two-stage model in complex scenes," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1121 – 1131, 2020, 10.1109/TITS.2020.3020556.

[5] R. Yang, H. Ma, J. Wu, Y. Tang, X. Xiao, M. Zheng, and X. Li, "Scalablevit: Rethinking the context-oriented generalization of vision transformer," *arXiv preprint arXiv:2203.10790*, 2022.

[6] G. Xu, X. Deng, X. Zhou, M. Pedersen, L. Cimmino, and H. Wang, "Fcfusion: Fractal component-wise modeling with group sparsity for medical image fusion," *IEEE Transactions on Industrial Informatics*, pp. 1–9, 2022, 10.1109/TII.2022.3185050.

[7] Y. Lu, C. He, Y.-F. Yu, G. Xu, H. Zhu, and L. Deng, "Vector co-occurrence morphological edge detection for colour image," *IET Image Processing*, vol. 15, no. 13, pp. 3063–3070, 2021.

[8] J. Yan, H. Chen, K. Wang, Y. Ji, Y. Zhu, J. Li, D. Xie, Z. Xu, J. Huang, S. Cheng *et al.*, "Hierarchical attention guided framework for multi-resolution collaborative whole slide image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 153–163.

[9] M. Ju, C. Ding, W. Ren, Y. Yang, D. Zhang, and Y. J. Guo, "Ide: Image dehazing and exposure using an enhanced atmospheric scattering model," *IEEE Transactions on Image Processing*, vol. 30, pp. 2180–2192, 2021.

[10] Z. Yue, Q. Zhao, L. Zhang, and D. Meng, "Dual adversarial network: Toward real-world noise removal and noise generation," in *European Conference on Computer Vision*. Springer, 2020, pp. 41–58.

[11] C. He, X. Wang, L. Deng, and G. Xu, "Image threshold segmentation based on glle histogram," in *2019 International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*. IEEE, 2019, pp. 410–415.

[12] G. Lu, C. He, L. Xu, J. Ren, G. Xu, and H. Zhao, "Infrared and visible image fusion based on local gradient constraints," in *2020 International Conferences on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics (Cybermatics)*. IEEE, 2020, pp. 571–575.

[13] H. Zhu, H. Ni, S. Liu, G. Xu, and L. Deng, "Tnlrs: Target-aware non-local low-rank modeling with saliency filtering regularization for infrared small target detection," *IEEE Transactions on Image Processing*, vol. 29, pp. 9546–9558, 2020.

[14] Y. Tian, J. Gelernter, X. Wang, J. Li, and Y. Yu, "Traffic sign detection using a multi-scale recurrent attention network," *IEEE Transactions on Intelligent Transportation systems*, vol. 20, no. 12, pp. 4466–4475, 2019.

[15] S. Ahmed, U. Kamal, and M. K. Hasan, "Dfr-tsd: A deep learning based framework for robust traffic sign detection under challenging weather conditions," *IEEE Transactions on Intelligent Transportation Systems*, no. 6, pp. 5150 – 5162, 2021.

[16] D. Temel, M.-H. Chen, and G. AlRegib, "Traffic sign detection under challenging conditions: A deeper look into performance variations and spectral characteristics," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 9, pp. 3663–3673, 2019.

[17] H. Zhu, H. Peng, G. Xu, L. Deng, Y. Cheng, and A. Song, "Bilateral weighted regression ranking model with spatial-temporal correlation filter for visual tracking," *IEEE Transactions on Multimedia*, vol. 24, pp. 2098–2111, 2021.

[18] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "The german traffic sign recognition benchmark: a multi-class classification competition," in *The 2011 International Joint Conference on Neural Networks*. IEEE, 2011, pp. 1453–1460.

[19] Z. Zhu, D. Liang, S. Zhang, X. Huang, B. Li, and S. Hu, "Traffic-sign detection and classification in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2110–2118.

[20] X. Sun, B. Wang, Z. Wang, H. Li, H. Li, and K. Fu, "Research progress on few-shot learning for remote sensing image interpretation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 2387–2402, 2021.

[21] F. Li, Fergus, and Perona, "A bayesian approach to unsupervised one-shot learning of object categories," in *Proceedings Ninth IEEE International Conference on Computer Vision*. IEEE, 2003, pp. 1134–1141.

[22] Q. Zheng, J. Zhu, H. Tang, X. Liu, Z. Li, and H. Lu, "Generalized label enhancement with sample correlations," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2021.

[23] X. Xu, J. Tian, K. Lin, H. Lu, J. Shao, and H. T. Shen, "Zero-shot cross-modal retrieval by assembling autoencoder and generative adversarial network," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 17, no. 1s, pp. 1–17, 2021.

[24] X. Xu, H. Lu, J. Song, Y. Yang, and H. T. Shen, "Ternary adversarial networks with self-supervision for zero-shot cross-modal retrieval," *IEEE Transactions on Cybernetics*, vol. 50, no. 6, pp. 2400–2413, 2019.

[25] H. Lu, M. Zhang, X. Xu, Y. Li, and H. T. Shen, "Deep fuzzy hashing network for efficient image retrieval," *IEEE Transactions on Fuzzy Systems*, vol. PP, no. 99, pp. 1–1, 2020.

[26] Q. Sun, Y. Liu, T.-S. Chua, and B. Schiele, "Meta-transfer learning for few-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 403–412.

[27] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International Conference on Machine Learning*, vol. 70. PMLR, 2017, pp. 1126–1135.

[28] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra *et al.*, "Matching networks for one shot learning," *Advances in Neural Information Processing Systems*, vol. 29, pp. 3630–3638, 2016.

[29] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *Advances in Neural Information Processing Systems*, vol. 30, pp. 4077–4087, 2017.

[30] L. Deng, G. Xu, Y. Dai, and H. Zhu, "A dual stream spectrum deconvolution neural network," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 5, pp. 3086–3094, 2021.

[31] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, "Human-level concept learning through probabilistic program induction," *Science*, vol. 350, no. 6266, pp. 1332–1338, 2015.

[32] J. Kim, T.-H. Oh, S. Lee, F. Pan, and I. S. Kweon, "Variational prototyping-encoder: One-shot learning with prototypical images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9462–9470.

[33] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[34] C. Li, K. Xu, J. Zhu, and B. Zhang, "Triple generative adversarial nets," *arXiv preprint arXiv:1703.02291*, 2017.

[35] J. Cao, L. Mo, Y. Zhang, K. Jia, C. Shen, and M. Tan, "Multi-marginal wasserstein gan," *Advances in Neural Information Processing Systems*, vol. 32, pp. 1776–1786, 2019.

[36] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1125–1134.

[37] T. Kaneko and T. Harada, "Noise robust generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8404–8414.

[38] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of wasserstein gans," *arXiv preprint arXiv:1704.00028*, 2017.

[39] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition," *Neural Networks*, vol. 32, pp. 323–332, 2012.

[40] J. Kim, S. Lee, T.-H. Oh, and I. S. Kweon, "Co-domain embedding using deep quadruplet networks for unseen traffic sign recognition," in *Thirty-Second AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[41] C. Xiao, N. Madapana, and J. Wachs, "One-shot image recognition using prototypical encoders with reduced hubness," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 2252–2261.

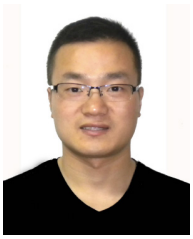
[42] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1026–1034.

[43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

Lizhen Deng (Member) received the B.S. degree in electronic information science and technology from Huaibei Coal Industry Teachers College, Huaibei, China, in 2007, and the M.S. degree in communication and information systems from Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2010. She received her Ph.D. degree in electrical engineering from Huazhong University of Science and Technology, China, in 2014. She is currently Associate Professor with the Nanjing University of Posts and Telecommunications, Nanjing. Her current research interests include computer vision.

Chunming He received the B.S. degree in communication engineering from Nanjing University of Posts and Telecommunications, Jiangsu Nanjing, China in 2021. Now, he is pursing his master degree in artificial intelligence, Tsinghua Shenzhen International Graduate School, Tsinghua University. His research interest includes computer vision.

Guoxia Xu (Member) received the B.S. degree in information and computer science from Yancheng Teachers University, Jiangsu Yancheng, China in 2015, and the M.S. degree in computer science and technology from Hohai University, Nanjing, China in 2018. Now, he is pursing his Ph.D. degree in computer science with Department of Computer Science, Norwegian University of Science and Technology, Gjovik Norway. His research interest includes image processing.



Hu Zhu (Member) received the B.S. degree in mathematics and applied mathematics from Huaibei Coal Industry Teachers College, Huaibei, China, in 2007, and the M.S. and Ph.D. degrees in pattern recognition and intelligent systems from Huazhong University of Science and Technology, Wuhan, China, in 2009 and 2013, respectively. He is currently Professor with the Nanjing University of Posts and Telecommunications, Nanjing. His research interests include pattern recognition.

Hao Wang (Senior Member) is an associate professor in the Department of Computer Science in Norwegian University of Science and Technology, Norway. He received his Ph.D. degree and B.Eng. degree, both in computer science and engineering, from South China University of Technology in 2006 and 2000, respectively. His research interests include big data analytics, industrial internet of things, high performance computing, and safety-critical systems. He has published 140+ papers in reputable international journals and conferences. He served as a TPC co-chair for IEEE CPSCom 2020, IEEE CIT 2017, ES 2017, and DataCom 2015, a senior TPC member for CIKM 2019, and reviewers/TPC members for many journals and conferences. He is the Chair for Sub-TC on Healthcare in IEEE IES Technical Committee on Industrial Informatics.