SJDL-Vehicle: Semi-supervised Joint Defogging Learning for Foggy Vehicle Re-identification

Wei-Ting Chen[1,3*], I-Hsiang Chen[2*], Chih-Yuan Yeh[2], Hao-Hsiang Yang[2], Jian-Jiun Ding[2], and Sy-Yen Kuo[2]

[1]Graduate Institute of Electronics Engineering, National Taiwan University, Taiwan
[2]Department of Electrical Engineering, National Taiwan University, Taiwan
[3]ASUS Intelligent Cloud Services, Taiwan
{f05943089, f09921058, f09921063, r05921014, jjding, sykuo}@ntu.edu.tw

Abstract

Vehicle re-identification (ReID) has attracted considerable attention in computer vision. Although several methods have been proposed to achieve state-of-the-art performance on this topic, re-identifying vehicle in foggy scenes remains a great challenge due to the degradation of visibility. To our knowledge, this problem is still not well-addressed so far. In this paper, to address this problem, we propose a novel training framework called Semi-supervised Joint Defogging Learning (SJDL) framework. First, the fog removal branch and the re-identification branch are integrated to perform simultaneous training. With the collaborative training scheme, defogged features generated by the defogging branch from input images can be shared to learn better representation for the re-identification branch. However, since the fog-free image of real-world data is intractable, this architecture can only be trained on the synthetic data, which may cause the domain gap problem between real-world and synthetic scenarios. To solve this problem, we design a semi-supervised defogging training scheme that can train two kinds of data alternatively in each iteration. Due to the lack of a dataset specialized for vehicle ReID in the foggy weather, we construct a dataset called FVRID which consists of real-world and synthetic foggy images to train and evaluate the performance. Experimental results show that the proposed method is effective and outperforms other existing vehicle ReID methods in the foggy weather. The code and dataset are available in https://github.com/Cihsaing/SJDL-Foggy-Vehicle-Re-Identification--AAAI2022.

Q1: What is the motivation for the work?

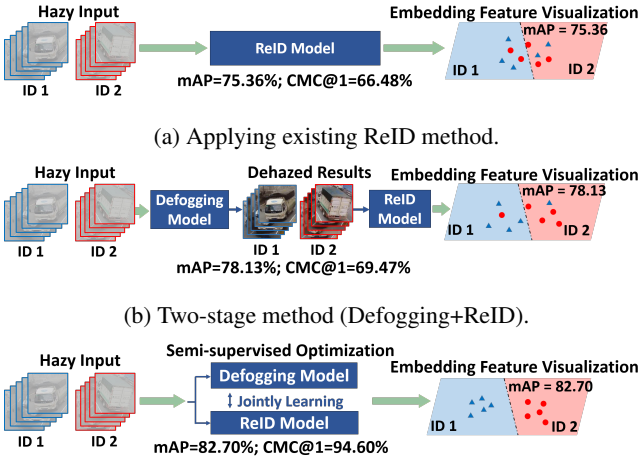
Introduction

With the prosperity of the deep convolutional neural network (DCNN) and comprehensive construction of dataset, vehicle re-identification (ReID) has attained great success in the past decade (Lou et al. 2019; He et al. 2019; Meng et al. 2020; Chen et al. 2020; He et al. 2021). Vehicle ReID is indispensable for building intelligent transportation and public security systems. Its goal is to find images of the same vehicle in a large gallery set based on a query image under multiple cameras and various viewpoints. Though several methods have shown superior performance on the normal images, they usually fail to perform vehicle ReID tasks under the foggy scenario, which is one of the most common weather types that appeared in the real world. This is because these methods are designed for clear images. Fog is an atmospheric phenomenon that consists of smoke, dust, and other floating particles, which may lead to poor visibility and degrade the features extracted from these images for the purpose of vehicle ReID.

Q1: Why doesn't the people's problem have a trivial solution?

Q1: What are the previous solutions and why are they inadequate?

A straightforward way to resolve this problem is to improve the visibility of input images via existing defogging strategies (e.g., the MPR-Net (Zamir et al. 2021) or the MS-BDN (Dong et al. 2020)) and then conduct the ReID afterward. However, this two-stage solution is not effective for the following reasons. First, the conventional image defogging methods are not trained for the purpose of ReID but



(a) Applying existing ReID method.

(b) Two-stage method (Defogging+ReID).

(c) Proposed Method.

Figure 1: Comparison of different strategies for vehicle ReID in the foggy weather in terms of the mean average precision (mAP) and CMC@1. We adopt (Meng et al. 2020) as the existing ReID method and MPR-Net (Zamir et al. 2021) as the defogging method. One can notice that the proposed method can achieve superior performance on this task compared to other methods. In this figure, we evaluate the real-world ReID dataset in the foggy weather.

* Equally-contributed first authors.

for human perception. Using these defogging methods as the pre-process cannot always guarantee the performance of ReID. Second, integrating the defogging and ReID models may increase the complexity of the entire system because most defogging methods require a heavy computational burden. Fig. 1 illustrates the limitation of the existing ReID methods and the aforementioned two-stage methods.

Q2: What is the proposed solution?

Q2: How is the solution achieved?

Q2: How does it represent an improvement?

Therefore, to tackle the vehicle ReID problem in the foggy scenario, in this paper, a novel joint defogging learning (JDL) paradigm to keep the fog-free feature representation is proposed. The proposed JDL mechanism is embedded in a two-branch network, which consists of a defogging model, a re-identification model, and a feature sharing module. Specifically, the proposed framework is trained in an end-to-end fashion to learn the defogging and ReID jointly. With the simultaneous optimization paradigm, clean features extracted by the defogging branch from foggy input images for visibility enhancement can be shared to learn better ReID features in the re-identification branch. Therefore, the performance of ReID in the foggy weather can be improved effectively by this design.

Moreover, since the fog-free ground truth in the real-world scenario is intractable, directly leveraging real-world foggy data on the aforementioned training framework is challenging. Therefore, to train the proposed network, a synthetic foggy vehicle ReID dataset is constructed. However, the performance of the network on real-world ReID is limited because the network is only trained on synthetic data. To well address the domain gap between real-world and synthetic foggy data, we proposed a semi-supervised defogging scheme to train our network in a supervised way and an unsupervised way alternatively. With this mechanism, the domain gap problem can be solved effectively and the proposed method can achieve state-of-the-art performance in both synthetic and real-world datasets.

Q5: What are the paper's contributions?

The contribution of this paper is summarized as follows:

- A novel training framework that unifies the defogging network and re-identification network is proposed. The joint defogging learning framework can preserve defogging features for the ReID to cope with the poor visibility problem.
- A semi-supervised defogging training mechanism is proposed to optimize the proposed network on both synthetic data and real-world data alternatively to address the domain gap problem.
- Since there is no existing dataset mainly for ReID in the foggy weather, we reorganize the existing benchmarks and construct a dataset called Foggy Vehicle ReID (FVRID).

Related Works

Vehicle Re-identification. With the development of the DCNN and the releases of several large-scale benchmarks (e.g., VehicleID (Liu et al. 2016), VeRi-776 (Liu et al. 2017), VERI-Wild (Lou et al. 2019), Vehicle-1M (Guo et al. 2018)), vehicle re-identification (vehicle ReID) has attracted more and more attention recently. Numerous methods have been proposed and achieved outstanding performance. Most of existing methods rely on DCNN techniques and can be divided into several classes. The first class leverages the meta-information to contribute to embedding feature fusion. Shen *et al.* (Shen et al. 2017) leveraged the spatial-temporal regularization and the visual-spatio-temporal path proposals to improve the accuracy of the vehicle ReID. Zheng *et al.* (Zheng et al. 2019) proposed a unified-attribute guided network which learns the global feature, the camera view and the vehicle type and color. The second class is to leverage the local information for representation learning. For example, He *et al.* (He et al. 2019) developed a part-regularized mechanism to preserve discriminative features based on local information (e.g., light bounding box and window). Khorramshahi *et al.* (Khorramshahi et al. 2020) applied the Variational Auto-Encoder (VAE) to generate the coarse output and the pixel-wise difference of the original input. This coarse output contains important details in local regions which can benefit the ReID process. Meng *et al.* (Meng et al. 2020) proposed a vehicle part parser to retrieve the common region information to conduct the common-visible attention, which enhances the vehicle embeddings under different views. The third class applied the Generative Adversarial Network to conduct feature learning. Lou *et al.* (Lou et al. 2019) proposed the FDA-Net to generate hard examples based on the GAN architecture to improve the ability of ReID. Zhou *et al.* (Zhou and Shao 2018) proposed to learn global multi-view feature representation based on the single-view input features by the viewpoint-aware attention model and the GAN. The last class is based on the Vision Transformer (ViT), a powerful neural network architecture. He *et al.* (He et al. 2021) leveraged the ViT to encode non-visual information as a vector for embedding representation. Then, vector projection was adopted to encode the correlation between patches to acquire robust feature representation.

Single Image Fog Removal. The formation of fog can be modeled by the atmospheric scattering model (Nayar and Narasimhan 1999; Narasimhan and Nayar 2003):

$$I(x) = J(x)\,t(x) + A\,(1 - t(x))\,, \tag{1}$$

where $I(x)$ is the foggy image acquired by the sensor, $J(x)$ is the fog-free image, A is the global atmospheric light, and $t(x)$ is the transmission map which can be defined as $t(x) = e^{-\beta d(x)}$, where β is the scattering coefficient of the atmosphere and $d(x)$ is the scene depth from the sensor to the object. Based on this model, several fog removal methods have been proposed in past decades. These methods can be categorized into two classes. The first class is to extract the haze-relevant features such as the dark channel (He, Sun, and Tang 2010), the color attenuation (Zhu, Mai, and Shao 2015), and the haze-line (Berman, Avidan et al. 2016) to perform fog removal based on images priors. The other class is to leverage the DCNN. For example, Zhang *et al.* (Zhang and Patel 2018) proposed a densely connected pyramid dehazing network (DCPDN) based on atmospheric model. Qu *et al.* (Qu et al. 2019) adopted global visual perception (Chen 2005; Yang et al. 2021) to recover the clear image from coarse to fine scales by multi-resolution generators and discriminators. Zamir *et al.* (Zamir et al.

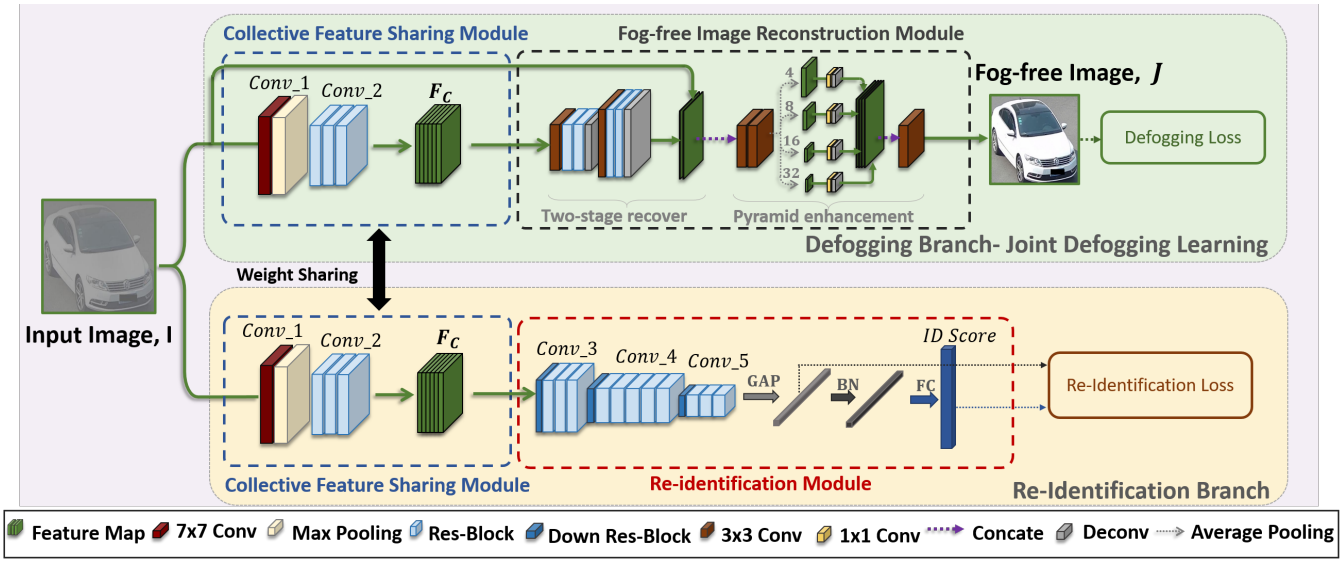


Figure 2: The architecture of the proposed joint defogging learning network for vehicle ReID. The proposed network consists of two branches: the re-identification branch and the defogging branch. Two branches share the Collective Feature Sharing Module to learn defogging and ReID simultaneously. Note that, the defogging branch is only involved in the network during the training stage.

2021) designed a multi-stage architecture to produce contextually enriched and spatially accurate outputs.

Methodology

Overview of the Proposed Architecture

As shown in Fig. 2, the proposed architecture mainly consists of two branches: the re-identification branch (ReID Branch) and the defogging branch. At the training stage, both ReID and defogging branches share a feature extraction module called the collective feature sharing module (CFSM) to ensure that the fog-free features generated by this module termed F_C can be applied to each branch in our joint defogging learning architecture. Then, the extracted features F_C are passed to the fog-free image reconstruction module (FIRM) and the re-identification module (ReIDM) to produce the corresponding outputs. At the inference stage, only the CFSM and the ReIDM are required to perform ReID. By this architecture, the performance of ReID in the foggy weather can be improved significantly without additional computational burden at the inference stage.

Re-identification Branch

In the re-identification branch, we adopt the ResNet-50 (He et al. 2016) as the backbone. The first two Conv blocks from ResNet-50 are assigned as the CFSM which is shared with the defogging branch for feature extraction while the rest blocks are assigned as the ReIDM. The detail of CFSM is illustrated in the next sub-section. We pass the features extracted by the CFSM (termed F_C) through rest ResBlocks and down-scale them by the global average pooling (GAP) and the batch normalization (BN) layer to generate 2048-d embedding features. Then, the fully connected layer (FC layer) is adopted to align the number of identities for the classification. We adopt triplet loss $\mathcal{L}_{Tri}$ and ID loss $\mathcal{L}_{ID}$ to optimize the ReID network. It can be defined as follows.

$$\mathcal{L}_{Tri} = \frac{1}{Q}\sum_{i=1}^{Q}\left[\max_{z_p\in\mathcal{P}(z_i)} D(z_i, z_p) - \min_{z_n\in\mathcal{N}(z_i)} D(z_i, z_n) + M\right]_+ \tag{2}$$

where Q denotes the batch size. $\mathcal{P}(z_i)$ and $\mathcal{N}(z_i)$ represent the positive and negative sample sets where z_i represents the extracted features from i^{th} input sample. M is the margin of the triplet loss, $D(\cdot,\cdot)$ is the Euclidean distance of two features, and $[\cdot]_+$ equals to $max(\cdot, 0)$. Second, the $\mathcal{L}_{ID}$ is defined as:

$$\mathcal{L}_{ID} = -\frac{1}{Q}\sum_{i=1}^{Q}\log\frac{\exp(\sigma_i^{y_i})}{\sum_{j=1}^{C}\exp(\sigma_i^j)} \tag{3}$$

where σ_i^j represents the output of the FC layer with the class j based on i^{th} input image. C presents the total number of the class. y_i donates the ground truth class.

Although this architecture can extract the feature for vehicle ReID in fog-free conditions effectively, the performance may be degraded dramatically in the foggy weather because the fog may deteriorate the ability of feature extraction. Thus, we propose a joint defogging learning strategy that simultaneously deals with fog removal and re-identification by introducing the defogging branch.

Defogging Branch

The defogging branch aims to improve the quality of common features F_C extracted by the CFSM to boost the performance of the re-identification branch in the foggy weather. To accomplish this goal, two modules are adopted in this branch, namely, the CFSM and the fog-free image reconstruction module (FIRM).

Collective Feature Sharing Module. The CFSM aims to extract the features of the input image which contains the crucial information for jointly learning defogging and vehicle ReID. The CFSM is designed based on some convolution blocks in the ReID network because we want to keep the architecture simple and prevent the network from increasing computational burden. Based on previous works (Chen et al. 2021; Hui et al. 2020), the features extracted from the shallower layer of the network contain more spatial and low-level information which can benefit the fog removal process while those of deeper layers contain more high-level information. Thus, the proposed CFSM is constituted by the first two convolution blocks (Conv_2) in the re-identification branch. The extracted features by the CFSM are delivered to the fog-free image reconstruction module for defogging simultaneously.

Fog-free Image Reconstruction Module. The features extracted by the CFSM may be deteriorated by fog, which may lead to limited performance on vehicle ReID. To reconstruct the F_C which are shared to the re-identification branch during the joint learning stage, the FIRM is proposed and its architecture is illustrated as follows. First, the extracted features F_C pass through one convolution block and two ResBlocks to extract more accurate features for defogging. Then, since the dimension of these features is reduced in the previous layers, the deconvolution operation is conducted to upsample the features for matching the resolution of the input. These two operations are repeated two times. Then, the upsampled features are concatenated with the input image and delivered to the pyramid enhancement block (Qu et al. 2019) to generate the final fog-free results. The operation of pyramid enhancement can extract the features based on different receptive fields and multi-scale learning, which can expand the representational ability of the network. This operation is based on pyramid pooling (Zhao et al. 2017) and the detail of it is illustrated as follows. Initially, there are two 3×3 front-end convolution layers. The output of the front-end convolution layer is passed through an average pooling layer to downsample by factors of 4×, 8×, 16×, 32× to build a four-scale pyramid. Then, 1×1 convolution is applied to reduce the dimension on each scale layer. Next, the features are up-sampled to the original size and concatenated together. Finally, the 3×3 convolution is used on the concatenated features to generate the output.

defogging processing

Semi-supervised Optimization for Joint Defogging Learning

Based on the proposed joint defogging learning architecture, although the performance of vehicle ReID can be significantly improved in the foggy weather, it may be limited in real-world scenarios. Specifically, since the ground truth of the fog-free image in real-world scenarios is intractable, the defogging branch can only be optimized on the synthetic data. Thus, the performance of vehicle ReID may have a domain gap between real-world and synthetic scenarios. To address this issue, we proposed a semi-supervised optimization scheme to train both real-world images and synthetic images alternatively in each iteration. The training process

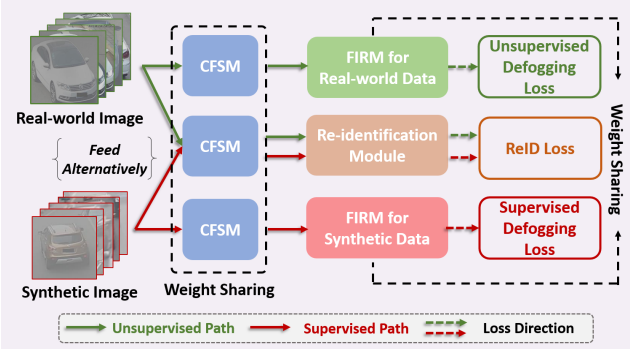


Figure 3: The overview of the proposed semi-supervised joint defogging learning. Based on different sources of input data, two optimized schemes are applied for the defogging branch, namely unsupervised and supervised defogging optimizations. Different sources of data are fed into the network alternatively at the training stage.

can be divided into two parts: (i) *the supervised defogging learning stage for the synthetic data*, and (ii) *the unsupervised defogging learning stage for the real-world data*. The detail is shown in Fig. 3.

Supervised Learning Stage. In this stage, the defogging branch is optimized based on the synthetic data in a supervised way. It is jointly optimized with the re-identification branch. The loss function of the defogging branch $\mathcal{L}_{DF_s}$ can be expressed as follows.

$$\mathcal{L}_{DF_s} = \frac{1}{Q}\sum_{i=1}^{Q} \|J_i - J_i^{GT}\|_2. \tag{4}$$

where $\|\cdot\|_2$ presents the $L2$ norm. J_i and J_i^{GT} represent the i^{th} predicted fog-free image and the corresponding ground truth in a batch, respectively. The total loss at this stage $\mathcal{L}_{Supervised}$ is:

$$\mathcal{L}_{Supervised} = \mathcal{L}_{Tri} + \mathcal{L}_{ID} + \lambda_1 \mathcal{L}_{DF_s}. \tag{5}$$

Unsupervised Learning Stage. To address the domain gap between the real-world and the synthetic data, we propose unsupervised defogging learning to optimize the defogging branch without the ground truth of foggy images. Four losses are involved at this stage: (1) the color entropy loss ($\mathcal{L}_{CE}$); (2) dark channel loss ($\mathcal{L}_{DC}$); (3) total variation loss ($\mathcal{L}_{TV}$); (4) self-constraint loss ($\mathcal{L}_{SC}$).

The former three losses aim to resolve the low-contrast, noise, and residual fog problem based on image priors. They enforce the defogging branch to learn the images that have the same statistical properties as clean images. First, the color entropy loss can enhance the contrast of the recovered images. It is defined as:

$$\mathcal{L}_{CE} = -\frac{1}{Q}\sum_{i=1}^{Q}\sum_{k=0}^{255} H^k(J_i) log(H^k(J_i)), \tag{6}$$

where $H^k(\cdot)$ denotes the normalized histogram counts at value k. We can assume that the fog-free image generally

contains clear image content and vivid color. The color variation may tend to have a higher value for a well-defogged result. On the other hand, an inappropriate fog removal output may usually have a lower value of color variation. Therefore, the color entropy can be considered as a metric to evaluate the clearness of the defogged result.

Second, the dark channel operation (He, Sun, and Tang 2010) has been proved as an effective metric to represent the density of fog (Chen, Ding, and Kuo 2019; Tang, Yang, and Wang 2014). It can be defined:

$$DC(J)(x) = \min_{y \in \Omega(x)} \left(\min_{c \in \{r,g,b\}} J^c(y) \right), \tag{7}$$

where $DC(\cdot)$ is the dark channel operation, $J^c(y)$ is the intensity in the color channel c, and $\Omega(x)$ is a local patch with a fixed size centered at x. He *et al.* (He, Sun, and Tang 2010) observed that, for most pixels in natural images, min $\{J^R(x), J^G(x), J^B(x)\}$ is close to zero. Thus, we apply the dark channel loss $\mathcal{L}_{DC}$ to constrain the defogged image to have less residual fog. Its definition is:

$$\mathcal{L}_{DC} = \frac{1}{Q} \sum_{i=1}^{Q} \|DC(J_i)\|_1. \tag{8}$$

By reducing the dark channel value in a recovered image, a desirable fog-free image is obtained.

Third, the total variation loss focuses on suppressing the noise while preserving the image content and structural information. It is defined as:

$$\mathcal{L}_{TV} = \frac{1}{Q} \sum_{i=1}^{Q} \| \bigtriangledown_x J_i\|_1 + \| \bigtriangledown_y J_i\|_1. \tag{9}$$

where $\bigtriangledown_x$ and $\bigtriangledown_y$ denote the gradient operation along the horizontal and vertical direction. $\|\cdot\|_1$ denotes the $L1$ norm.

Last, though the $\mathcal{L}_{CE}$, $\mathcal{L}_{DC}$, and $\mathcal{L}_{TV}$ can enhance the image quality effectively, the network may still need further optimization. Specifically, undesirable results may be produced to achieve lower values of three loss functions because they are not optimized with the ground truths of fog-free images. It may degrade the performance of the CFSM and the fog-free image reconstruction module on real-world images. Thus, to solve this issue, we applied the self-constraint loss to prevent the network from learning undesired features. The self-constraint loss $\mathcal{L}_{SC}$ is:

$$\mathcal{L}_{SC} = -\frac{1}{Q} \sum_{i=1}^{Q} \frac{\langle \mathcal{F}(J_i), \mathcal{F}(I_i) \rangle}{\|\mathcal{F}(J_i)\|_2 \|\mathcal{F}(I_i)\|_2}, \tag{10}$$

where I_i, $\mathcal{F}(\cdot)$, $\langle , \rangle$ denote the i^{th} input foggy image, Fourier transformation, and dot product, respectively. Our idea is that, in the Fourier domain, the amplitude component usually contains the style information while the phase component often contains the structural and content information (Yang and Soatto 2020; Yang et al. 2020). Based on previous literature (Li et al. 2018b), image defogging can be treated as a process of style transformation, which implies that the content information of the defogged result should be similar to that of the input. Thus, by using the self-constraint loss, the structural information can be constrained and the network can be prevented from generating undesired results.

Figure 4: Examples of the images in FVRID_Real and FVRID_Syn datasets for vehicle ReID.

Finally, the loss function for the defogging branch $\mathcal{L}_{DFu}$ and the total loss $\mathcal{L}_{Unsupervised}$ at the unsupervised stage can be presented as:

$$\mathcal{L}_{DFu} = \mathcal{L}_{CE} + \lambda_2 \mathcal{L}_{DC} + \lambda_3 \mathcal{L}_{TV} + \lambda_4 \mathcal{L}_{SC}, \tag{11}$$

and

$$\mathcal{L}_{Unsupervised} = \mathcal{L}_{DFu} + \mathcal{L}_{Tri} + \mathcal{L}_{ID}. \tag{12}$$

Experiments

Dataset and Evaluation Protocols

Real-world Dataset. For the real-world dataset, we investigate all existing benchmarks and find that only VERI-Wild and Vehicle-1M datasets contain the cases in the foggy weather. Thus, in our experiments, the real-world dataset is constituted based on these two datasets and we called it the Foggy Vehicle ReID for real-world scenes (FVRID_Real) dataset. We carefully pick the vehicle images in the foggy scenarios from the two datasets and organize these images to the dataset. The details of this dataset are presented in Table 1 and the lower part of Fig. 4. We leverage this dataset for unsupervised defogging learning in our network.

Synthetic Dataset. Due to the limited number of vehicle ReID data in foggy weather, to train the proposed network, we construct a synthetic training set called FVRID_Syn for the training process. We select fog-free images from VERI-Wild and Vehicle-1M datasets. Then, we synthesize these images based on the fog synthesis process in (Li et al. 2018a). First, we apply the (Liu et al. 2015) to estimate the depth map d for each image. Second, based on these depth maps, we synthesize the fog on these clear images by (1). We set $\beta \in [0.4, 1.6]$ and $A \in [0.5, 1]$. The examples and detailed constitution of this dataset are shown in the upper part of Fig. 4 and Table 2, respectively. We leverage this dataset for the supervised defogging learning in our network.

Evaluation Protocols. The experiments are performed on our FVRID_Syn and FVRID_Real datasets. We follow the protocol proposed in (Lou et al. 2019; Guo et al. 2018) for

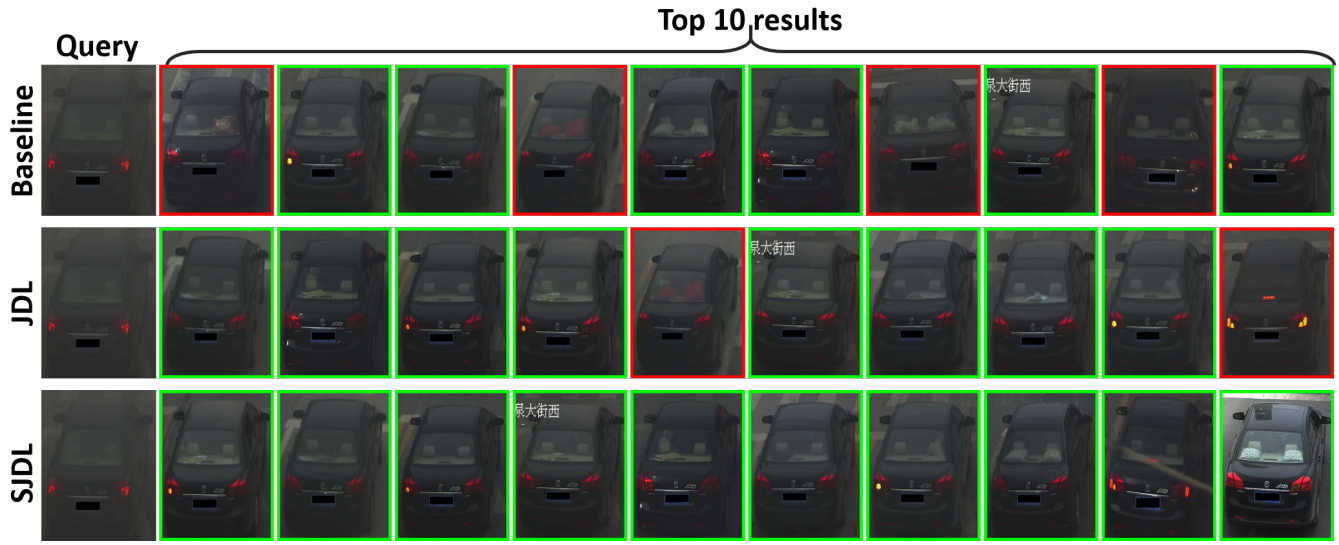


Figure 5: Visualization of the ranking list on FVRID_Real dataset. The images in the first column are the query images while the rest are retrieved top-10 ranking results. The false instances are on the red border while correct retrieved images are on the green border.

Figure 6: Defogged results based on different optimization schemes.

Set	Train	Probe	Gallery
VERI-Wild	156/2472	389/389	389/5985
Vehicle-1M	247/2579	611/611	611/6242
FVRID_Real	403/5051	1000/1000	1000/12227

Table 1: The detailed constitution of the FVRID_Real dataset. (IDs/Images)

Q3:
What is the author's evaluation of the solution

evaluation. Specifically, we randomly select one foggy image for each vehicle and put it into the probe set. The remained images form the gallery set. We apply the cumulative matching characteristic (CMC) curve and mean average precision (mAP) to evaluate the performance.

Implementation Details

Training Stage. For the re-identification branch, the ResNet-50 is adopted as the backbone, whose weights are initialized from the model pre-trained on the ImageNet. We apply two ID classifications and the dimensions of these FC layers are set to 3000 (synthetic) and 403 (real-world), respectively. The weights of the restoration branch are initialized by Kaiming normalization (He et al. 2015). The whole network is trained in an end-to-end fashion based on the training sets of FVRID_Syn and FVRID_Real for learning defogging, vehicle ReID and ID classification simultaneously. The input image is resized to 384 $\times$ 384 and the training batch size Q is set to 36. We apply horizontal flip and random crop to prevent the overfitting problem due to the limited number of training data. We train models for 120 epochs with a warm-up strategy. The initial learning rate is 1.09×10^{-5}, which increases to 10^{-4} after the 10^{th} epoch. The Adam optimizer is adopted to optimize the model with a decay rate of 0.6. The hyper-parameters λ_1, λ_2, λ_3, and λ_4 are set as to 1, 10^{-5}, 10^{-5}, and 300. The network is trained on an Nvidia Tesla V100 GPU for 20 hours and we implement it on the Pytorch platform.

weight
initailize

Set	Train	Probe	Gallery
VERI-Wild	1167/19532	389/389	389/6125
Vehicle-1M	1833/23026	611/611	611/7093
FVRID_Syn	3000/42558	1000/1000	1000/13218

Table 2: The detailed constitution of the FVRID_Syn dataset. (IDs/Images)

Inference Stage. At the inference stage, the defogging branch is not involved. The computational burden caused by it can be ignored. We calculate the Euclidean distance D through embedding features to evaluate the performance.

Comparison with the Existing Methods

We compare our method with other existing ReID methods, including the original Triplet (Hermans, Beyer, and Leibe 2017), the original VRCF (Gao et al. 2020), the original VOC (Zhu et al. 2020), the original VEHICLEX (Yao

Method	mAP		CMC@1		CMC@5		CMC@10	
	S	R	S	R	S	R	S	R
Triplet	35.70	36.10	65.10	60.30	82.00	79.20	87.80	85.10
Triplet-defog	51.20	39.00	76.80	62.10	90.60	80.80	94.00	86.30
Triplet-fog	69.10	52.80	87.80	72.50	95.60	89.40	97.80	94.20
VRCF	25.90	36.60	61.70	63.70	76.50	78.80	81.30	83.20
VRCF-defog	61.50	50.80	85.40	78.00	95.10	92.00	97.20	95.40
VRCF-fog	69.00	58.00	88.60	81.10	97.60	93.80	98.40	96.80
VOC	59.70	57.40	86.10	82.80	94.30	94.00	95.60	96.60
VOC-defog	63.40	49.20	87.00	74.10	94.80	89.90	96.50	94.30
VOC-fog	67.10	59.90	88.70	83.50	95.10	94.00	96.50	97.20
VEHICLEX	63.64	61.56	86.50	83.20	95.00	95.20	97.40	97.90
VEHICLEX-defog	73.06	64.82	89.70	83.90	96.70	95.10	98.20	97.60
VEHICLEX-fog	77.86	69.01	91.20	84.80	97.10	96.10	98.70	98.10
DMT	73.90	71.70	93.40	93.20	97.20	97.40	97.90	98.50
DMT-defog	75.10	71.60	93.40	92.40	96.90	97.50	98.30	98.40
DMT-fog	77.30	73.40	94.00	93.40	97.60	97.60	98.60	98.80
PVEN	72.83	75.36	63.73	66.48	84.39	86.53	89.65	91.20
PVEN-defog	81.70	78.13	73.29	69.47	92.50	89.16	96.04	93.43
PVEN-fog	84.55	81.92	76.60	74.09	95.02	92.15	97.84	95.66
TransReID	62.90	64.00	82.40	77.70	92.30	88.80	98.40	94.00
TransReID-defog	66.80	65.30	83.00	76.60	94.10	89.90	98.10	94.60
TransReID-fog	73.90	72.10	84.80	82.60	95.20	90.70	98.70	95.60
Ours	**85.36**	**82.70**	**94.60**	**94.60**	**97.90**	**98.10**	**98.90**	**99.20**

Table 3: Quantitative evaluation on the foggy ReID datasets. The texts 'S' and 'R' denote FVRID_Syn and FVRID_Real datasets. (The FVRID_Real dataset was constructed from the foggy images in VERI-Wild and Vehicle-1M datasets) The words with boldface indicate the best results, and the words with underline indicate the second-best results.

Module	mAP			CMC@1			CMC@5			CMC@10		
	S	R	Δ	S	R	Δ	S	R	Δ	S	R	Δ
Baseline	81.88	76.17	5.71	94.40	93.40	1.0	97.60	97.50	0.5	98.70	98.50	0.2
JDL	83.04	79.47	3.57	94.50	93.80	0.7	**98.10**	97.90	0.2	98.80	99.00	-0.2
SJDL w/o $\mathcal{L}_{SC}$	84.39	81.50	2.89	94.50	94.20	0.3	98.00	98.00	0.0	98.80	99.00	-0.2
SJDL	**85.36**	**82.70**	**2.66**	**94.60**	**94.60**	**0.0**	97.90	**98.10**	**-0.2**	**98.90**	**99.20**	**-0.3**

Table 4: Effectiveness of the proposed joint defogging learning and semi-supervised defogging optimization. The text 'JDL' denotes the joint defogging learning only with supervised defogging optimization, while the text 'SJDL' presents the JDL mechanism with the semi-supervised defogging optimization, respectively. The 'Baseline' presents the ResNet-50. The symbol 'Δ' presents the difference between the results of synthetic data and real-world data (The smaller value indicates better performance for addressing the domain gap problem).

Method	mAP		CMC@1		CMC@5		CMC@10	
	S	R	S	R	S	R	S	R
Conv_2	**85.36**	**82.70**	**94.60**	**94.60**	**97.90**	**98.10**	**98.90**	**99.20**
Conv_3	84.93	81.94	94.60	94.30	97.90	98.00	98.90	99.10
Conv_4	84.76	81.57	94.40	94.10	97.90	98.00	98.70	99.00
Conv_5	82.57	79.39	93.70	94.10	97.80	97.80	98.50	99.00

Table 5: Comparison of performance for using different blocks as collective feature sharing module.

et al. 2020), the original DMT (He et al. 2020), the original PVEN (Meng et al. 2020), original TransReID (He et al. 2021). We also retrained these models in the foggy scenarios using the same training sets as the proposed one and the obtained ReID models are denoted by Triplet-fog, VRCF-fog, VOC-fog, VEHICLEX-fog, DMT-fog, PVEN-fog, and TransReID-fog, respectively. Moreover, Triplet-defog, VRCF-defog, VOC-defog, VEHICLEX-defog, DMT-defog, PVEN-defog, and TransReID-defog denote the two-stage solutions that are the combinations of the defogging method with the original ReID models. The adopted defogging method is MPR-Net (Zamir et al. 2021). The results are reported in Table 3, which show that the proposed method can achieve the best performance on vehicle ReID in foggy

Q3: experiments are presented in support of the idea

Q4: What are the most interesting or controversial ideas?

weather on FVRID_Syn and FVRID_Real datasets in terms of mAP and CMC. The proposed algorithm outperforms existing ReID models, no matter whether they are training in the foggy scenarios or combined with a defogging method.

Ablation Studies

Effectiveness of the Joint Defogging Learning. Table 4 presents the effectiveness of the proposed joint defogging learning strategy. We also present the visualization of the ranking list on FVRID_Real dataset in Fig. 5. One can see that, the baseline may retrieve the wrong instance because the crucial features such as light and window may become ambiguous, which may degrade the feature extraction of the network. However, with the proposed JDL, clear features can be extracted and the performance vehicle ReID in the foggy weather can be improved. Furthermore, we present the comparison of using different convolution blocks as the CFSM in the ReID backbone in Table 5. One can see that using Conv_2 can achieve the best performance.

Effectiveness of the Semi-supervised Optimization. The experiments in Table 4 and Fig. 5 show that better performance on real-world scenarios can be achieved if semi-supervised learning is applied. Specifically, compared to the baseline and the module only applies supervised learning (i.e., JDL), the performance is improved. Moreover, with semi-supervised learning, the domain gap between real-world and synthetic scenarios is reduced effectively, that is, the Δ value in Table 4 is decreased in each metric. Based on the aforementioned results, the proposed semi-supervised defogging training technique can mitigate the domain problem between real-world and synthetic scenarios. Moreover, in Fig. 6, we present a visual comparison of using different optimization schemes. One can see that the proposed semi-supervised optimization can generate more desirable defogged results in real-world scenarios, which may further benefit the performance of vehicle ReID.

In Table 4, we evaluate the effectiveness of using enhancement loss (i.e., $\mathcal{L}_{DC}$, $\mathcal{L}_{TV}$, and $\mathcal{L}_{CE}$) and the self-constraint loss $\mathcal{L}_{SC}$ in unsupervised branch. One can see that, with the use of three enhancement losses (i.e., SJDL w/o $\mathcal{L}_{SC}$), the ReID performance can be improved compared to the JDL module. Moreover, the usage of $\mathcal{L}_{SC}$ can preserve the content and textural information in the recovered results, which is beneficial to defogging and vehicle ReID simultaneously.

Conclusion

In this paper, to alleviate the vehicle ReID problem in the foggy weather, we proposed a semi-supervised joint defogging learning (SJDL) system that can conduct defogging and vehicle ReID simultaneously. Moreover, this framework can solve the performance gap between real-world and synthetic scenarios, Furthermore, to train the proposed network, we construct a dataset called FVRID which contains synthetic and real-world foggy images. Experimental results indicate the proposed method can achieve superior performance than existing methods and each of the proposed modules contributes to the performance of the network.

Acknowledgements

We thank to National Center for High-performance Computing (NCHC) for providing computational and storage resources. This research was supported by the Ministry of Science and Technology, Taiwan under Grants MOST 108-2221-E-002-072-MY3 and MOST 108-2638-E-002-002-MY2. Portions of the research in this paper use the Vehicle-1M dataset collected under the sponsor of the National Natural Science Foundation of China.

References

Berman, D.; Avidan, S.; et al. 2016. Non-local image dehazing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1674–1682.

Chen, L. 2005. The topological approach to perceptual organization. *Visual Cognition*, 12(4): 553–637.

Chen, T.-S.; Liu, C.-T.; Wu, C.-W.; and Chien, S.-Y. 2020. Orientation-aware vehicle re-identification with semantics-guided part attention network. In *European Conference on Computer Vision*, 330–346. Springer.

Chen, W.-T.; Ding, J.-J.; and Kuo, S.-Y. 2019. PMS-Net: Robust haze removal based on patch map for single images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11681–11689.

Chen, W.-T.; Lou, H.-L.; Fang, H.-Y.; Chen, I.-H.; Chen, Y.-W.; Ding, J.-J.; and Kuo, S.-Y. 2021. DesmokeNet: A Two-stage Smoke Removal Pipeline Based on Self-Attentive Feature Consensus and Multi-Level Contrastive Regularization. *IEEE Transactions on Circuits and Systems for Video Technology*.

Dong, H.; Pan, J.; Xiang, L.; Hu, Z.; Zhang, X.; Wang, F.; and Yang, M.-H. 2020. Multi-scale boosted dehazing network with dense feature fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2157–2167.

Gao, C.; Hu, Y.; Zhang, Y.; Yao, R.; Zhou, Y.; and Zhao, J. 2020. Vehicle re-identification based on complementary features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 590–591.

Guo, H.; Zhao, C.; Liu, Z.; Wang, J.; and Lu, H. 2018. Learning coarse-to-fine structured feature embedding for vehicle re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

He, B.; Li, J.; Zhao, Y.; and Tian, Y. 2019. Part-regularized near-duplicate vehicle re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3997–4005.

He, K.; Sun, J.; and Tang, X. 2010. Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence*, 33(12): 2341–2353.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, 1026–1034.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

He, S.; Luo, H.; Chen, W.; Zhang, M.; Zhang, Y.; Wang, F.; Li, H.; and Jiang, W. 2020. Multi-domain learning and identity mining for vehicle re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 582–583.

He, S.; Luo, H.; Wang, P.; Wang, F.; Li, H.; and Jiang, W. 2021. Transreid: Transformer-based object re-identification. *arXiv preprint arXiv:2102.04378*.

Hermans, A.; Beyer, L.; and Leibe, B. 2017. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*.

Hui, Z.; Li, J.; Wang, X.; and Gao, X. 2020. Image fine-grained inpainting. *arXiv preprint arXiv:2002.02609*.

Khorramshahi, P.; Peri, N.; Chen, J.-c.; and Chellappa, R. 2020. The devil is in the details: Self-supervised attention for vehicle re-identification. In *European Conference on Computer Vision*, 369–386. Springer.

Li, B.; Ren, W.; Fu, D.; Tao, D.; Feng, D.; Zeng, W.; and Wang, Z. 2018a. Benchmarking single-image dehazing and beyond. *IEEE Transactions on Image Processing*, 28(1): 492–505.

Li, R.; Pan, J.; Li, Z.; and Tang, J. 2018b. Single image dehazing via conditional generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8202–8211.

Liu, F.; Shen, C.; Lin, G.; and Reid, I. 2015. Learning depth from single monocular images using deep convolutional neural fields. *IEEE transactions on pattern analysis and machine intelligence*, 38(10): 2024–2039.

Liu, H.; Tian, Y.; Wang, Y.; Pang, L.; and Huang, T. 2016. Deep Relative Distance Learning: Tell the Difference Between Similar Vehicles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2167–2175.

Liu, X.; Liu, W.; Mei, T.; and Ma, H. 2017. Provid: Progressive and multimodal vehicle reidentification for large-scale urban surveillance. *IEEE Transactions on Multimedia*, 20(3): 645–658.

Lou, Y.; Bai, Y.; Liu, J.; Wang, S.; and Duan, L. 2019. Veri-wild: A large dataset and a new method for vehicle re-identification in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3235–3243.

Meng, D.; Li, L.; Liu, X.; Li, Y.; Yang, S.; Zha, Z.-J.; Gao, X.; Wang, S.; and Huang, Q. 2020. Parsing-based view-aware embedding network for vehicle re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7103–7112.

Narasimhan, S. G.; and Nayar, S. K. 2003. Contrast restoration of weather degraded images. *IEEE transactions on pattern analysis and machine intelligence*, 25(6): 713–724.

Nayar, S. K.; and Narasimhan, S. G. 1999. Vision in bad weather. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, 820–827. IEEE.

Qu, Y.; Chen, Y.; Huang, J.; and Xie, Y. 2019. Enhanced Pix2pix Dehazing Network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shen, Y.; Xiao, T.; Li, H.; Yi, S.; and Wang, X. 2017. Learning deep neural networks for vehicle re-id with visual-spatio-temporal path proposals. In *Proceedings of the IEEE International Conference on Computer Vision*, 1900–1909.

Tang, K.; Yang, J.; and Wang, J. 2014. Investigating haze-relevant features in a learning framework for image dehazing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2995–3000.

Yang, H.-H.; Chen, W.-T.; Luo, H.-L.; and Kuo, S.-Y. 2021. Multi-modal Bifurcated Network for Depth Guided Image Relighting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

Yang, Y.; Lao, D.; Sundaramoorthi, G.; and Soatto, S. 2020. Phase consistent ecological domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9011–9020.

Yang, Y.; and Soatto, S. 2020. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4085–4095.

Yao, Y.; Zheng, L.; Yang, X.; Naphade, M.; and Gedeon, T. 2020. Simulating content consistent vehicle datasets with attribute descent. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, 775–791. Springer.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; Yang, M.-H.; and Shao, L. 2021. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14821–14831.

Zhang, H.; and Patel, V. M. 2018. Densely connected pyramid dehazing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3194–3203.

Zhao, H.; Shi, J.; Qi, X.; Wang, X.; and Jia, J. 2017. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2881–2890.

Zheng, A.; Lin, X.; Li, C.; He, R.; and Tang, J. 2019. Attributes guided feature learning for vehicle re-identification. *arXiv preprint arXiv:1905.08997*.

Zhou, Y.; and Shao, L. 2018. Aware attentive multi-view inference for vehicle re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6489–6498.

Zhu, Q.; Mai, J.; and Shao, L. 2015. A fast single image haze removal algorithm using color attenuation prior. *IEEE transactions on image processing*, 24(11): 3522–3533.

Zhu, X.; Luo, Z.; Fu, P.; and Ji, X. 2020. VOC-ReID: Vehicle re-identification based on vehicle-orientation-camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 602–603.