FoggyFormer: Learning Spatial-Channel Enhanced Transformer for Object Categorization under Foggy Conditions

Qi Bi Shaodi You Theo Gevers
Computer Vision Research Group, University of Amsterdam
Amsterdam, the Netherlands
{q.bi, s.you, th.gevers}@uva.nl

Abstract

Fog represents a customary meteorological phenomenon capable of exerting a detrimental impact on the efficacy of computer vision tasks like autonomous driving. Categorizing fog within images poses considerable challenges, primarily due to the obscuration of key objects by fog, thereby impeding the determination of their respective categories. This obscuration, coupled with the resultant reduction in contrast, further compounds the complexity of the categorization. Therefore, in this paper, we propose a novel vision Transformer-based network, named FoggyFormer, to address the aforementioned challenges. FoggyFormer incorporates a distinctive spatial-channel enhancement strategy to enhance the performance in foggy image categorization tasks. Firstly, the channel-wise enhancement technique mitigates the adverse effects of low contrast prevalent in foggy images. Secondly, the spatial-wise enhancement employs a spatial attention mechanism to discern and emphasize the spatial positions of crucial objects within the foggy images.

To facilitate a comprehensive investigation of the task at hand, we introduce the Foggy-CODaN dataset, a meticulously curated collection comprising 10,000 samples distributed across 10 distinct categories. By employing this dataset, we aim to establish a solid foundation for the systematic examination of foggy image categorization algorithms. We demonstrate that the proposed FoggyFormer surpasses the performance of existing vision Transformer models, thereby achieving state-of-the-art results. The experimental results substantiate the model's robust generalization capabilities, as evidenced by its proficient handling of unlabeled real-world driving scenes.

1. Introduction

Vision in adverse weather conditions such as rain, snow, and haze plays a pivotal role in numerous safety-critical applications, including autonomous driving [3, 4, 23]. Nevertheless, existing models face substantial challenges in this domain due to their reliance on training data that predominantly consists of well-conditioned images [21, 22]. Consequently, these models struggle to generalize effectively when confronted with real-world images exhibiting unfavorable weather conditions [3]. Amongst these challenging conditions, fog poses one of the most formidable obstacles [19].

Visibility enhancement techniques, encompassing dehazing, deraining, and desnowing, have been widely employed as straightforward approaches in addressing adverse weather conditions in images [15, 17, 35, 40]. However, recent studies have revealed that image enhancement methodologies have the potential to alter the content of the original image, thereby potentially impeding high-level computer vision tasks such as categorization [14, 32, 39]. Consequently, in contrast to the conventional visibility enhancement paradigm, we propose an alternative approach that focuses on learning a robust representation capable of facilitating high-level vision tasks in challenging weather conditions.

In this paper, the focus is on the visual categorization of foggy conditions. The task has significant challenges due to the presence of extremely low contrast and the obscuration of key visual cues, such as objects, by the fog (as depicted in Fig. 1.(a)) [27]. Consequently, the direct learning of a robust representation becomes problematic [14, 32, 39].

Hence, we propose a spatial-channel enhanced vision Transformer named FoggyFormer. A channel-wise enhancement strategy is proposed to learn representative features from low contrast image parts. A spatial-wise enhancement is introduced to focus on regions belonging to scene semantics.

Our proposed task is novel and existing datasets [1, 26, 27] cannot be used to evaluate the task at hand. Therefore, a novel dataset Foggy-CODaN is created. The Foggy-CODaN dataset is built based on the Common Objects Day and Night (CODaN) dataset [12], which contains 10,000 samples from 10 categories in total. The original data is

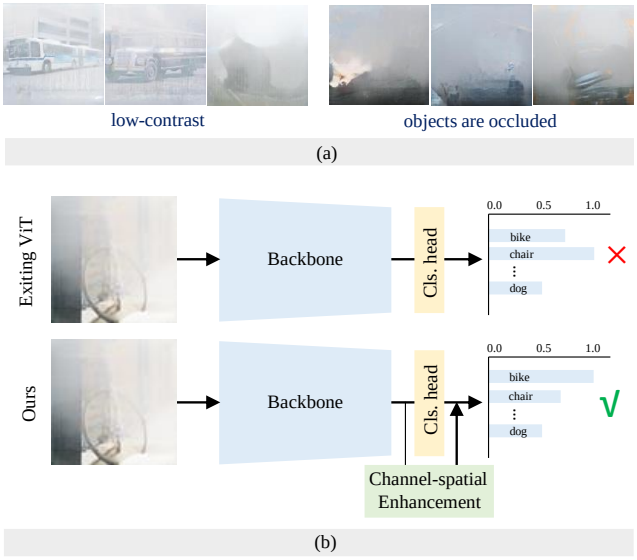

Figure 1. (a) Challenges for visual categorization for foggy images. Low contrast and fuzziness. (b) Pipeline difference between existing ViT and the proposed FoggyFormer.

modified using Foggy-CycleGAN [37]. We change the foggy density and provide photo-realistic foggy rendering effects. Extensive experiments are conducted and show that the proposed method state-of-the-art performance.

Our contribution is summarized as follows:

- A novel network FoggyFormer is proposed for visual categorization considering strong foggy conditions.
- A spatial-channel enhancement strategy is proposed for vision Transformer to learn robust semantics from foggy images.
- A novel Foggy-CODaN dataset is created which enables a systematic study of the new task.
- Extensive experiments show state-of-the-art performance and good generalization of the proposed FoggyFormer.

2. Related Work

Image De-hazing & De-fogging In the past decade, methods are proposed to leverage convolutional neural networks (CNNs) for image de-hazing tasks. Some typical works include MSCNN [25], DehazeNet [2], AOD-Net [13], FAMED-Net [38], EPDN [24], and *etc*. More recently, due to the introduction of vision Transformer, feature representation are proposed for adverse weather removal [6,18,34]. However, works that focus on fog removal are still relatively few so far [8,30].

Also, attention is paid to the development of model-driven image de-fogging algorithms. Specially, Jiang *et al*. [7] leverage depth priors for image fog removal. Liang *et al*. [16] propose a defogging method based on bilateral hybrid filtering. Hu *et al*. [5] propose an iterative de-fogging framework with a physical prior of illumination decomposition. Kang *et al*. [9] introduce a hyperspectral image de-fogging pipeline by modeling pre-band reflectance. Liu *et al*. create a large-scale real-captured image de-fog dataset [19]. Zhou *et al*. propose a de-hazing framework by using a polarization light model [40].

Vision Tasks under Foggy Conditions There are multiple benchmarks for vision tasks under haze conditions. For example, *UG 2+ Challenge Track 2* is a 2-D object detection dataset under hazy conditions [20,33]. Five object categories, namely, car, bus, bicycle, motorcycle and pedestrian, are considered. *NYU-depth-V2* [28] is a depth estimation dataset. Its synthesized version [11] is especially designed for the estimation of haze. However, a dataset, focusing on foggy conditions, is still rare.

For foggy conditions, Sakaridis *et al*. [26] focus on semantic segmentation and object detection in foggy scenarios, and proposed a Foggy CityScapes dataset. More recently, a semantic segmentation dataset from real-world scenarios, Adverse Conditions Dataset with Correspondences (ACDC) [27], is proposed. It can benchmark the semantic segmentation not only for foggy conditions, but also for rain, snow and nigh-time. *PTAW172Real* [1] is a person tracking dataset of real-world sequences, under fog, rain and snow conditions.

However, so far, there is no dataset available that can be used to systematically benchmark the performance of methods for foggy conditions.

Vision Transformer Due to the self-attention mechanism to capture long-dependencies and high-level semantic information, it is shown that vision Transformer (ViT) have stronger feature representation capabilities than convolutional neural networks (CNN) [31]. A variety of advanced vision Transformer models are proposed, *e.g*., BiT [10], DeiT [29], Metaformer [36]. More recently, Swin Transformer [22] and its updated versions [21] demonstrate stronger feature representation capabilities [31].

In conclusion, so far, only a few methods investigated the performance of vision Transformers under foggy conditions.

3. Foggy-CODaN Dataset

3.1. Dataset Overview & Statistics

We propose a Foggy-CODaN dataset to benchmark the new task. It uses the Common Objects Day and Night (CODaN) dataset [12] and rendered by Foggy-CycleGAN.

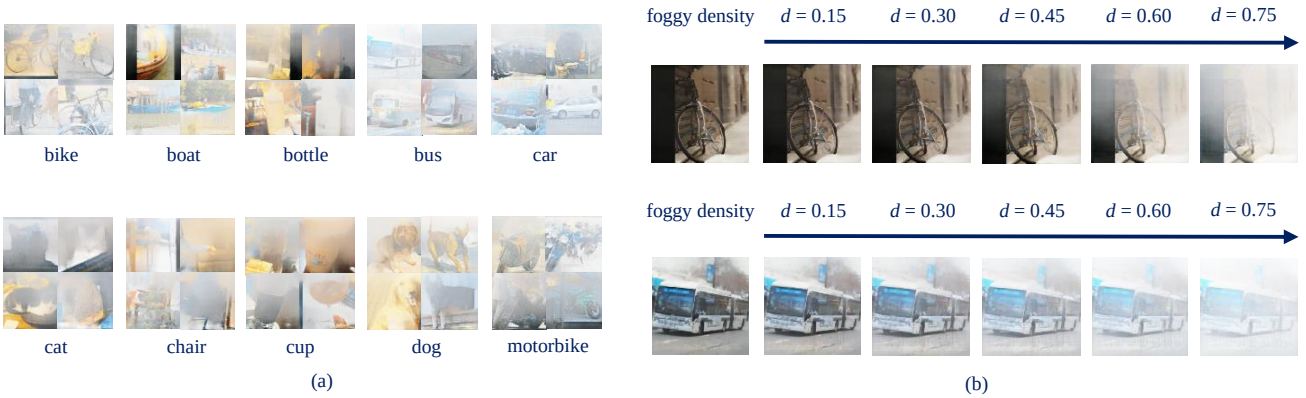

Figure 2. (a) Some selected samples for each category in the Foggy-CoDaN dataset, namely, *bycicle*, *boat*, *bottle*, *bus*, *car*, *cat*, *chair*, *cup*, *dog* and *motorbike*; (b) illustration of modifying the foggy density parameter and its visual effect on the rendered foggy images.

Common Objects Day and Night (CODaN) is an image categorization dataset, which consists of 10 common object categories from a variety of day and night conditions [12]. For each category, it has 1,000 training images, and 600 testing images in total.

Ten object categories in this dataset are *bycicle*, *boat*, *bottle*, *bus*, *car*, *cat*, *chair*, *cup*, *dog* and *motorbike*, respectively.

3.2. Rendering Details

Foggy-CycleGAN [37] uses CycleGAN [41] to synthesize foggy conditions in images. Apart from the default parameters of CycleGAN, an additional parameter *foggy density* (denoted by d) steers the density of the fog in an image. We use the official software package [1] of the Foggy-CycleGAN [37] for rendering.

Some examples with changing the *foggy density* d on the same image are shown in Fig. 2. It is shown that a change in parameter d corresponds to the fog density in the image. The higher d, the stronger the foggy effect is in the image.

3.3. Evaluation Metrics

Both overall accuracy (OA) and average accuracy (AA) are used as the evaluation metrics.

The overall accuracy is the ratio between the number of correctly categorized images and the number of all the images. The average accuracy is the average of the per-category accuracy. For per-category accuracy, it is the ratio between the number of correctly categorized images of a certain category and the number of the images in this category.

4. Methodology

4.1. Framework Overview

Fig. 3 shows the proposed spatial-channel enhanced Transformer for visual categorization under the foggy condition. After feature extraction, using Swin-Tiny as backbone [22], the method consists of three components, namely, channel-wise enhancement, spatial-wise enhancement, and channel-spatial enhanced feature fusion.

4.2. Channel-wise Enhancement

Given the low contrast in foggy images, the challenge is how to enhance the features of the key objects. To this end, we propose a channel-wise enhancement strategy for the foggy images. The first step is to normalize the feature representation in a channel-wise manner, so that the channel-wise difference are magnified. The second step is to implement channel-wise attention of the normalized feature maps, so that the per-channel contribution of the feature maps, determining the object category, can be measured.

Specifically, given the feature representation $\mathbf{X} \in \mathbb{R}^{(W \cdot H) \times C}$ from the backbone, for the first step, an instance normalization is conducted, denoted by

$$\mathbf{X}'^{(W \cdot H),c} = \frac{\mathbf{X}^{(W \cdot H),c} - \mu}{\sigma + \epsilon} \cdot \gamma + \beta, \tag{1}$$

$$\mu = \frac{1}{C}\sum_{c=1}^{C} \mathbf{X}^{(W \cdot H),c}, \sigma = \sqrt{\frac{1}{C}\sum_{i=1}^{C}(\mathbf{X}^{(W \cdot H),c} - \mu)^2}, \tag{2}$$

where $c = 1, 2, \cdots, C$. Here W, H and C denote the width, height and channel size of the feature map $\mathbf{X}$.

For the second step, a channel-wise attention is computed on the refined feature map $\mathbf{X}'$, to measure the importance of each channel in determining the scene category.

[1] https://github.com/ghaiszaher/Foggy-CycleGAN

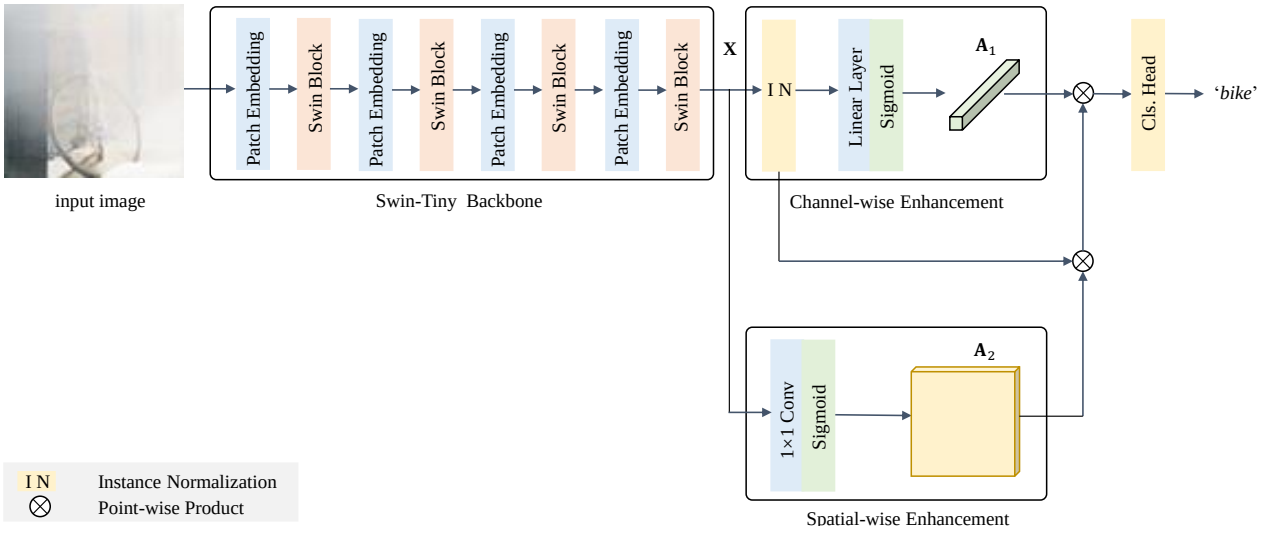

Figure 3. Framework of the proposed Foggy-Former. After feature extraction using Swin-Tiny as backbone, there are three key steps involved. The first step is the channel-wise enhancement, which learns a channel-wise weight matrix to tolerate low contrast in foggy images. The second step is the spatial-wise enhancement, which highlights the spatial position of the key object in images. In the third step, both weight matrices are fused to generate a final prediction.

The channel-wise weight matrix $\mathbf{A}_1 \in \mathbb{R}^{(W \cdot H) \times 1}$ is computed by

$$\mathbf{A}_1 = \text{Sigmoid}(\mathbf{W}_1 \mathbf{X}^{'} + \mathbf{b}_1), \tag{3}$$

where $\mathbf{W}_1 \in \mathbb{R}^{(W \cdot H) \times 1}$ and $\mathbf{b}_1 \in \mathbb{R}^{(W \cdot H) \times 1}$ are the weight and bias matrix of a linear layer. Sigmoid denotes the Sigmoid activation function.

4.3. Spatial-wise Enhancement

Due to the poor visibility caused by fog, usually only a limited number of parts of the key objects may appear in foggy images. Hence, it is particularly important to highlight these regions in foggy images. This objective is realized in a simple and straight-forward way. We reshape the normalized feature $\mathbf{X}' \in \mathbb{R}^{(W \cdot H) \times C}$ into the organization manner of convolutional features, as $\mathbf{X}' \in \mathbb{R}^{W \times H \times C}$. Then, an one-layer spatial attention module is used to extract the attention weight matrix $\mathbf{A}_2 \in \mathbb{R}^{W \times H \times 1}$, which highlights the parts of the key objects. This process is defined by

$$\mathbf{A}_2 = \text{Sigmoid}(\mathbf{W}_2 \otimes \mathbf{X}^{'} + \mathbf{b}_2), \tag{4}$$

where $\otimes$ denotes the convolutional operation function. $\mathbf{W}_2 \in \mathbb{R}^{W \times H \times 1}$ and $\mathbf{b}_2 \in \mathbb{R}^{W \times H \times 1}$ denote the weight and bias matrix of the convolutional layer.

4.4. Channel-spatial Enhanced Feature Fusion

Both the channel-wise and spatial-wise attention matrix are used to enhance the feature representation. In this way, low contrast is tolerated and the spatial positions of the key objects are emphasized. This fusion process is given by

$$\mathbf{X}_{final} = \mathbf{A}_2 \mathbf{A}_1 \mathbf{X}^{'}, \tag{5}$$

where $\mathbf{X}_{final}$ is the final feature representation that is used for visual categorization.

Then, after processing $\mathbf{X}_{final}$ into a category-wise probability vector, the conventional cross-entropy loss function is used to optimize the entire framework in an end-to-end manner.

5. Experiments

5.1. Implementation Details

The Swin-tiny Transformer [22] uses the pre-trained weights of ImageNet as initial parameters. A random initialization is used to initialize the parameters of the rest part of the FoggyFormer.

The experiments are conducted on two GeForce RTX 2080 Ti GPUs. The batch size is set 2 per GPU. The Adam optimizer is used with an initial learning rate of 1×10^{-4}. The weight decay is set 0.05. The training terminates after 50 epochs.

5.2. Comparison with Existing Methods

We compare the proposed FoggyFormer with a number recent vision Transformer models on the proposed Foggy-CoDaN dataset, including ViT [31], BiT [10], DeiT [29], Swin [22] and Swin-V2 [21].

Method	Year	param.	FLOPs	bicycle	boat	bottle	bus	car	cat	chair	cup	dog	motorbike	AA	OA
ViT [31]	2017	307M	190.7G	81.48	95.35	83.67	90.38	90.91	90.00	82.14	85.71	93.33	85.45	87.84	87.53
BiT [10]	2020	44M	8.3G	88.00	94.12	90.70	97.96	91.67	87.50	87.50	86.67	81.13	87.04	89.23	89.13
DeiT [29]	2021	86M	55.4G	94.12	97.83	90.00	**100.00**	90.74	92.31	93.48	91.84	93.75	96.08	94.01	93.96
Swin [22]	2021	197M	103.9G	95.83	97.96	**95.74**	98.04	97.96	90.74	93.75	90.00	97.87	92.59	95.05	94.97
Swin-V2 [21]	2022	197M	104.3G	82.69	95.24	78.43	98.04	95.65	91.67	78.18	84.00	89.80	88.68	88.24	87.93
Ours	2023	198M	104.1G	**96.00**	**100.00**	92.00	96.15	**100.00**	**98.04**	**97.96**	**93.88**	**100.00**	**96.15**	**97.02**	**96.98**

Table 1. Performance comparison of the proposed FoggyFormer with a number of state-of-the-art vision Transformer models on the Foggy-CoDaN dataset. Evaluation metrics include per-category accuracy, average accuracy (AA) and overall accuracy (OA). All these metrics are presented in percentage (%). The comparison is evaluated for an image size of 384×384.

Overall accuracy & average accuracy The last two columns of Table 1 report the overall accuracy (OA) and average accuracy (AA) of the proposed FoggyFormer and the other vision Transformer models.

Our method outperforms the second best method Swin Transformer [22] by 1.97% and 2.01% in terms of the average accuracy and the overall accuracy, and reaches 97.02% and 96.98% in terms of the average accuracy and the overall accuracy. In contrast, ViT [31], BiT [10] and Swin-V2 [21] only achieves an average accuracy of 87.84%, 89.23% and 88.24%, respectively. Also, their overall accuracy metric is 87.53%, 89.13% and 87.93%, which is also far behind the 96.98% achieved by FoggyFormer. The more than 10% accuracy performance gain of the proposed FoggyFormer against the existing vision Transformer models is noteworthy for autonomous driving in foggy conditions.

Per-category performance Table 1 reports the per-category accuracy of the proposed FoggyFormer and the compared vision Transformer models. The proposed FoggyFormer shows the best classification performance on eight out of ten object categories. Compared with the second best classification method, the proposed FoggyFormer shows a performance gain of 0.17% on bicycle, 2.04% on boat, 2.04% on car, 5.73% on cat, 4.21% on chair, 2.04% on cup, 2.13% on dog and 0.07% on motorbike. However, on the bottle and bus category, its performance is somewhat inferior to one or several vision Transformer models.

Notably, the performance gain in the foggy conditions on categories such as car is also important for the application of autonomous driving.

Computational Cost We also compare the proposed FoggyFormer against existing vision Transformer models. It only leads to a 1M parameter number gain and 0.2G increase of FLOPs, when compared with the backbone Swin Transformer. Its parameter number and FLOPs are similar to Swin-V2 [21], and are much smaller than that of ViT [31].

Although BiT [10] and DeiT [29] models have much less parameters and GFLOPs, these two models are actually based on a hyrid CNN-ViT paradigm and use CNN to extract features. Also, their performance in terms of both average accuracy and overall accuracy shows a great decline (>9%) against the proposed FoggyFormer.

Component			Metric (%)	
Backbone	CE	SE	AA	OA
✓			95.05	94.97
✓	✓		95.91	95.86
✓		✓	96.37	96.28
✓	✓	✓	**97.02**	**96.98**

Table 2. Ablation studies for each component of the proposed PN-Former. CE / SE: channel-wise enhancement / spatial-wise enhancement. Evaluation metrics average accuracy (AA) and overall accuracy (OA) are presented in percentage (%).

5.3. Ablation Studies

Table 2 provides an ablation study for each component of the proposed FoggyFormer. The channel-wise enhancement and the spatial-wise enhancement are denoted as CE and SE, respectively.

It is shown that both the channel-wise enhancement and spatial-wise enhancement provides a positive impact on the proposed framework. Compared to the baseline, the channel-wise enhancement leads to a performance gain of 0.86% and 0.89% on average accuracy and overall accuracy, respectively. In contrast, the spatial-wise enhancement leads to a performance gain of 1.32% and 1.31% on average accuracy and overall accuracy, respectively. The joint use of both enhancement leads to a performance gain of 1.97% and 2.01% on average accuracy and overall accuracy, respectively.

5.4. Inference on Real-world Images

It is practical to validate the generalization ability of the proposed FoggyFormer and the proposed Foggy-CoDaN dataset on driving scenarios. To this end, we directly use the FoggyFormer pre-trained on Foggy-CoDaN to infer the category-wise prediction on some un-labeled real-world foggy images. The output is the 10-dimensional probability distribution for each individual category. The results are

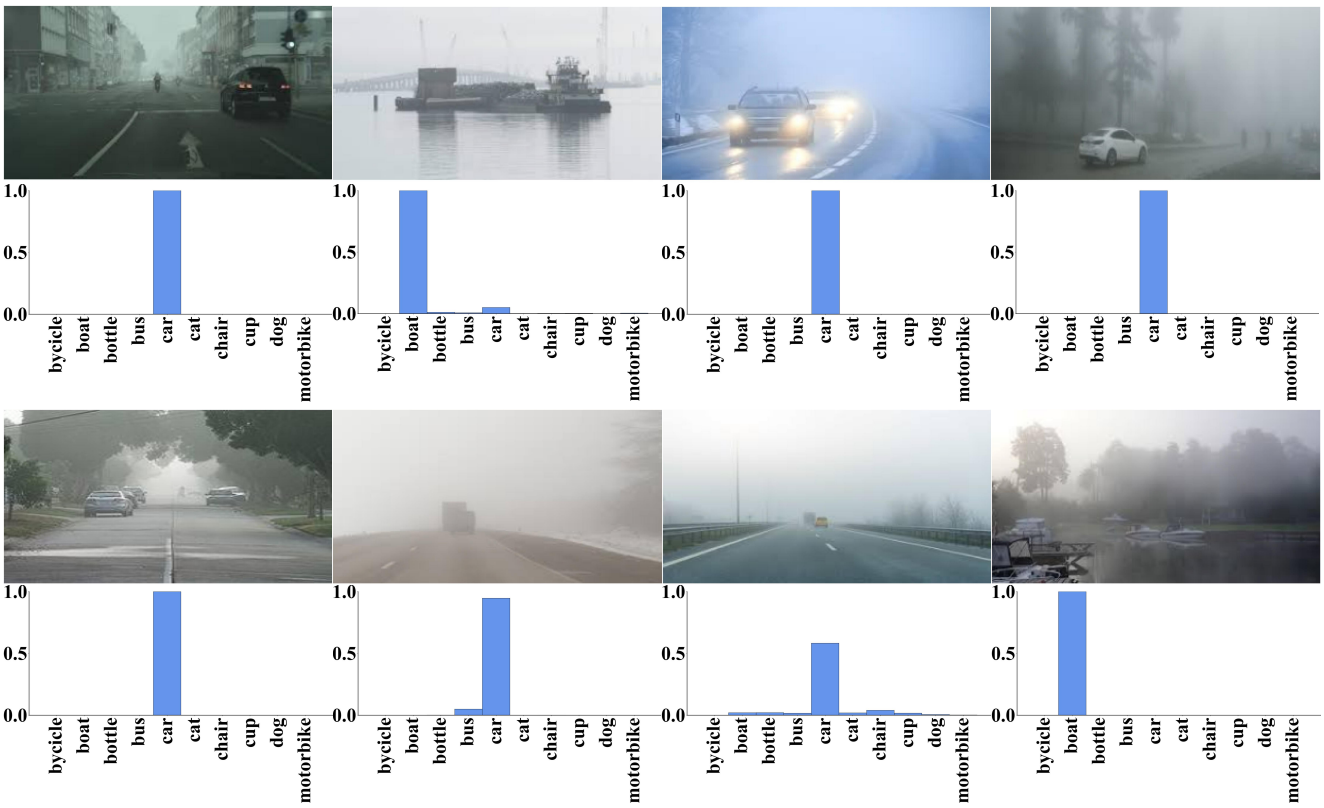

Figure 4. Some real-world un-labeled image inference of the proposed FoggyFormer after pre-trained on the Foggy-CoDaN dataset. The probability distribution of the ten categories is visualized by a histogram.

shown in Fig. 4.

It can be derived that, in both foggy road scenarios and foggy river scenarios, the probability response of car or boat is very high, and is nearly close to 1. In contrast, the other object categories have very little chance to obtain a clear probability estimation. It indicates that the FoggyFormer pre-trained on the proposed Foggy-CoDaN dataset has a good generalization ability on real-world driving scenes.

6. Conclusion

In this paper, the task of visual categorization under the foggy condition is considered. For this task, a FoggyFormer is proposed using a spatial-channel enhancement strategy to highlight the semantic information in foggy images. To benchmark this task, a Foggy-CoDaN dataset is collected. Extensive experiments show state-of-the-art performance of the proposed FoggyFormer. In addition, the reasonable categorization inference on real-world unlabeled images shows good generalization of the FoggyFormer and the Foggy-CoDaN dataset. As future research, the method is further explored on driving scenes under foggy conditions.

References

[1] Kerim Abdulrahman, Celikcan Ufuk, Erdem Erkut, and Erdem Aykut. Using synthetic data for person tracking under adverse weather conditions. *Image and Vision Computing*, page 111, 2020. 1, 2

[2] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao. Dehazenet: An end-to-end system for single image haze removal. *IEEE Transactions on Image Processing*, 25(11):5187–5198, 2016. 2

[3] Sungha Choi, Sanghun Jung, Huiwon Yun, Joanne T Kim, Seungryong Kim, and Jaegul Choo. Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11580–11590, 2021. 1

[4] Carlos A Diaz-Ruiz, Youya Xia, Yurong You, Jose Nino, Junan Chen, Josephine Monica, Xiangyu Chen, Katie Luo, Yan Wang, Marc Emond, et al. Ithaca365: Dataset and driving perception under repeated and challenging weather conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21383–21392, 2022. 1

[5] Hai-Miao Hu, Qiang Guo, Jin Zheng, Hanzi Wang, and Bo Li. Single image defogging based on illumination decompo-

sition for visual maritime surveillance. *IEEE Transactions on Image Processing*, 28(6):2882–2897, 2019. 2

[6] Nanfeng Jiang, Kejian Hu, Ting Zhang, Weiling Chen, Yiwen Xu, and Tiesong Zhao. Deep hybrid model for single image dehazing and detail refinement. *Pattern Recognition*, 136:109227, 2023. 2

[7] Yutong Jiang, Changming Sun, Yu Zhao, and Li Yang. Fog density estimation and image defogging based on surrogate modeling for optical depth. *IEEE Transactions on Image Processing*, 26(7):3397–3409, 2017. 2

[8] Yeying Jin, Wending Yan, Wenhan Yang, and Robby T Tan. Structure representation network and uncertainty feedback learning for dense non-uniform fog removal. In *Computer Vision–ACCV 2022: 16th Asian Conference on Computer Vision, Macao, China, December 4–8, 2022, Proceedings, Part III*, pages 155–172. Springer, 2023. 2

[9] Xudong Kang, Zhengyao Fei, Puhong Duan, and Shutao Li. Fog model-based hyperspectral image defogging. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–12, 2021. 2

[10] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 491–507. Springer, 2020. 2, 4, 5

[11] B. Lee, K. Lee, J. Oh, and I. Kweon. Cnn-based simultaneous dehazing and depth estimation. *ICRA*, pages 9722–9728, 2020. 2

[12] Attila Lengyel, Sourav Garg, Michael Milford, and Jan C van Gemert. Zero-shot day-night domain adaptation with a physics prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4399–4409, 2021. 1, 2, 3

[13] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng. Aod-net: All-in-one dehazing network. *ICCV*, pages 4770–4778, 2017. 2

[14] Ruoteng Li, Robby T Tan, Loong-Fah Cheong, Angelica I Aviles-Rivero, Qingnan Fan, and Carola-Bibiane Schonlieb. Rainflow: Optical flow under rain streaks and rain veiling effect. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7304–7313, 2019. 1

[15] Yu Li, Shaodi You, Michael S Brown, and Robby T Tan. Haze visibility enhancement: A survey and quantitative benchmarking. *Computer Vision and Image Understanding*, 165:1–16, 2017. 1

[16] Wei Liang, Jing Long, Kuan-Ching Li, Jianbo Xu, Nanjun Ma, and Xia Lei. A fast defogging image recognition algorithm based on bilateral hybrid filtering. *ACM transactions on multimedia computing, communications, and applications (TOMM)*, 17(2):1–16, 2021. 2

[17] Yuanchu Liang, Saeed Anwar, and Yang Liu. Drt: A lightweight single image deraining recursive transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 589–598, 2022. 1

[18] Huan Liu, Zijun Wu, Liangyan Li, Sadaf Salehkalaibar, Jun Chen, and Keyan Wang. Towards multi-domain single image dehazing via test-time training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5831–5840, 2022. 2

[19] Wei Liu, Fei Zhou, Tao Lu, Jiang Duan, and Guoping Qiu. Image defogging quality assessment: Real-world database and method. *IEEE Transactions on image processing*, 30:176–190, 2020. 1, 2

[20] Y. Liu, G. Zhao, B. Gong, Y. Li, R. Raj, N. Goel, S. Kesav, S. Gottimukkala, Z. Wang, W. Ren, and D. Tao. Improved techniques for learning to dehaze and beyond: A collective study. *arXiv:1807.00202*, 2018. 2

[21] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12009–12019, 2022. 1, 2, 4, 5

[22] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, pages 10012–10022, 2021. 1, 2, 3, 4, 5

[23] M Jehanzeb Mirza, Marc Masana, Horst Possegger, and Horst Bischof. An efficient domain-incremental learning approach to drive in all weather conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3001–3011, 2022. 1

[24] Y. Qu, J. Chen, and Y. Huang. Enhanced pix2pix dehazing network. *CVPR*, pages 8160–8168, 2019. 2

[25] W. Ren, S. Liu, H. Zhang, J. Pan, X. Cao, and M. Yang. Single image dehazing via multi-scale convolutional neural networks. *ECCV*, pages 154–169, 2016. 2

[26] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126:973–992, 2018. 1, 2

[27] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10765–10775, 2021. 1, 2

[28] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. *ECCV*, pages 746–760, 2012. 2

[29] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. 2, 4, 5

[30] Jeya Maria Jose Valanarasu, Rajeev Yasarla, and Vishal M Patel. Transweather: Transformer-based restoration of images degraded by adverse weather conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2353–2363, 2022. 2

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia

Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 4, 5

[32] Wending Yan, Aashish Sharma, and Robby T Tan. Optical flow in dense foggy scenes using semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13259–13268, 2020. 1

[33] W. Yang, Y. Yuan, W. Ren, and et al. Advancing image understanding in poor visibility environments: A collective benchmark study. *IEEE Transactions on Image Processing*, 29:5737–5752, 2020. 2

[34] Yang Yang, Chaoyue Wang, Risheng Liu, Lin Zhang, Xiaojie Guo, and Dacheng Tao. Self-augmented unpaired image dehazing via density and depth decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2037–2046, 2022. 2

[35] Shaodi You, Robby T Tan, Rei Kawakami, Yasuhiro Mukaigawa, and Katsushi Ikeuchi. Adherent raindrop modeling, detection and removal in video. *IEEE transactions on pattern analysis and machine intelligence*, 38(9):1721–1733, 2015. 1

[36] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10819–10829, 2022. 2

[37] Ghais Zaher. Simulating weather conditions on digital images. 2020. 2, 3

[38] J. Zhang and D. Tao. Famed-net: A fast and accurate multi-scale end-to-end dehazing network. *IEEE Transactions on Image Processing*, 29:72–84, 2019. 2

[39] Yinqiang Zheng, Mingfang Zhang, and Feng Lu. Optical flow in the dark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6749–6757, 2020. 1

[40] Chu Zhou, Minggui Teng, Yufei Han, Chao Xu, and Boxin Shi. Learning to dehaze with polarization. *Advances in Neural Information Processing Systems*, 34:11487–11500, 2021. 1, 2

[41] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 3