Single-Domain Generalized Object Detection in Urban Scene via Cyclic-Disentangled Self-Distillation: Supplementary Material

Aming Wu, Cheng Deng
School of Electronic Engineering, Xidian University, Xi'an, China
amwu@xidian.edu.cn, chdeng@mail.xidian.edu.cn

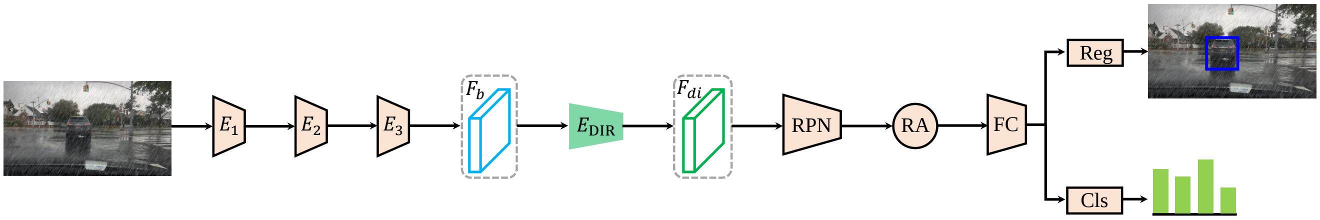

Figure 1. Illustration of the proposed method for the testing stage. During inference, we directly utilize the output of the extractor E_{DIR} to make predictions. The cyclic-disentangled operation and self-distillation module are not used in the testing stage.

In supplementary material, we provide additional quantitative results to examine the impact of hyper-parameters. Meanwhile, more visualization and qualitative results are given to further analyze the proposed method.

1. Inference Stage

Fig. 1 shows the details of the model for testing. During inference, given an input image, we directly utilize the disentangled output of the extractor E_{DIR} to make predictions. The cyclic-disentangled operation and self-distillation module are not required to use in the inference process. And E_{DSR} is also not used. This further shows that promoting E_{DIR} to own the ability of disentangling domain-invariant features is important for Single-DGOD.

2. Analysis of Hyper-Parameters

In Eq. (3) and (4) of the submission, we separately define a global-level and instance-level contrastive loss to enhance the disentangled ability. Here, based on the night-sunny scene, we analyze the impact of the hyper-parameter τ. When τ is separately set to 1.5, 1.0, and 0.5, the corresponding performance is 35.8%, 36.6%, and 36.2%. We can see that different settings of τ affect the performance. When τ is set to 1.0, the performance is the best.

Next, we analyze the impact of the hyper-parameter λ in Eq. (7) of the submission. Based on the night-sunny scene, when λ is separately set to 0.1, 0.01, and 0.001, the corresponding performance is 35.4%, 36.6%, and 35.9%.

3. Visualization Analysis of Self-Distillation

Taking F_{di} as the teacher representation, self-distillation is employed to promote the feature maps generated by the middle layers of the backbone network to contain more domain-invariant information, which is beneficial for further enhancing the generalization ability.

In Fig. 2, we show two visualization examples. The second column indicates the extracted F_b based on self-distillation. The third column is the F_b without using self-distillation. We can see that compared with F_b of the third column, the F_b of the second column contains much more object-related information and less background that mainly reflects domain-specific information. Moreover, our method detects the objects in the rainy images accurately. This shows that taking the disentangled domain-invariant features as the teacher, employing self-distillation is indeed helpful for facilitating the features generated by the middle layers of the backbone network to involve more domain-invariant information containing intrinsical object characteristics, which further improves the disentangled ability and detection performance.

4. Further Analysis of Disentangled Ability

This section further analyzes the disentangled ability of the proposed cyclic-disentangled self-distillation. In Fig. 3, we show three examples with different weather conditions. Compared with F_{di}, F_{i2i} contains much stronger object-related information. Meanwhile, compared with F_{di} and

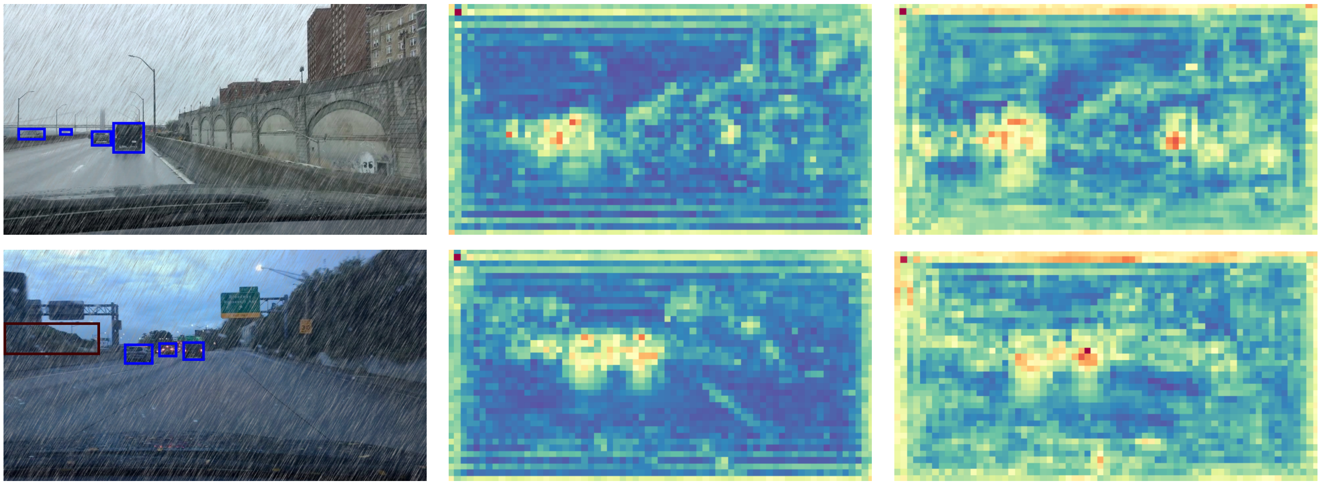

Figure 2. Visualization analysis of self-distillation. Here, the first column is the detection results of our method. The second column denotes the output F_b of E_3 based on self-distillation. And the third column indicates the output F_b of E_3 without using self-distillation. For each feature map, the channels corresponding to the maximum value are selected for visualization.

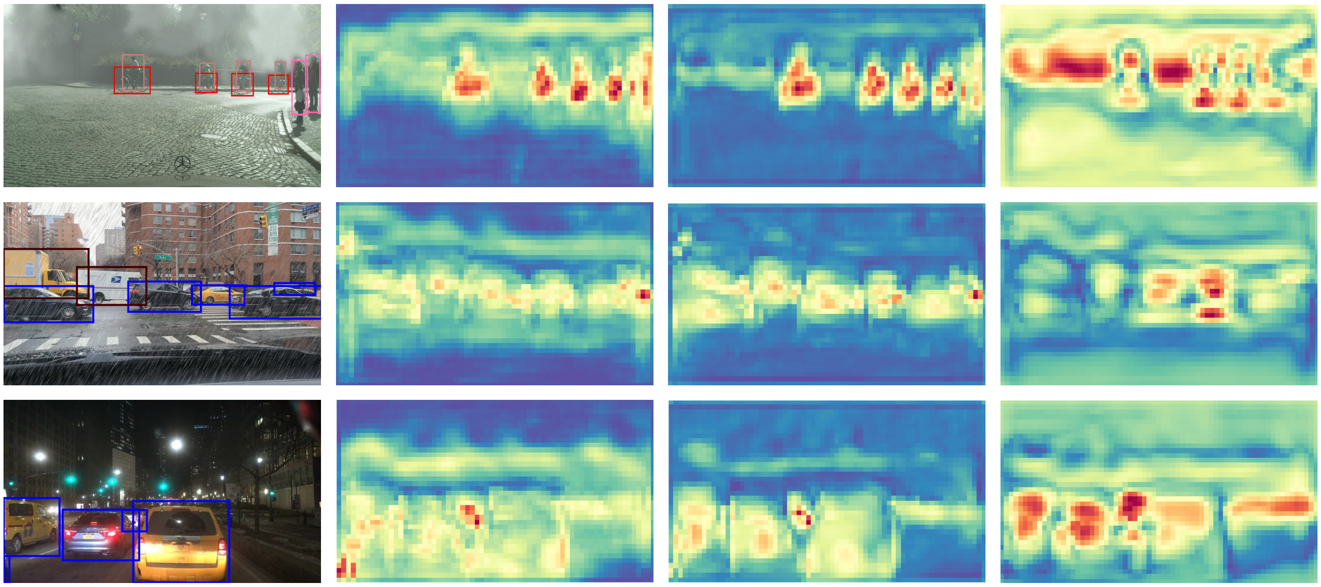

Figure 3. Visualization analysis of the disentangled ability of the proposed method. The first column is the detection results of our method. The second, third, and fourth column separately denote the output F_{di}, F_{i2i}, and F_{i2s} (as shown in Eq. (2) of the submission). For each feature map, the channels corresponding to the maximum value are selected for visualization.

F_{i2i}, F_{i2s} involves more background that reflects domain-specific information. This further shows that our method is indeed helpful for promoting the detector to own the disentangled ability, which improves the generalization ability. Finally, in Fig. 4, 5, and 6, we show more visualization results. And Fig. 7, 8, 9, and 10 give more detection results based on five different weather conditions. We can see that our method disentangles domain-invariant features and detects objects effectively. This further indicates that the proposed cyclic-disentangled self-distillation is indeed conductive to enhancing the generalization ability.

Limitations. Our method is evaluated on a built Diverse-Weather Dataset consisting of five different weather conditions, i.e., daytime-sunny, night-sunny, dusk-rainy, night-rainy, and daytime-foggy. However, this dataset does not cover all weather conditions for urban-scene object detection. In the future, we will collect more data (e.g., the snowy and dust weather) to further complete this dataset.

Potential Negative Societal Impact. In this paper, our method only utilizes the daytime-sunny data to train the object detector. However, to improve the detection performance, our method still requires humans to accurately annotate the daytime-sunny data. Alleviating human labeling efforts will be future work.

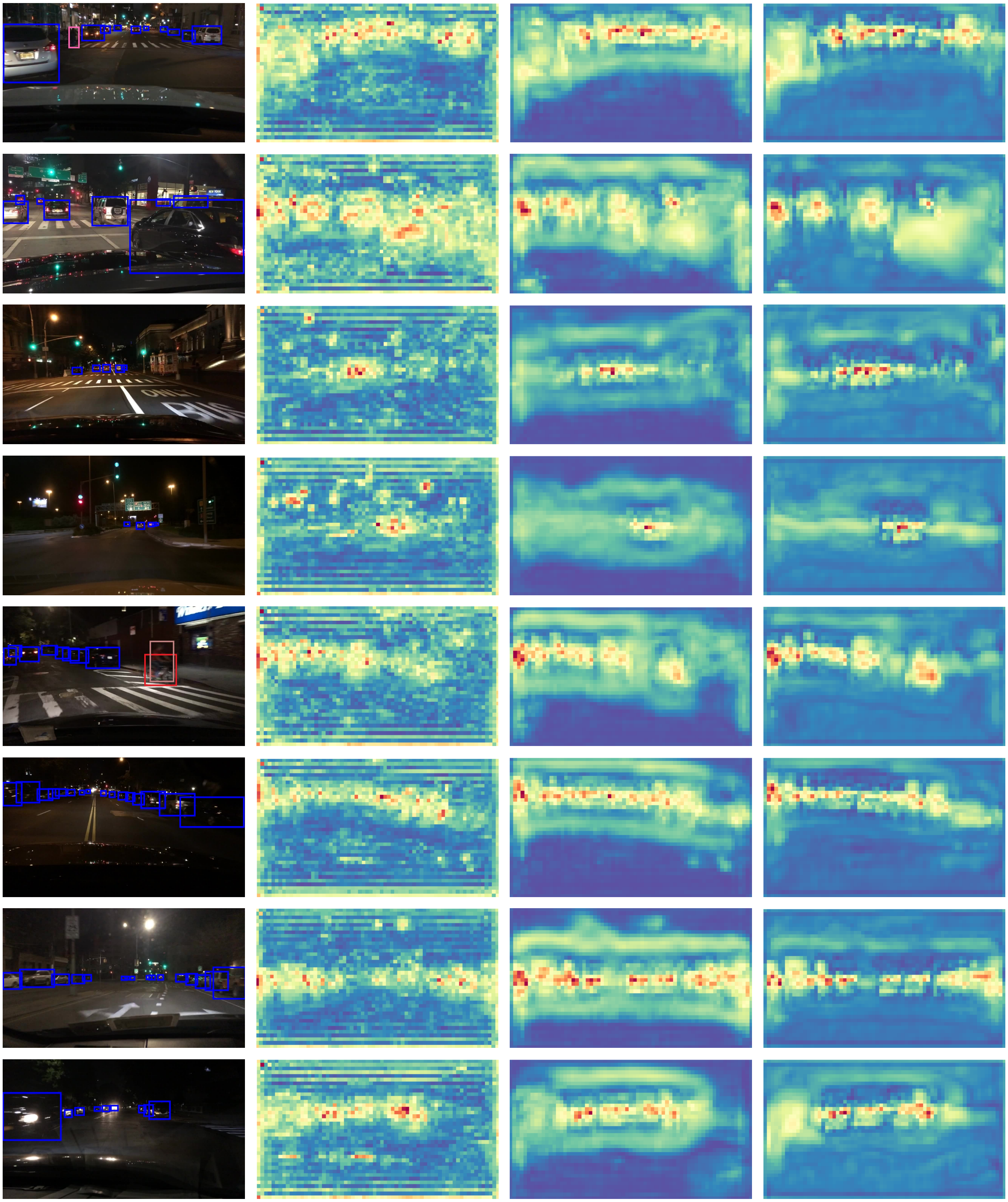

Figure 4. Visualization analysis of our method based on the night-sunny scene. The first column is the detection results of our method. The second, third, and fourth columns separately indicate the features F_b, F_{di}, and F_{i2i}. For each feature map, the channels corresponding to the maximum value are selected for visualization.

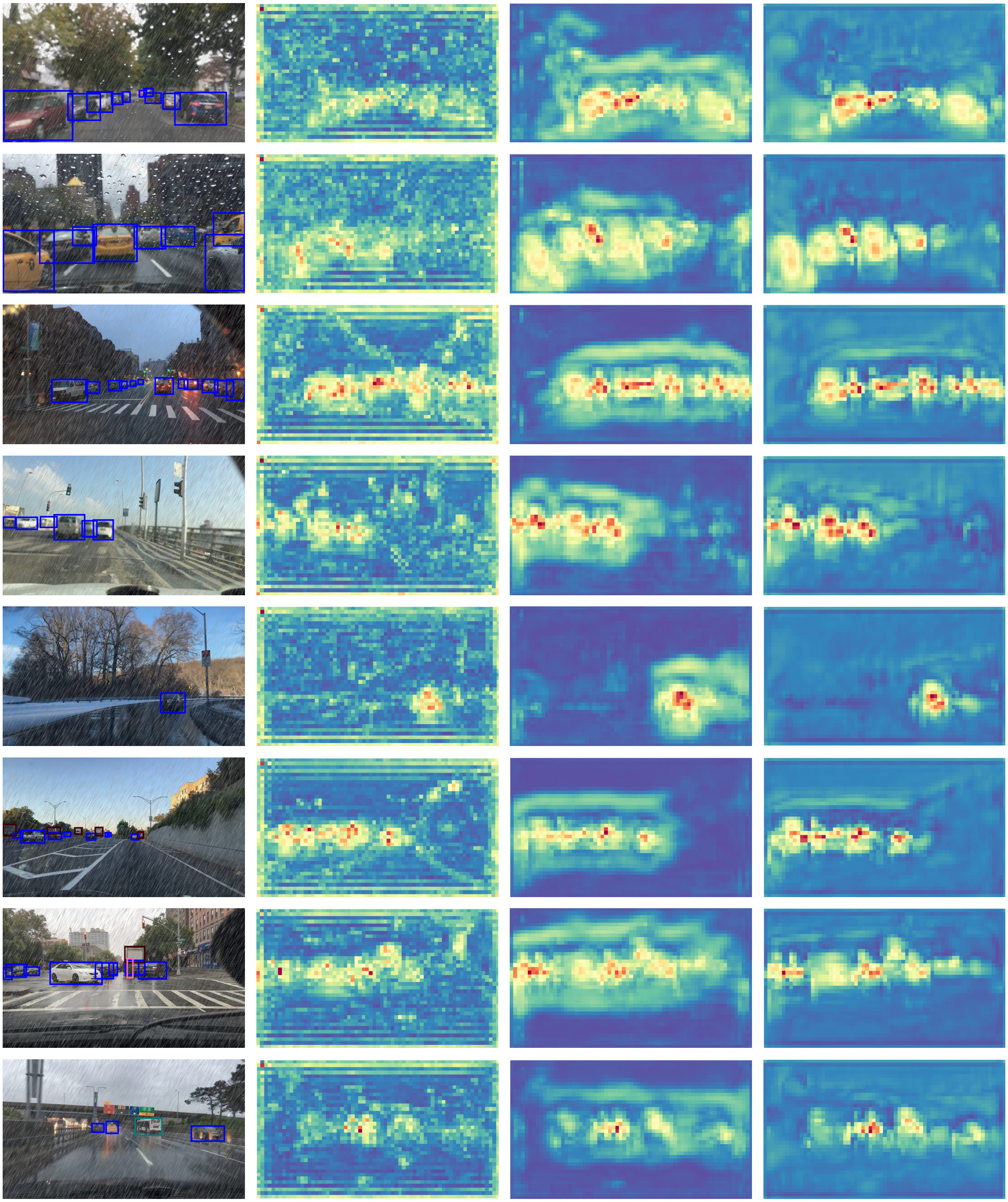

Figure 5. Visualization analysis of our method based on the dusk-rainy scene. The first column is the detection results of our method. The second, third, and fourth columns separately indicate the features F_b, F_{di}, and F_{i2i}. For each feature map, the channels corresponding to the maximum value are selected for visualization.

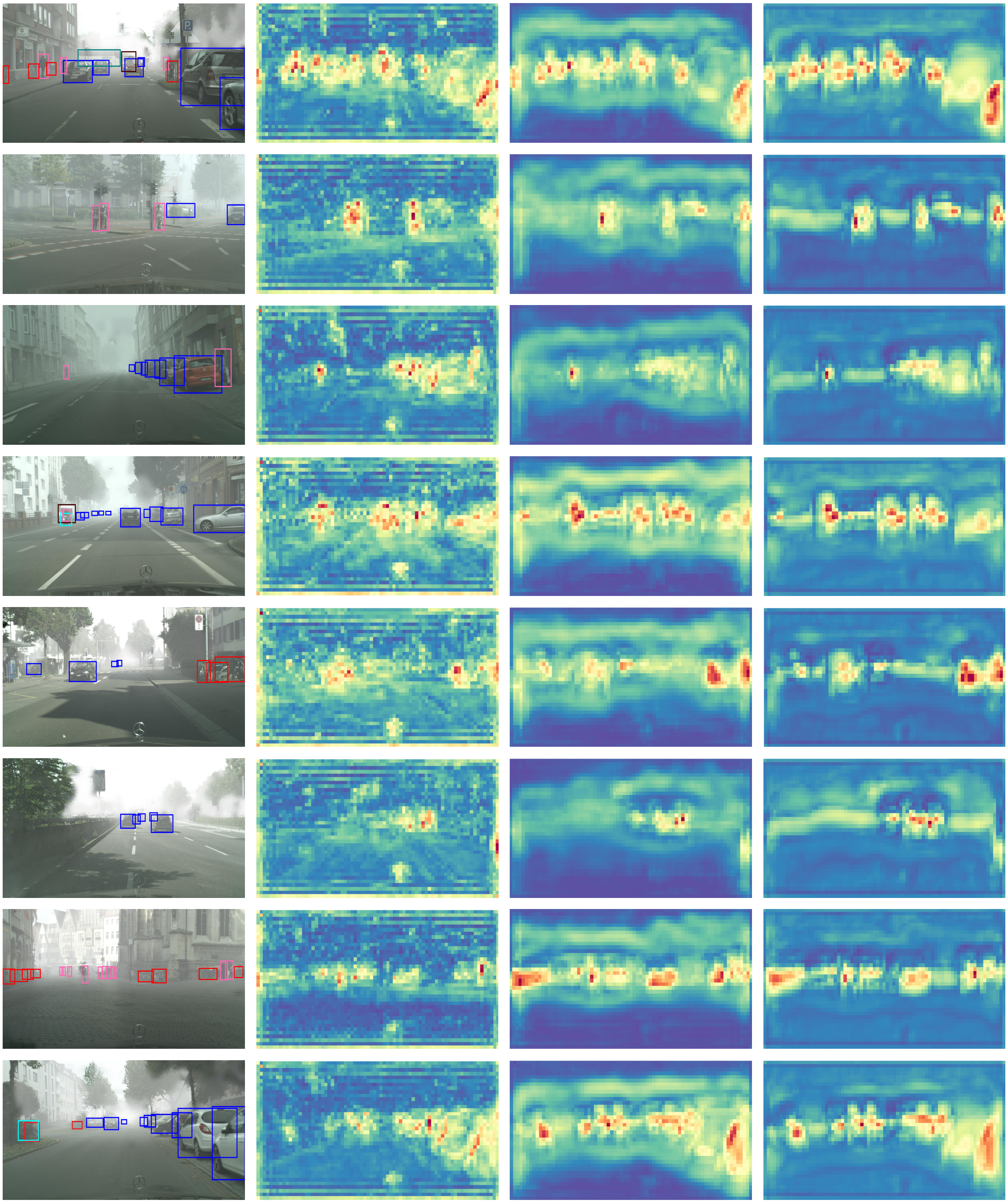

Figure 6. Visualization analysis of our method based on the daytime-foggy scene. The first column is the detection results of our method. The second, third, and fourth columns separately indicate the features F_b, F_{di}, and F_{i2i}. For each feature map, the channels corresponding to the maximum value are selected for visualization.

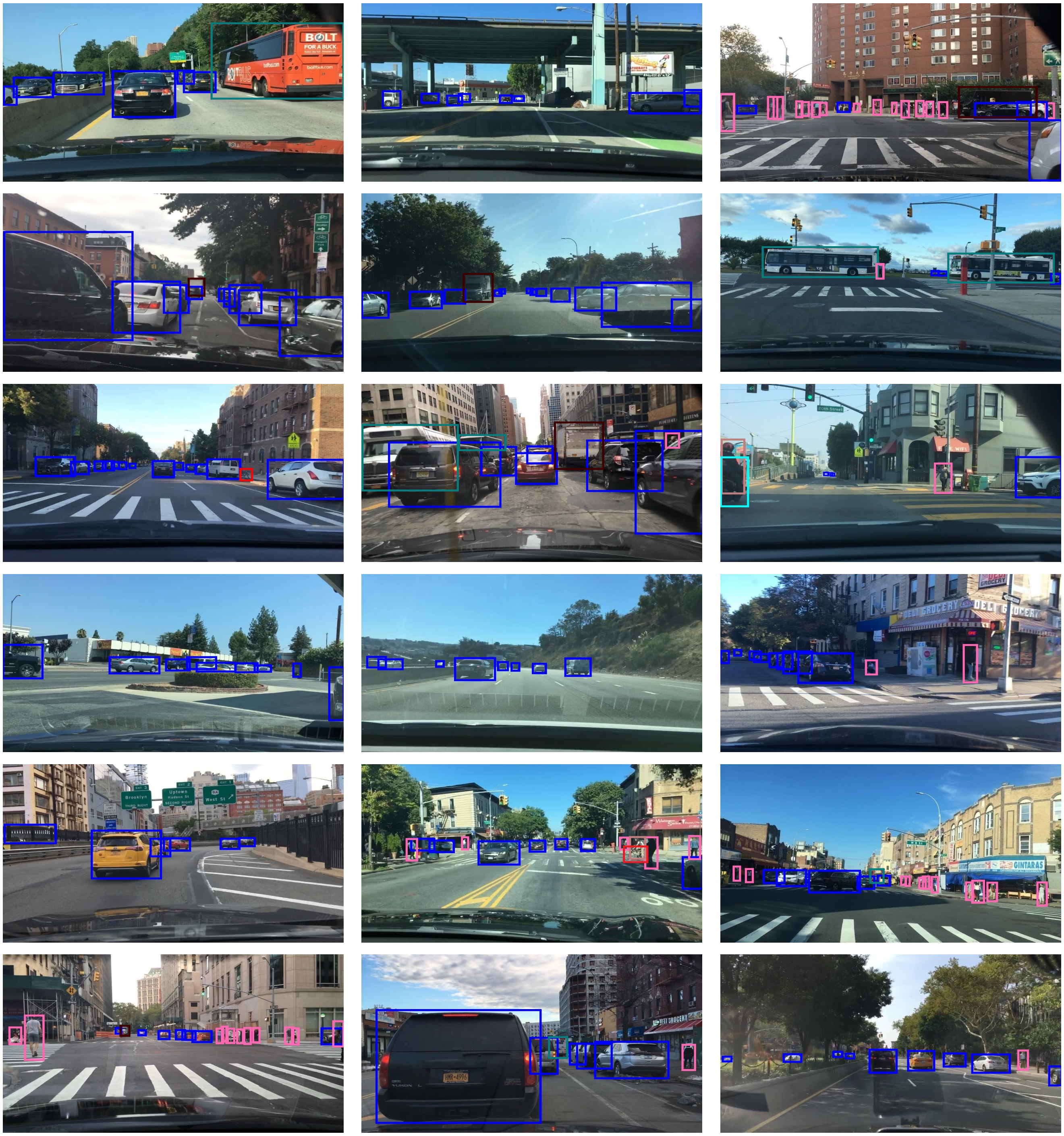

Figure 7. Detection results on the daytime-sunny scene. We can see that our method accurately detects objects in the daytime-sunny images, which shows that our method is still effective for the current domain.

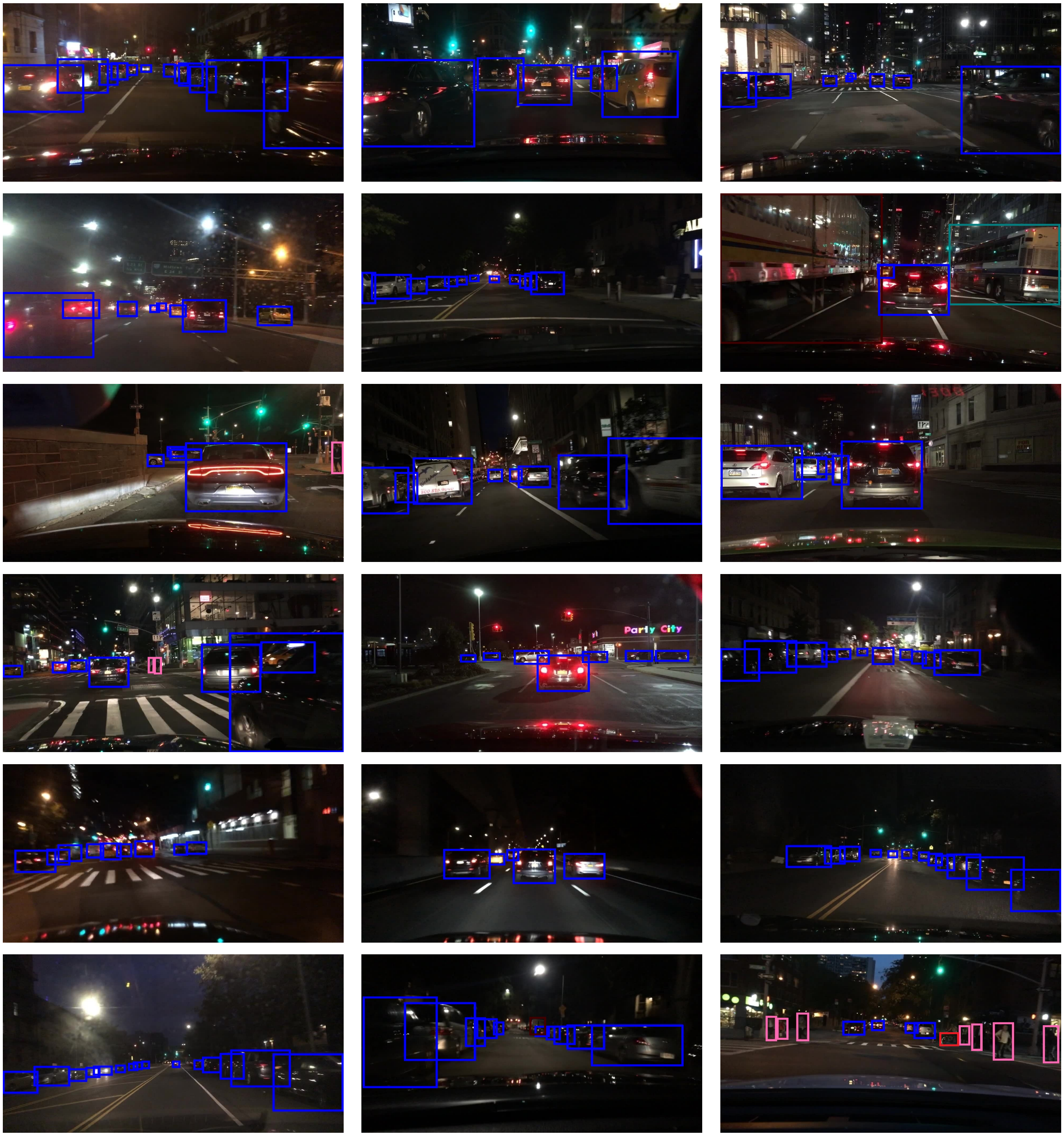

Figure 8. Detection results on the night-sunny scene. We can see that our method accurately detects objects in the night images, which shows that our method is indeed helpful for enhancing the generalization ability.

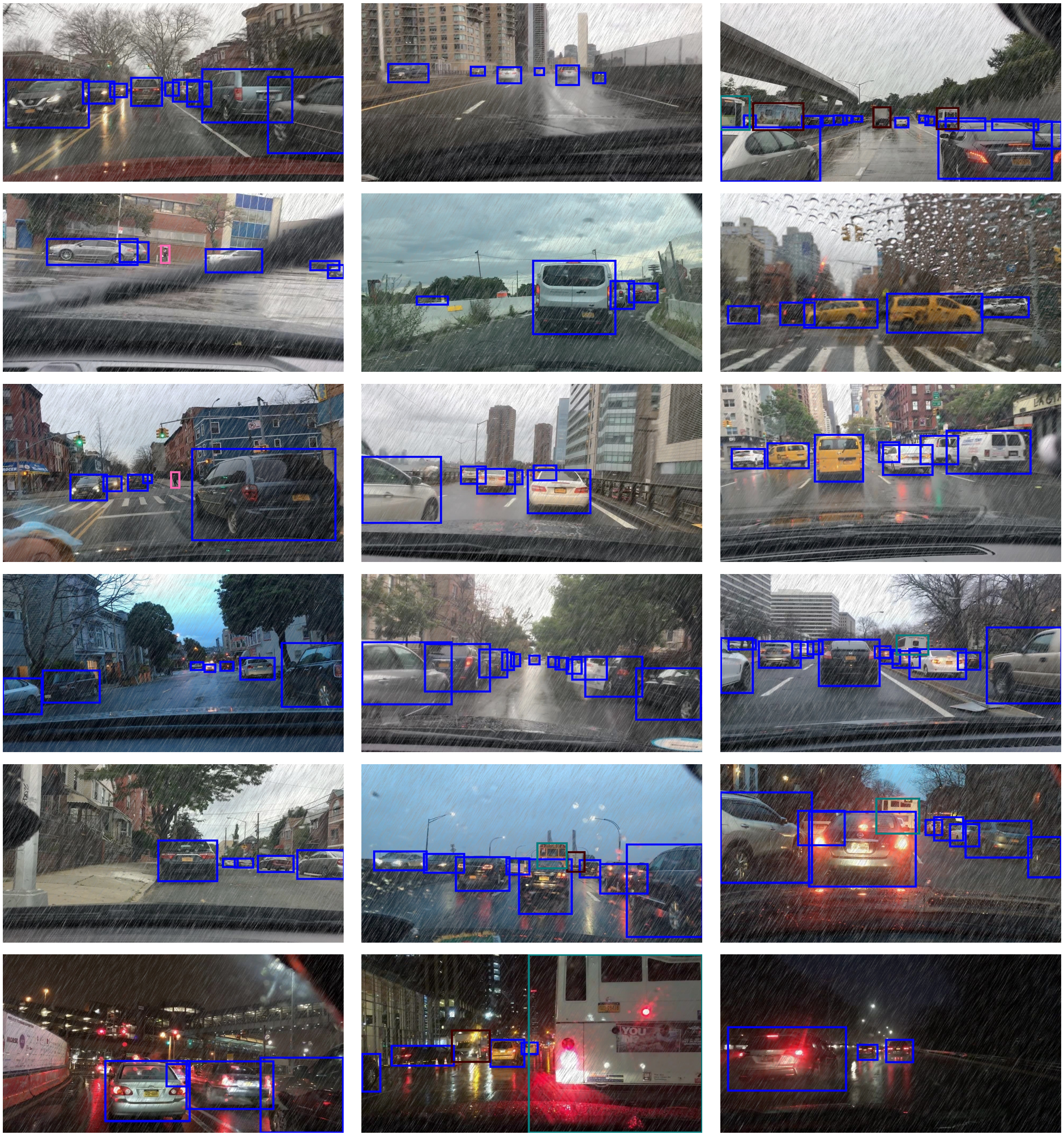

Figure 9. Detection results on the dusk-rainy and night-rainy scenes. The last two rows show the night-rainy results. We can see that our method accurately detects objects in the rainy images, which shows the effectiveness of our method.

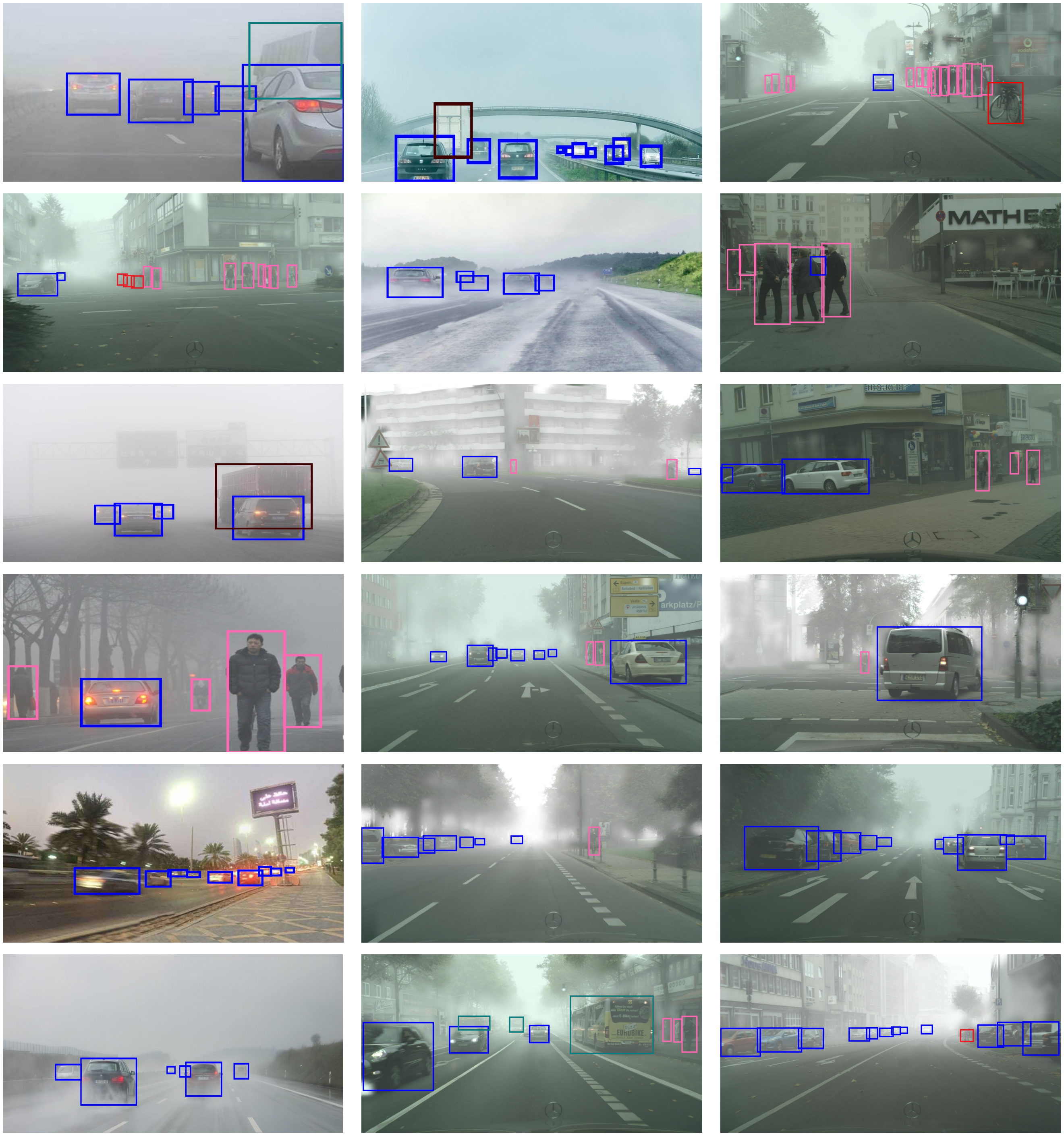

Figure 10. Detection results on the daytime-foggy scene. We can see that our method accurately detects objects in the foggy images, which shows that our method is indeed helpful for enhancing the generalization ability.