MT-DETR: Robust End-to-end Multimodal Detection with Confidence Fusion Supplementary Material

Shih-Yun Chu
National Taiwan University
r09922115@csie.ntu.edu.tw

Ming-Sui Lee
National Taiwan University
mslee@csie.ntu.edu.tw

1. Fog Synthesis Algorithm on Camera

1.1. Algorithm Details

As mentioned in Section 3.4.1, [2, 5] use the following composition function to synthesize foggy images. Given a clear image I captured by the camera, the corresponding depth map D, and the atmosphere light A, the foggy image $I^{'}$ can be generated by:

$$\begin{aligned} T &= e^{-\beta \times D} \,, \\ I^{'} &= T * I + (1-T) * A \,, \end{aligned} \tag{1}$$

where β is a weight representing the density of fog, which is set to 1.0, $*$ and + denote pixel-wise multiplication and addition, respectively.

Two modifications are introduced based on previous methods [2, 5]. (1) The sampling interval is adjusted over time to consider the difference between day and night. (2) Real foggy images have apparent glare effects, so the atmosphere light A varies by referring to the local brightness of the camera image I. In other words, the atmosphere light A is rather than a single value for the whole frame. It is brighter in the lighting parts (e.g., the regions of the street lights), and the rest remains unchanged. With the above designs, the fog synthesis algorithm is improved to generate more realistic foggy images. The details are shown in Algorithm 1.

Algorithm 1 Foggy Image Synthesis Algorithm

```
Input: clear image $I$, $time \in$ {day,night}
Output: foggy image $I_{\text{foggy}}$
 1: $depth \leftarrow$ GPT($I$)   ▷ predict depth by the pretrained model
 2: $T \leftarrow e^{-1.3 \times depth}$   ▷ obtain the transmission map
 3: if $time$ is day then
 4:     $light \sim U(0.4, 0.75)$
 5: else if $time$ is night then
 6:     $light \sim U(0.3, 0.65)$
 7:     $bright \leftarrow$ (grayscale($I$) $- 205)/50$
            ▷ find the bright part of the image
 8:     $glare \leftarrow bright$
            ▷ initialize the glare effect by bright part
 9:     $glare \leftarrow$ max(blur($glare$), $glare$)
            ▷ stronger effect with more iterations
10:     $light \leftarrow light + (0.95 - light) \times glare$
            ▷ adjust light with the glare effect
11: end if
12: $I_{\text{foggy}} \leftarrow I \times T + light \times (1 - T)$
            ▷ generate the foggy image
```

1.2. Visualization of the Fog Synthesis Process

Two examplar images along with the intermediate results are demonstrated in Figure 1. The numbers in the captions represent their corresponding line number in the algorithm.

Two examplar images along with the intermediate results are demonstrated in Figure 1. The numbers in the captions represent their line number in the algorithm. Figure 1(a) shows clear camera images, which are the input to the synthesis algorithm. Figure 1(b) are the depth maps predicted by the pretrained depth estimation model. In Figure 1(c), the lighting parts of the image are labeled in white color. For those marked white regions, we iteratively spread the lighting effect until the diffusion is as desired. The resultant glare effect is demonstrated in Figure 1(d) and added to the constant atmospheric light as shown in Figure 1(e). Finally, the clear image and atmospheric light are combined using the transmission map as transparency. Figure 1(f) are the synthesized foggy images.

2. Complete Loss Function

Due to insufficient space in the main paper, the complete loss functions and weights are supplemented here. Let P_m be the prediction of MT-DETR from each modality and $\hat{P}$ denote the ground truth. For $m \in \{\text{fusion, camera, depth}\}$, the complete loss function is as follows:

$$\mathcal{L}_m = 2\mathcal{L}_{\text{focal}}(P_m, \hat{P}) + 5\mathcal{L}_{l_1}(P_m, \hat{P}) + 2\mathcal{L}_{\text{GIoU}}(P_m, \hat{P}) \ , \tag{2}$$

$$\mathcal{L}_{\text{total}} = \lambda_{\text{fusion}}\mathcal{L}_{\text{fusion}} + \lambda_{\text{camera}}\mathcal{L}_{\text{camera}} + \lambda_{\text{depth}}\mathcal{L}_{\text{depth}} \ , \tag{3}$$

where $\mathcal{L}_{\text{focal}}(\cdot,\cdot)$ indicates focal loss [3], $\mathcal{L}_{l_1}(\cdot,\cdot)$ indicates l_1 loss, $\mathcal{L}_{\text{GIoU}}(\cdot,\cdot)$ indicates GIoU loss [4], and their corresponding weights are follow Deformable DETR [6]. (λ_{fusion}, λ_{camera}, λ_{depth}) are set to (1, 1, 0.5) due to our experiments.

3. Idea of RFM's and REM's design

As mentioned in Section 3.2, Residual Fusion Module (RFM) and Residual Enhanced Module (REM) have similar structures. The difference lies in the importance of branches and the depth of convolution blocks. RFM pays more attention to the more substantial branch, so there is a residual connection for the lidar features. As REM focuses more on the branch with less information, it has a residual connection for the unimodal features. Since REM also attempts to extract more valuable features from the more substantial branch, it contains more convolution blocks than RFM.

4. Ablation Study of Synthetic Fog Density

The camera-lidar foggy image pairs are generated using the approach mentioned in Section 3.4. Different densities of the synthesized fog effects are tested to determine which setting makes the MT-DETR more adaptable to adverse weather. Those synthesized data and the original clear data form a new training dataset. The experimental results are shown in Table 1, with AP_{75} as the metric. It can be observed that the synthetic data with density fog 0.010 provides the most effective help in the detection task under various weather conditions. It also proves that the synthetic algorithm with glare effect is better than without glare effect.

Table 1: **Ablation study of fog density for data synthesis.** The best result is highlighted in **bold**, and the one with * uses the synthesis algorithm without glare effect.

Training Data					Testing Data							
synthetic fog density					clear		light fog		dense fog		snow	
0.000	0.005	0.010	0.020	0.030	day	night	day	night	day	night	day	night
✓					64.7	62.2	67.0	63.7	71.3	**70.7**	65.9	64.8
✓	✓				65.3	62.1	67.3	64.7	70.5	70.3	66.7	65.4
✓			✓		65.7	62.6	65.3	65.5	70.5	68.7	66.8	**65.7**
✓				✓	65.0	62.4	67.4	65.0	71.3	70.0	66.4	65.0
✓		✓*			65.7	**63.2**	67.2	65.3	70.3	68.6	66.9	65.6
✓		✓			**66.2**	63.1	**68.0**	**65.8**	**71.7**	70.1	**67.2**	65.6

MT-DETR with synthetic training data may perform worse than the baseline in dense fog conditions. This is because the synthetic fog is too light to simulate the "dense fog" level, so the model overfits the lighter weather. Although each synthetic data density performs differently, it is worth noting that using synthetic data of any density can achieve better model performance than not using synthetic data.

5. Qualitative Visualization

The predictions of the MT-DETR on the STF [1] dataset are exhibited in Figure 2 to 9 according to weather and time. Two examplar images are selected for each condition and compared to the ground truths. The results demonstrate that the MT-DETR successfully detects even small objects under difficult conditions.

References

[1] Mario Bijelic, Tobias Gruber, Fahim Mannan, Florian Kraus, Werner Ritter, Klaus Dietmayer, and Felix Heide. Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11682–11692, 2020.

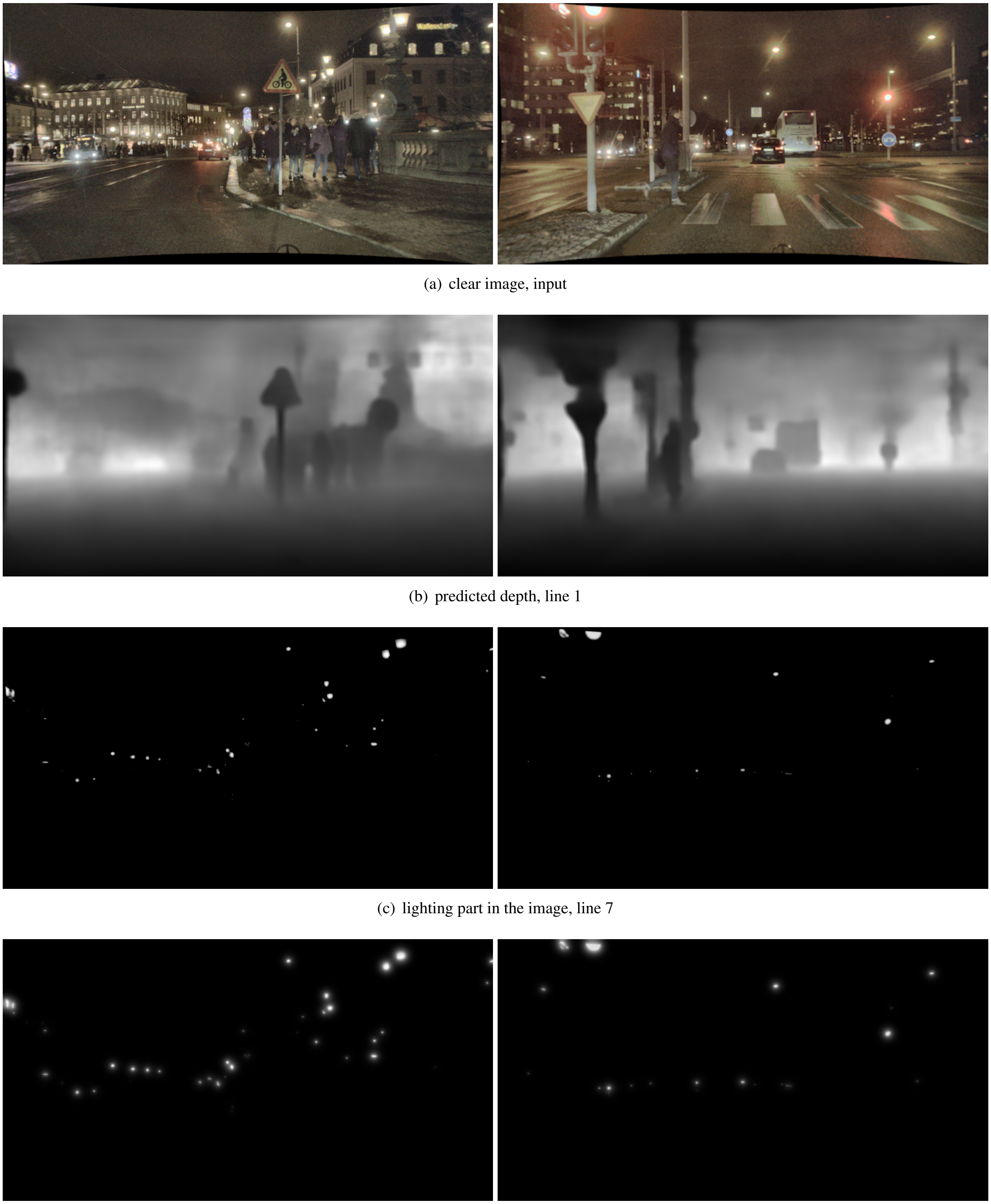
[2] Ruoteng Li, Loong-Fah Cheong, and Robby T Tan. Heavy rain image restoration: Integrating physics model and conditional adversarial learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1633–1642, 2019.

[3] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[4] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019.

[5] Christos Sakaridis, Dengxin Dai, Simon Hecker, and Luc Van Gool. Model adaptation with synthetic and real data for semantic dense foggy scene understanding. In *Proceedings of the european conference on computer vision (ECCV)*, pages 687–704, 2018.

[6] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.



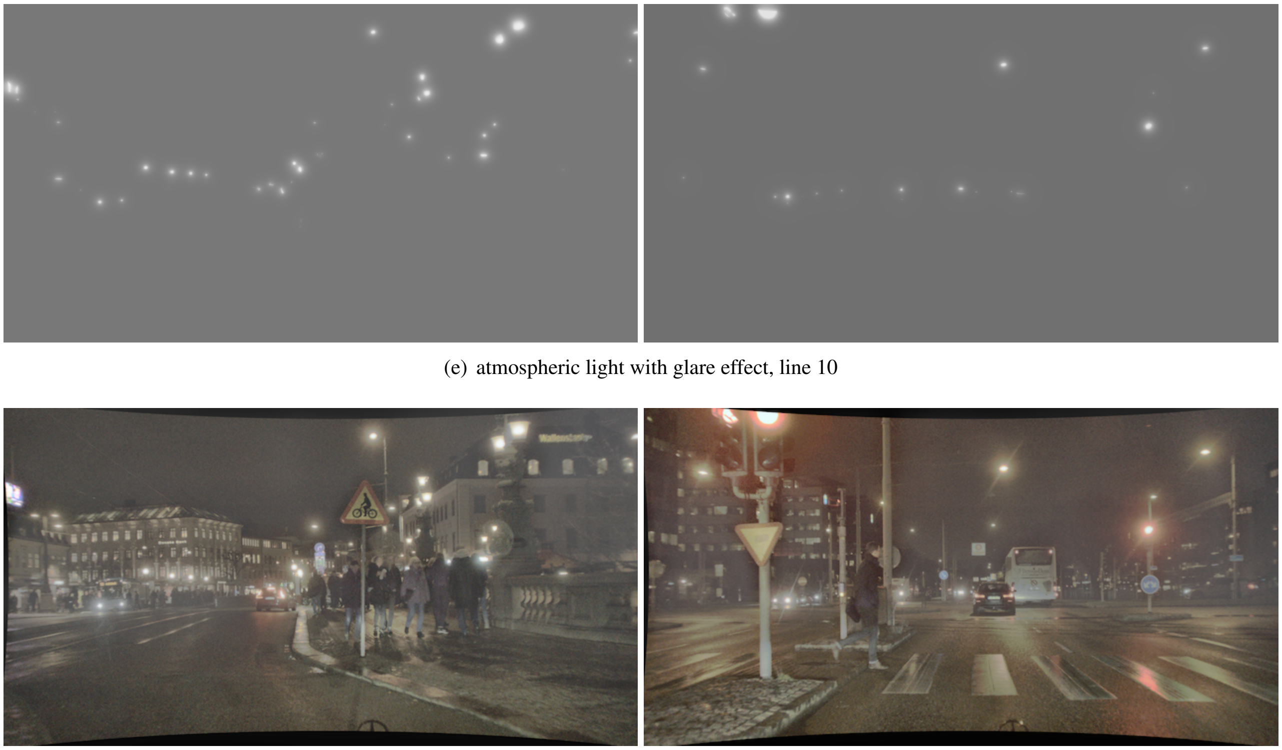
(a) clear image, input

(b) predicted depth, line 1

(c) lighting part in the image, line 7

(d) lighting part with glare effect, line 9

Figure 1: **Visualized results at each stage of fog synthesis algorithm process.**



(e) atmospheric light with glare effect, line 10

(f) fog synthesis result, line 12

Figure 1: **Visualized results at each stage of fog synthesis algorithm process. (cont.)**

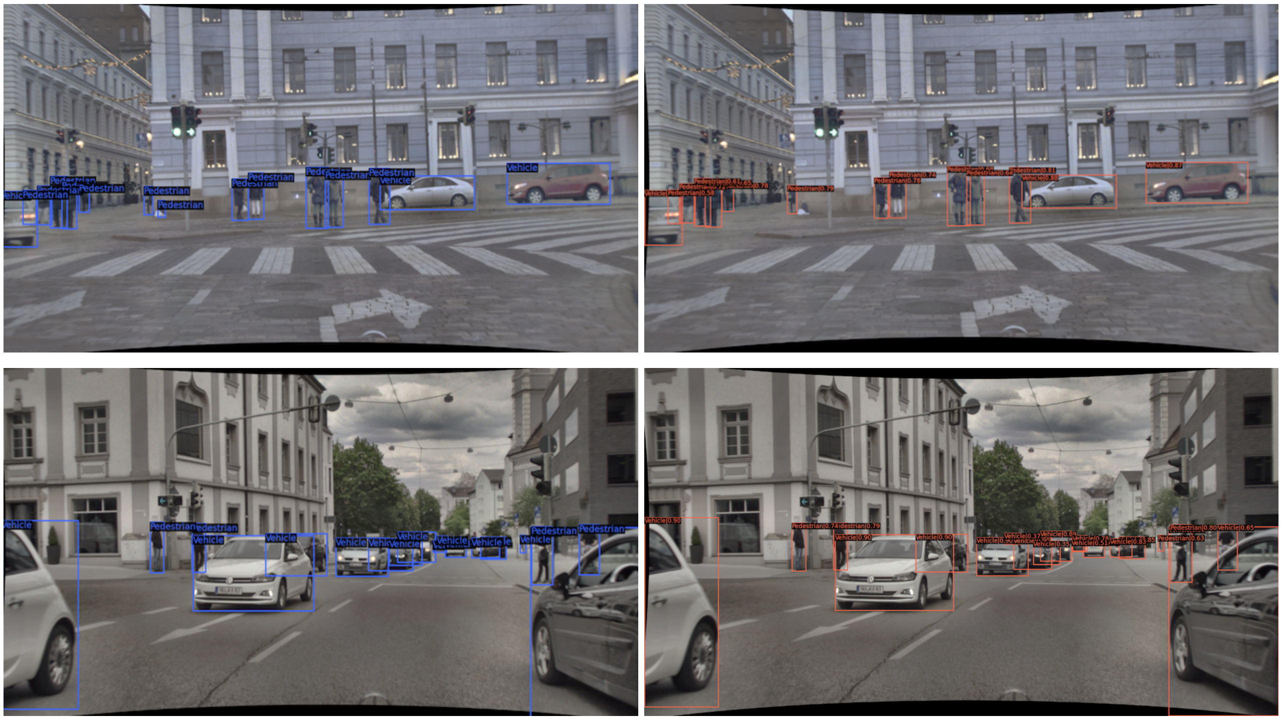


Figure 2: **Visualized results on STF clear day test split.** The left images are the ground truth, and the right images are the predictions of MT-DETR.

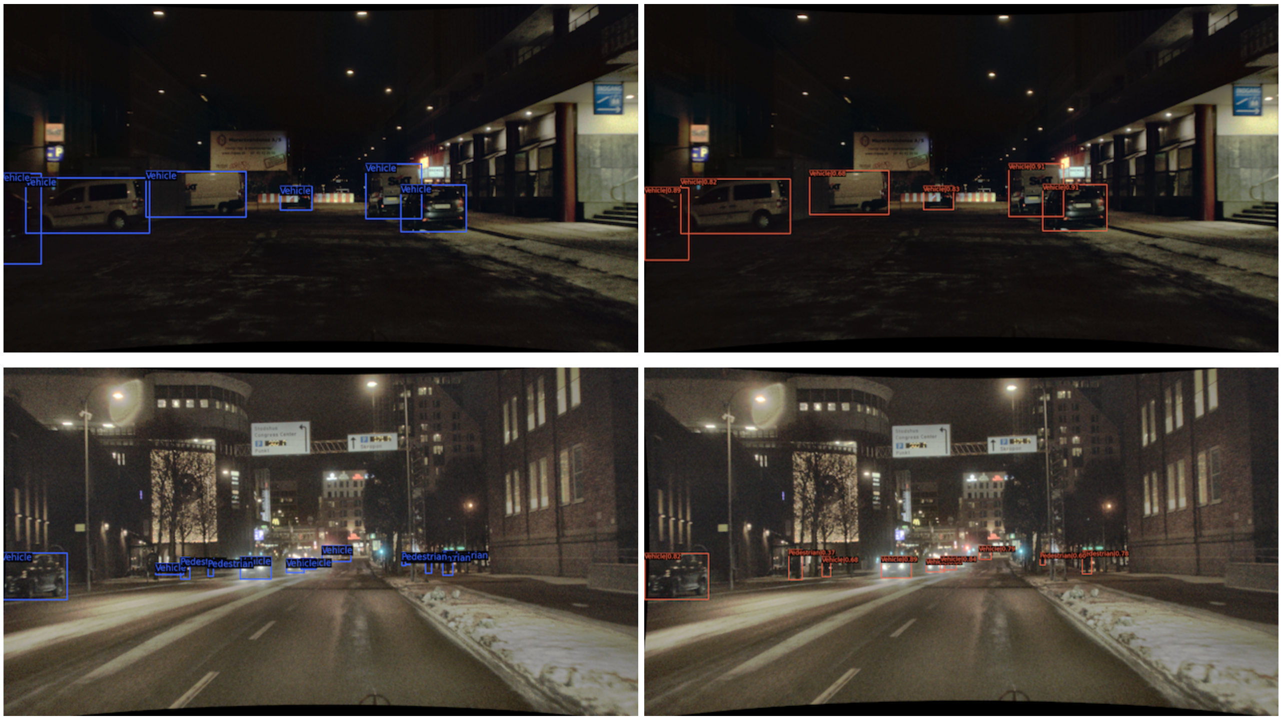


Figure 3: **Visualized results on STF clear night test split.** The left images are the ground truth, and the right images are the predictions of MT-DETR.

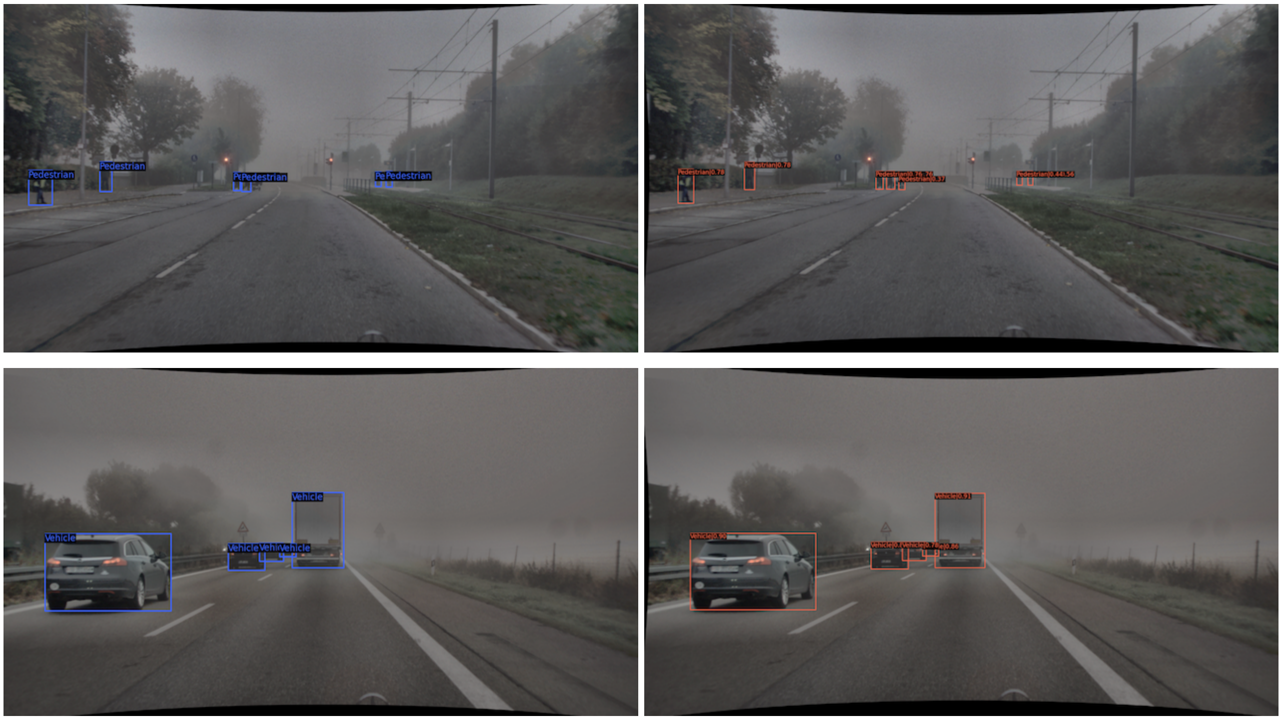


Figure 4: **Visualized results on STF light fog day test split.** The left images are the ground truth, and the right images are the predictions of MT-DETR.

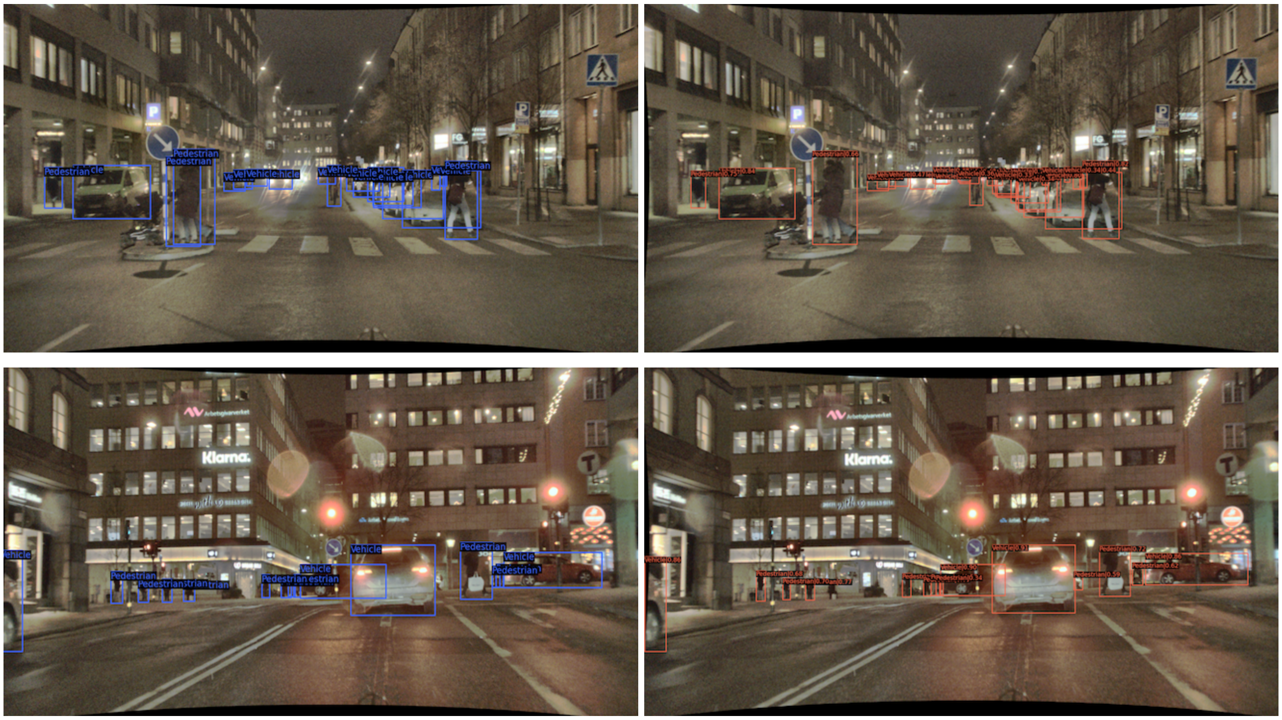


Figure 5: **Visualized results on STF light fog night test split.** The left images are the ground truth, and the right images are the predictions of MT-DETR.

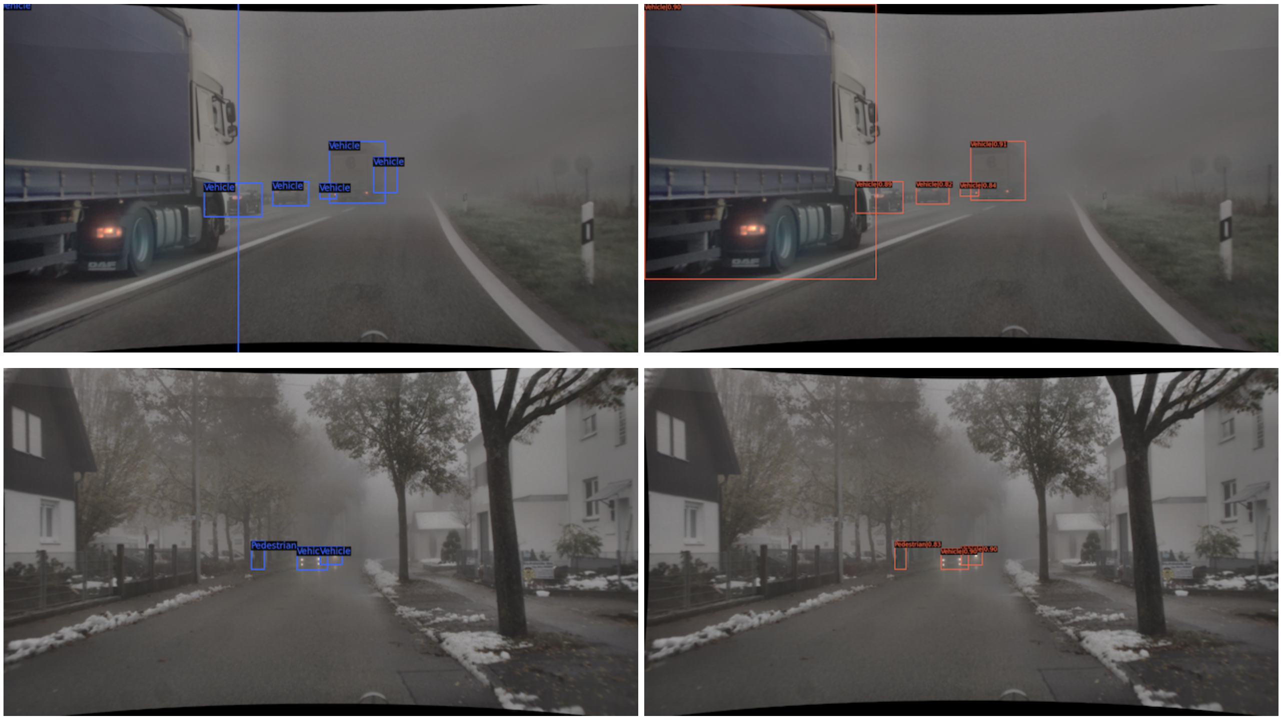


Figure 6: **Visualized results on STF dense fog day test split.** The left images are the ground truth, and the right images are the predictions of MT-DETR.

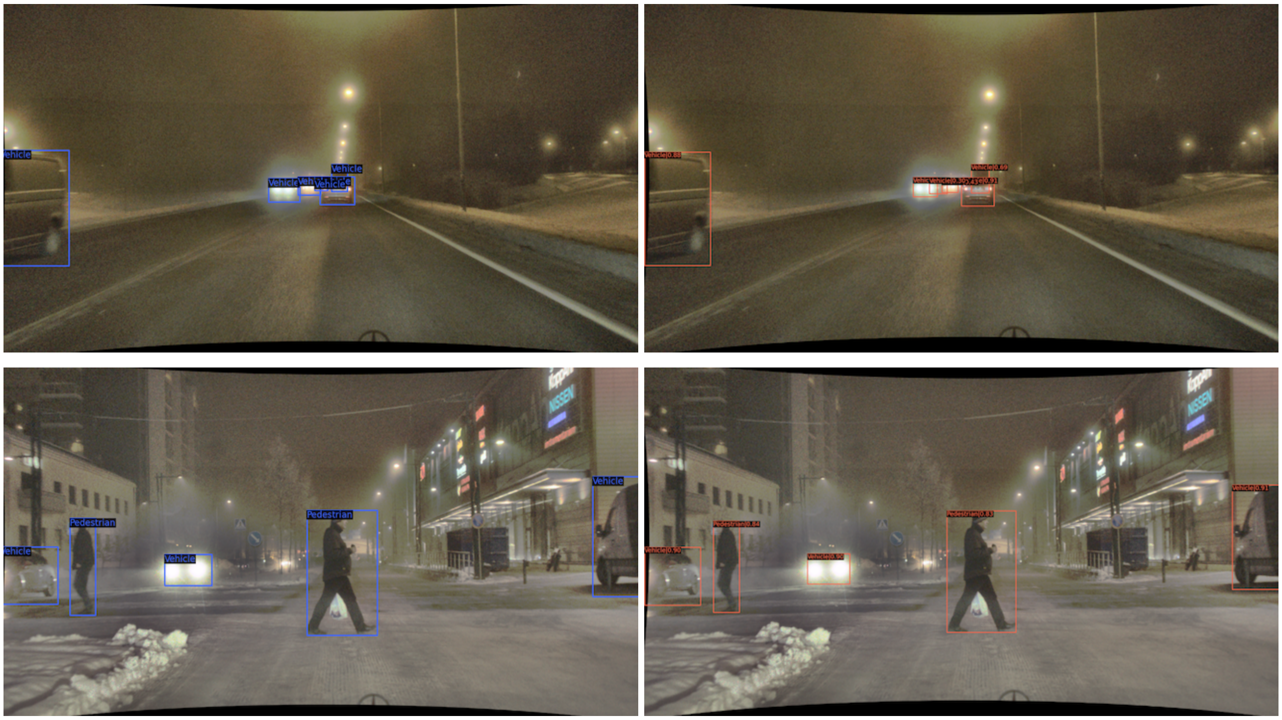


Figure 7: **Visualized results on STF dense fog night test split.** The left images are the ground truth, and the right images are the predictions of MT-DETR.

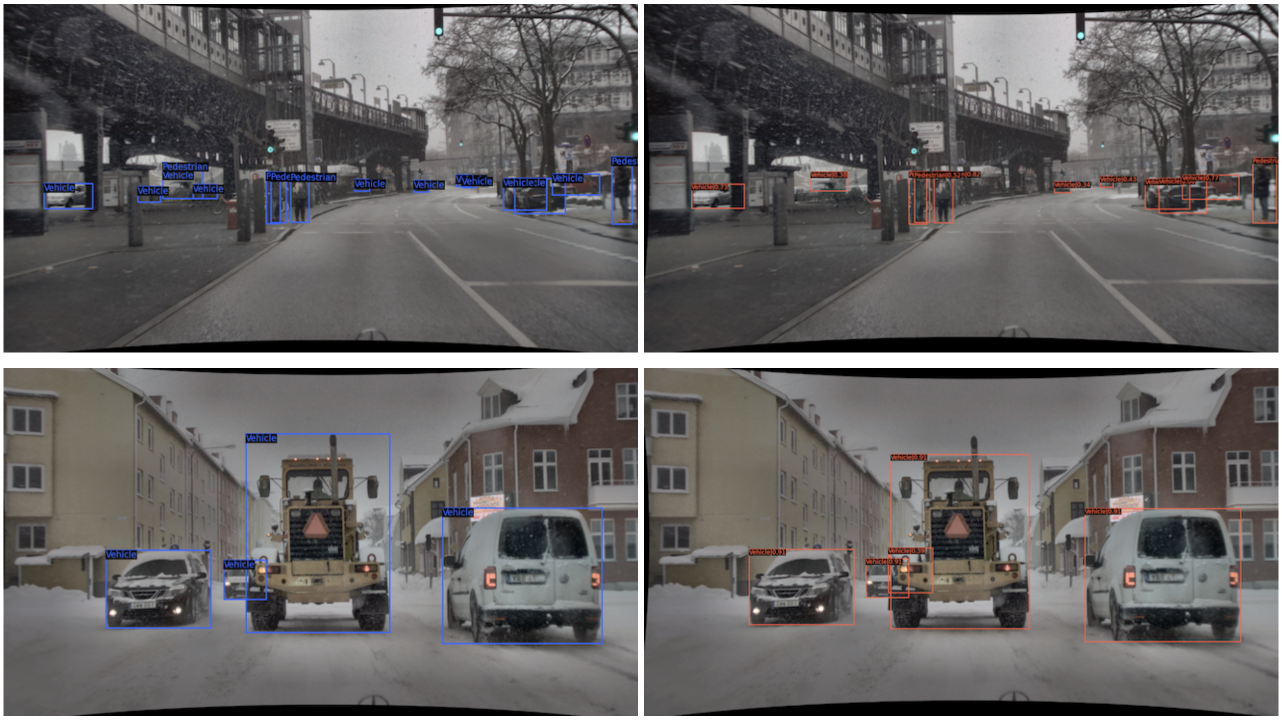


Figure 8: **Visualized results on STF snow day test split.** The left images are the ground truth, and the right images are the predictions of MT-DETR.

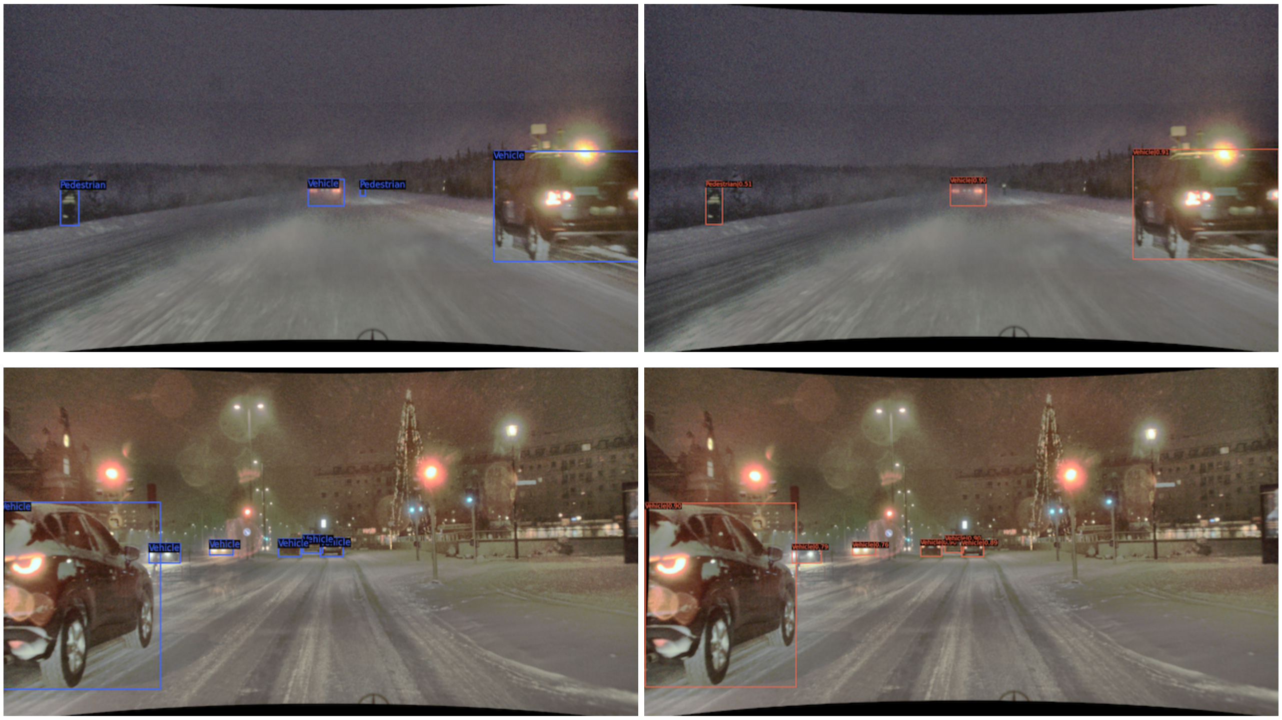


Figure 9: **Visualized results on STF snow night test split.** The left images are the ground truth, and the right images are the predictions of MT-DETR.