Supplementary Material for "Causal Mode Multiplexer: A Novel Framework for Unbiased Multispectral Pedestrian Detection"

A. Related Work

We provide a related work section in two-fold.

A.1. Multispectral Pedestrian Detection

Current research on multispectral pedestrian detection mainly focuses on developing effective fusion strategies. For instance, MBnet [13] adaptively fuses the RGBT complementary features according to illumination conditions. Recent works improve each modal feature through cross-modal learning. Kim et al. [5] proposed an uncertainty-aware feature fusion (UFF) network that alleviates miscalibration and modality discrepancy problems. Cross-Modality Fusion Transformer [9] introduces a new cross-modality fusion mechanism based on self-attention. However, they all suffer from the modality bias problem, constraining practical applicability.

A.2. Causality-Inspired Machine Learning

Causal inference and counterfactual reasoning encourage machines to explore causality behind observational likelihood, a proven and effective analytical approach in many machine learning problems [7, 11]. Several works leveraged counterfactual reasoning aiming to endow models with the capability to explore and understand causal effects. Niu et al. [8] mitigated the direct language effect on visual question answering (VQA) by guiding the model to learn the total indirect effect (TIE). Zhang et al. [12] introduced the layout-based soft Total Direct Effect (L-sTDE) to adjust the prediction of the navigation policy in object navigation. Different from them, we propose a Causal Mode Multiplexer (CMM) framework that interchangeably learns between total effect (TE) and total indirect effect (TIE) depending on the data type.

B. Comparison to Single-modal Models

Although our paper deals with modality bias problems in multispectral pedestrian detectors, we provide an evaluation of single-modal detectors to make our problem statement clear. Single-modal models refer to pedestrian detection models that operate with only one modality (e.g., only RGB or only thermal). We train the Faster-RCNN [10] model on the KAIST [4] train set and evaluate performance on the KAIST test set and ROTX-MP. Here, KAIST and ROTX-MP are multispectral pedestrian datasets, with RGB and T image pairs. We use only one of the RGBT pairs (e.g., only RGB for RGB single-modal detector) during train/test.

First of all, multispectral pedestrian detectors ([5], Ours) perform superiorly on KAIST data as these models use both RGB and T information. Low miss rates indicate higher performance. In particular, the Faster-RCNN trained with RGB data performs poorly at night as RGB sensors degrade in the dark. The thermal single-modal model performs well at night but degrades in the daytime. Compared to them, multispectral pedestrian detectors ([5], Ours) perform well on both day and night.

Second, RGB single-modal detector works well on ROTX-MP, as the dataset mainly contains pedestrians clearly visible in RGB but obscured in thermal. The thermal single-modal detector performs very poorly because most of the pedestrians are obscured in thermal. The conventional multispectral pedestrian detection model (Kim et al. [5]) performs poorly in ROTX-MP due to the modality bias. This model learns the statistical co-occurrence between the pedestrian and their thermal features, thus failing to detect pedestrians in ROTX as their thermal features are obscured. Compared to them, our CMM framework achieves high performance in ROTX-MP, effectively complementing RGB and T information through causality.

Overall, our CMM framework performs superiorly compared with single-modal models that leverage only RGB or a thermal sensor because CMM effectively fuses complementary information from RGB and T. Especially, CMM outperforms single-modal models where all day/night scenarios are required such as in KAIST data. Also, CMM achieves higher performance than RGB single-modal models in ROTX-MP in which RGB single-modal models can perform well. Moreover, our CMM framework solves the modality bias problem that persists in conventional models and performs well on ROTX-MP even when trained from a biased training data.

Table 1. Comparison to single-modal models which use only RGB or T. Models are trained on the KAIST data and tested on KAIST and ROTX-MP. **Best** results obtained are highlighted in bold.

Train		KAIST			
Test		KAIST			ROTX-MP
Metric		MR(↓)			AP(↑)
Model	Modality	Day	Night	All	All
Faster R-CNN	Only RGB	27.45	42.27	32.18	60.61
Faster R-CNN	Only T	26.70	9.73	20.79	5.33
Kim et al. [5]	RGB+T	10.11	**5.05**	8.67	21.69
CMM (Ours)	RGB+T	**9.60**	5.93	**8.54**	**70.44**

C. Implementation Details

For all models, the setting is kept the same across training KAIST [4], CVC [2], and FLIR [1] datasets.

CMM: The Uncertainty-guided model [5] is used as the baseline model. This model is designed with the Feature Pyramid Networks (FPN) with a backbone network of ResNet-50 [3]. Stochastic gradient descent (SGD) is used for optimization. The Pytorch library is used, and we used 4 GTX 1080 Ti GPUs for training the model. Each GPU processes 2 images, thus a total of 8 images are processed per mini-batch. We train the model for 2 epochs. The initial learning rate is 0.007 and there is a 0.1 learning rate decay for each epoch. The number of Region of Interests (RoIs) per image is set to 300.

Kim et al. [5]: The initial learning rate is 0.006 for the first 2 epochs. Then there is a 0.1 decay. We train the model for 3 epochs. The other implementation settings are identical to CMM.

Halfway Fusion [6]+Faster RCNN [10] (HFF): The initial learning rate is 0.008 for the first 2 epochs and there is a 0.1 learning rate decay. The model is trained for 3 epochs. We used the SGD optimizer is used.

CFT [9]: 0.01 initial learning rate, 0.937 momentum, 0.0005 weight decay, and 32 batch size. 200 epochs are trained we use the initial YOLO-v5 weight pre-trained on the COCO dataset. The code we used is from the GitHub page provided by the original authors. These implementation details are the same as the original work.

MBNet [13]: We use the same setting as the original paper. Resnet-50 trained on Imagenet is used as the backbone network. We used the official code from GitHub. The model is trained for 7 epochs. The learning rate is set to 0.0001 and batch size is 10. The Adam optimizer is used. We used the official code from GitHub.

D. No-treatment Condition

We introduce the implementation method of the no-treatment condition in Section 4.1. The no-treatment is defined as blocking (e.g., nullifying) the input from RGB or thermal. Denote the RGB feature as X_R and the thermal feature as X_T. Then we can write the no-treatment condition as $X_R = x_R* = \phi$ and $X_T = x_T* = \phi$. Note that neural networks can't deal with a no-treatment condition in which inputs have a null value. Under the no-treatment condition, we make the assumption that the neural model will sample inputs using a learnable parameter c which is initialized to zero. In this case, Y_{x_R}, Y_{x_T}, and Y_M can be represented as:

$$Y_{x_R} = \begin{cases} y_{x_R} = H_{\theta_{X_R}}(x_R) & \text{if } X_R = x_R \\ y_{x_R*} = c & \text{if } X_R = \phi \end{cases} \quad (1)$$

$$Y_{x_T} = \begin{cases} y_{x_T} = H_{\theta_{X_T}}(x_T) & \text{if } X_T = x_T \\ y_{x_T*} = c & \text{if } X_T = \phi \end{cases} \quad (2)$$

$$Y_m = \begin{cases} y_m = H_{\theta_M}(m) & \text{if } X_R = x_R \text{ and } X_T = x_T \\ y_{m*} = c & \text{if } X_R = \phi \text{ or } X_T = \phi. \end{cases} \quad (3)$$

E. More Visualized Results

We provide more visualized results in Fig.1-5. Fig.1-2 shows results on ROTX data. Fig.3 shows results on ROTO data. Fig.4 shows results on RXTO data. Fig.5 shows some failure cases.

Table 2. **Ilumination-aware weighting [a] can improve CMM**. CMM+I: CMM incorporated with illumination-aware weighting. The performances are evaluated on different datasets.

Train	KAIST				CVC-14				FLIR	
Test	ROTX-MP	KAIST			ROTX-MP	CVC-14			ROTX-MP	FLIR
Metric	AP (↑)	MR (↓)(Day/Night/All)			AP (↑)	MR (↓)(Day/Night/All)			AP (↑)	AP (↑)
CMM	70.44	9.60	**5.93**	8.54	34.96	27.81	7.71	17.13	57.09	87.80
CMM+I	**75.22**	**8.65**	6.72	**8.09**	**41.81**	**27.62**	**6.12**	**16.05**	**61.86**	**88.01**

Table 3. **Performance comparison with ProbEn [b]** on KAIST, CVC-14, FLIR, and ROTX-MP datasets.

Train	KAIST				CVC-14				FLIR	
Test	ROTX-MP	KAIST			ROTX-MP	CVC			ROTX-MP	FLIR
Metric	AP (↑)	MR (↓)(Day/Night/All)			AP (↑)	MR (↓)(Day/Night/All)			AP (↑)	AP (↑)
ProbEn [b]	18.8	9.93	**5.41**	**8.50**	16.64	**23.01**	21.02	22.23	15.98	87.65
CMM	**70.44**	**9.60**	5.93	8.54	**34.96**	27.81	**7.71**	**17.13**	**57.09**	**87.80**

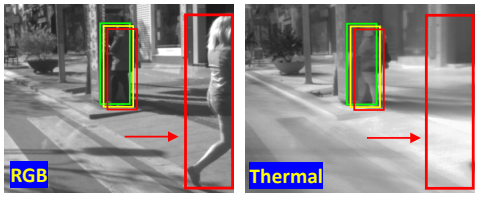


Figure i. We zoom in a frame of CVC-14-Day. The annotation (green) and conventional model ([14,27,40,19]) detections (yellow) miss the ROTX person (red arrow). Though CMM (red) can detect these miss-labeled objects, they will be evaluated as wrong results. This shows the issues in the CVC-14-Day dataset.

F. Rebuttal Questions & Answers

[dcFe] **Incorporate the illumination-aware weighting method [a] into CMM and evaluate the performance.**
The results are in Table i. It shows that CMM can be further improved by incorporating such techniques. We will cite the paper [a] and update these results and discussions.
[dcFe] **In Table 2, why does CMM underperform on CVC-14-Day compared to [14,27,40]?**
CVC-14 annotations are inconsistent across frames (misalignments & miss-labels), especially in CVC-14-Day. Such deficiencies make it hard for CMM to achieve better Miss-Rates (MR) than existing methods, despite making better detections. Figure i shows an example of this issue.
[dcFe] **Include comparison with ProbEn [b].**
Table ii shows the comparisons of CMM with ProbEn [b] (RGB+T version with v-avg). CMM outperforms ProbEn on the CVC-14, FLIR, and ROTX-MP datasets. Similar performances are obtained on KAIST. The results on the ROTX-MP highlight the generalizability of CMM. We will cite ProbEn [b] and update these results and discussions.
[dcFe] **Provide specific references for "Conventional models" in Figure 7.**
We will add the following reference: Kim et al. [14].
[besY] **As shown in Table 1, CMM already senses the ROTX feature pattern. It is not fair to compare CMM with previous detectors on the ROTX-MP dataset.**
K_{mode} (Eq.10) values in Table 1 are depicted in a consequential perspective, i.e., what value will K_{mode} obtain *"if ROTX is given as input?"*. This does not mean that CMM senses ROTX pattern to determine K_{mode}. To make this clear, we point out that our training objective (Eq.11) with respect to K_{mode} is based on sensing the modality discrepancy (whether $\Delta\pi_R$ and $\Delta\pi_T$ have the same value or not), and not ROTX.
[besY] **The terms of ROTO, RXTO, and ROTX should be explained at the very beginning of the paper.**
We will move the footnote regarding terms to the abstract.
[besY] **How many neural networks are contained in CMM? Are they trained jointly or separately?**
CMM jointly trains three sub-networks (two uni-modal, one multi-modal) sharing the same RPN and the head network.
[besY] **What is the computational cost of CMM? Is the inference burden higher than the previous method?**
For CMM, the inference speed is 0.11 (s) per image, and the number of parameters is 99.5M. The baseline model Kim et al. [14] shows 0.09 (s) and 71.7M, respectively.
[besY] **Does the performance gain mainly from the designed CMM framework or extra computational cost?**
The performance gain is from learning causality, especially the switchable total indirect effect (sTIE) we proposed. Table 3 and the supplementary Figures 1-4 demonstrate the effectiveness of learning sTIE.
[besY] **How do you fuse the RGB and T image in CMM?**
RGB and T features are extracted from each encoder, and mid-fused by concatenating them inside the network.
[besY] **The backbones of CMM and previous works are diverse, which should be debiased for a fair comparison.**
We used the same Resnet-50 for CMM and the comparison methods [14,40,19] (same as original papers).
[vqAR] **In ablation studies, it seems unreasonable that MR becomes worse when TE+TIE is added to baseline.**
TE+TIE performs counterfactual intervention for all test cases. However, for ROTO data in which RGB and T both perform well, subtracting the thermal direct effect results in low confidence scores. Such effects make TE+TIE have worse MR on general datasets (KAIST, CVC-14, and FLIR) that largely contain ROTO data.
[vqAR] **The model based only on RGB data achieves a good AP on the ROTX-MP dataset. Is the ROTX-MP dataset largely influenced by RGB rather than both?**
Yes. ROTX-MP is largely influenced by RGB, rather than both, i.e., complementary to existing RGBT datasets in which thermal is more influential. Thus, models using only RGB can perform well. RGBT models, ideally, should be able to perform well on ROTX-MP based on their RGB signals. But in practice, they learn modality biases toward thermal, making them vulnerable to ROTX data. ROTX-MP is proposed to evaluate such vulnerabilities. CMM leverages *causality* to resolve this issue (Table 4) and retains the advantages of the RGBT models (Table 2).
(Additional References) [a] Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection. Information Fusion, 2019.
[b] Multimodal object detection via probabilistic ensembling. ECCV, 2022.

References

[1] Inc FLIR Systems. Free teledyne flir thermal dataset for algorithm training. https://www.flir.com/oem/adas/adas-dataset-form/, 2021. Accessed: 2022-08-05. 2

[2] Alejandro González, Zhijie Fang, Yainuvis Socarras, Joan Serrat, David Vázquez, Jiaolong Xu, and Antonio M López. Pedestrian detection at day/night time with visible and fir cameras: A comparison. *Sensors*, 16(6):820, 2016. 2

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[4] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1037–1045, 2015. 1, 2

[5] Jung Uk Kim, Sungjune Park, and Yong Man Ro. Uncertainty-guided cross-modal learning for robust multispectral pedestrian detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3):1510–1523, 2022. 1, 2

[6] Jingjing Liu, Shaoting Zhang, Shu Wang, and Dimitris N Metaxas. Multispectral deep neural networks for pedestrian detection. *arXiv preprint arXiv:1611.02644*, 2016. 2

[7] Miguel Monteiro, Fabio De Sousa Ribeiro, Nick Pawlowski, Daniel C Castro, and Ben Glocker. Measuring axiomatic soundness of counterfactual image models. In *The Eleventh International Conference on Learning Representations*, 2022. 1

[8] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12700–12710, 2021. 1

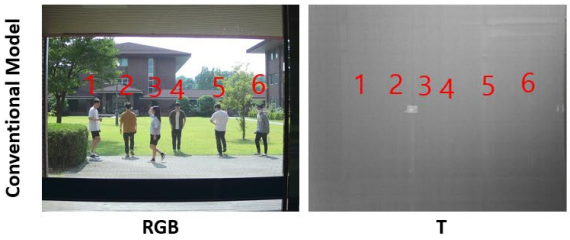
[9] Fang Qingyun, Han Dapeng, and Wang Zhaokui. Cross-modality fusion transformer for multispectral object detection, 2022. 1, 2

[10] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 1, 2

[11] Alexander P Wu, Rohit Singh, and Bonnie Berger. Granger causal inference on dags identifies genomic loci regulating transcription. In *International Conference on Learning Representations*, 2021. 1

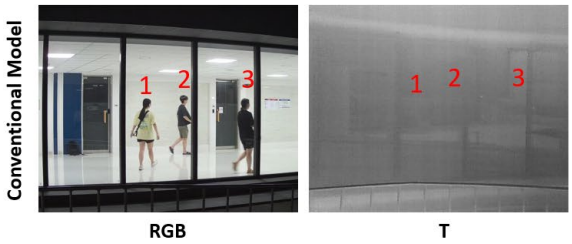
[12] Sixian Zhang, Xinhang Song, Weijie Li, Yubing Bai, Xinyao Yu, and Shuqiang Jiang. Layout-based causal inference for object navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10792–10802, 2023. 1

[13] Kailai Zhou, Linsen Chen, and Xun Cao. Improving multispectral pedestrian detection by addressing modality imbalance problems. In *European Conference on Computer Vision*, pages 787–803. Springer, 2020. 1, 2

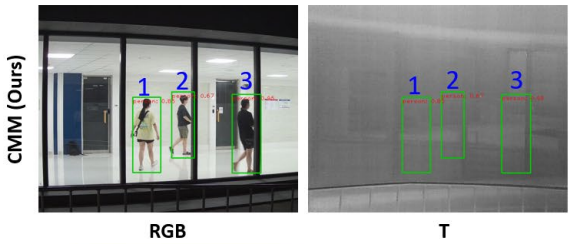


	Conf
index	
person 1	< 0.05
person 2	< 0.05
person 3	< 0.05
person 4	< 0.05
person 5	< 0.05
person 6	< 0.05

	Type (K_mode)	TE		TE_d	Conf (Before)	NDE		NDE_d	sTIE		sTIE_d	CE*	Conf (After)
index		0	1			0	1		0	1			
person 1	ROTX (-1)	-5.64	-7.73	-2.09	0.11	-1.39	-5.49	-4.10	-4.25	-2.24	2.01	= TIE	0.88
person 2	ROTX (-1)	-6.47	-6.86	-0.39	0.40	-1.39	-5.04	-3.65	-5.09	-1.82	3.26	= TIE	0.96
person 3	ROTX (-1)	-5.85	-7.44	-1.60	0.17	-1.39	-4.95	-3.56	-4.46	-2.50	1.96	= TIE	0.88
person 4	ROTX (-1)	-4.15	-7.94	-3.79	0.02	-1.39	-5.18	-3.79	-2.76	-2.76	0.00	= TIE	0.50
person 5	ROTX (-1)	-4.18	-7.41	-3.23	0.04	-1.39	-5.09	-3.71	-2.79	-2.31	0.48	= TIE	0.62
person 6	ROTX (-1)	-5.67	-6.80	-1.13	0.24	-1.39	-5.02	-3.63	-4.28	-1.78	2.50	= TIE	0.92



	Conf
index	
person 1	< 0.05
person 2	< 0.05
person 3	< 0.05



	Type (K_mode)	TE		TE_d	Conf (Before)	NDE		NDE_d	sTIE		sTIE_d	CE*	Conf (After)
index		0	1			0	1		0	1			
person 1	ROTX (-1)	-5.94	-5.73	0.21	0.55	-1.39	-4.21	-2.83	-4.55	-1.52	3.03	=TIE	0.95
person 2	ROTX (-1)	-4.37	-7.39	-3.02	0.05	-1.39	-5.14	-3.75	-2.98	-2.25	0.73	=TIE	0.67
person 3	ROTX (-1)	-4.78	-6.05	-1.26	0.22	-1.39	-4.40	-3.02	-3.40	-1.64	1.75	=TIE	0.85

	Conf
person 1	< 0.05
person 2	< 0.05
person 3	< 0.05
person 4	< 0.05

	Type (K_mode)	TE		TE_d	Conf (Before)	NDE		NDE_d	sTIE		sTIE_d	CE*	Conf (After)
index		0	1			0	1		0	1			
person 1	ROTX (-1)	-3.97	-6.77	-2.80	0.06	-1.39	-4.65	-3.26	-2.59	-2.13	0.46	=TIE	0.61
person 2	ROTX (-1)	-4.26	-6.20	-1.94	0.13	-1.39	-5.23	-3.84	-2.87	-0.97	1.90	=TIE	0.87
person 3	ROTX (-1)	-3.38	-7.22	-3.85	0.02	-1.39	-5.32	-3.93	-1.99	-1.90	0.09	=TIE	0.52
person 4	ROTX (-1)	-4.88	-5.83	-0.95	0.28	-1.39	-4.80	-3.42	-3.49	-1.03	2.46	=TIE	0.92

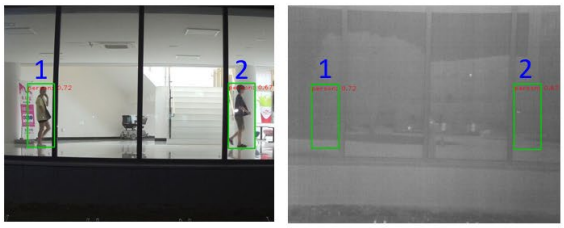
Figure 2. **Visualized examples of multispectral pedestrian detection on ROTX data**. The table on the right indicates each value for TE, NDE, sTIE, and confidence scores. We compare the conventional model and CMM (ours).

Conventional Model

RGB T

	Conf
index	
person 1	< 0.05
person 2	< 0.05

CMM (Ours)



RGB T

	Type (K_mode)	TE		TE_d	Conf (Before)	NDE		NDE_d	sTIE		sTIE_d	CE*	Conf (After)
index		0	1			0	1		0	1			
person 1	ROTX (-1)	-4.57	-6.49	-1.92	0.13	-1.39	-4.28	-2.89	-3.19	-2.22	0.97	=TIE	0.72
person 2	ROTX (-1)	-4.22	-6.12	-1.90	0.13	-1.39	-4.01	-2.63	-2.83	-2.11	0.73	=TIE	0.67

Conventional Model

RGB T

	Conf
person 1	1.00
person 2	0.98
person 3	0.43

CMM (Ours)

RGB T

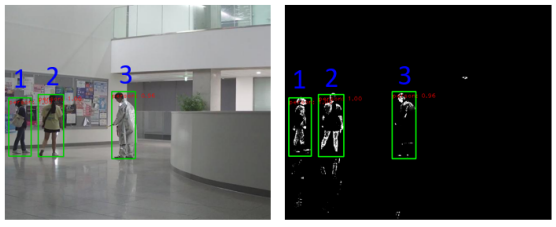
	Type (K_mode)	TE		TE_d	Conf (Before)	NDE		NDE_d	sTIE		sTIE_d	CE*	Conf (After)
index		0	1			0	1		0	1			
person 1	ROTO (1)	-11.16	-0.85	10.31	1.00	-4.86	-1.39	3.46	-11.16	-0.85	10.31	=TE	1.00
person 2	ROTO (1)	-7.62	-0.49	7.13	1.00	-3.42	-1.46	1.96	-7.62	-0.49	7.13	=TE	1.00
person 3	ROTX (-1)	-3.96	-3.66	0.29	0.57	-1.39	-2.94	-1.55	-2.57	-0.72	1.85	=TIE	0.86

Conventional Model

RGB T

	Conf
person 1	0.97
person 2	1.00
person 3	0.41

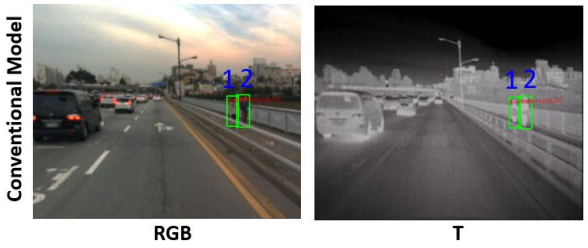
CMM (Ours)



RGB T

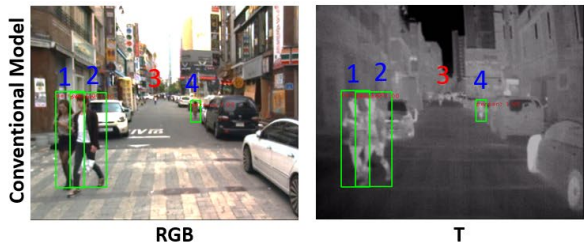
	Type (K_mode)	TE		TE_d	Conf (Before)	NDE		NDE_d	sTIE		sTIE_d	CE*	Conf (After)
index		0	1			0	1		0	1			
person 1	ROTO (1)	-10.22	-0.48	9.74	1.00	-3.92	-1.43	2.49	-10.22	-0.48	9.74	=TE	1.00
person 2	ROTO (1)	-7.42	-2.76	4.66	0.99	-1.39	-2.97	-1.58	-6.02	0.21	6.24	=TE	1.00
person 3	ROTX (-1)	-4.97	-4.20	0.77	0.68	-1.39	-3.81	-2.43	-3.58	-0.38	3.20	=TIE	0.96

Figure 3. **Visualized examples of multispectral pedestrian detection on ROTX data**. The table on the right indicates each value for TE, NDE, sTIE, and confidence scores. We compare the conventional model and CMM (ours).



	Conf
person 1	0.95
person 2	1.00

	Type (K_mode)	TE		TE_d	Conf (Before)	NDE		NDE_d	sTIE		sTIE_d	CE*	Conf (After)
index		0	1			0	1		0	1			
person 1	ROTO (1)	-9.60	-0.36	9.24	1.00	-5.40	-1.42	3.99	-9.60	-0.36	9.24	=TE	1.00
person 2	ROTO (1)	-7.98	-0.47	7.52	1.00	-4.21	-1.40	2.81	-7.98	-0.47	7.52	=TE	1.00



	Conf
person 1	1.00
person 2	1.00
person 3	0.43
person 4	1.00

	Type (K_mode)	TE		TE_d	Conf (Before)	NDE		NDE_d	sTIE		sTIE_d	CE*	Conf (After)
index		0	1			0	1		0	1			
person 1	ROTO (1)	-9.24	-0.29	8.94	1.00	-4.73	-1.39	3.34	-9.24	-0.29	8.94	=TE	1.00
person 2	ROTO (1)	-9.06	-0.38	8.68	1.00	-4.35	-1.40	2.94	-9.06	-0.38	8.68	=TE	1.00
person 3	ROTO (1)	-9.27	-0.61	8.66	1.00	-3.77	-1.40	2.38	-9.27	-0.61	8.66	=TE	1.00
person 4	ROTO (1)	-4.78	-1.70	3.07	0.96	-1.98	-1.53	0.45	-4.78	-1.70	3.07	=TE	0.96

	Conf
person 1	1.00
person 2	1.00
person 3	1.00

	Type (K_mode)	TE		TE_d	Conf (Before)	NDE		NDE_d	sTIE		sTIE_d	CE*	Conf (After)
index		0	1			0	1		0	1			
person 1	ROTO (1)	-12.70	-0.43	12.27	1.00	-6.21	-1.39	4.82	-12.70	-0.43	12.27	=TE	1.00
person 2	ROTO (1)	-10.04	-0.41	9.63	1.00	-5.05	-1.40	3.66	-10.04	-0.41	9.63	=TE	1.00
person 3	ROTO (1)	-9.72	-0.43	9.30	1.00	-5.67	-1.40	4.27	-9.72	-0.43	9.30	=TE	1.00

Figure 4. **Visualized examples of multispectral pedestrian detection on ROTO data**. The table on the right indicates each value for TE, NDE, sTIE, and confidence scores. We compare the conventional model and CMM (ours).

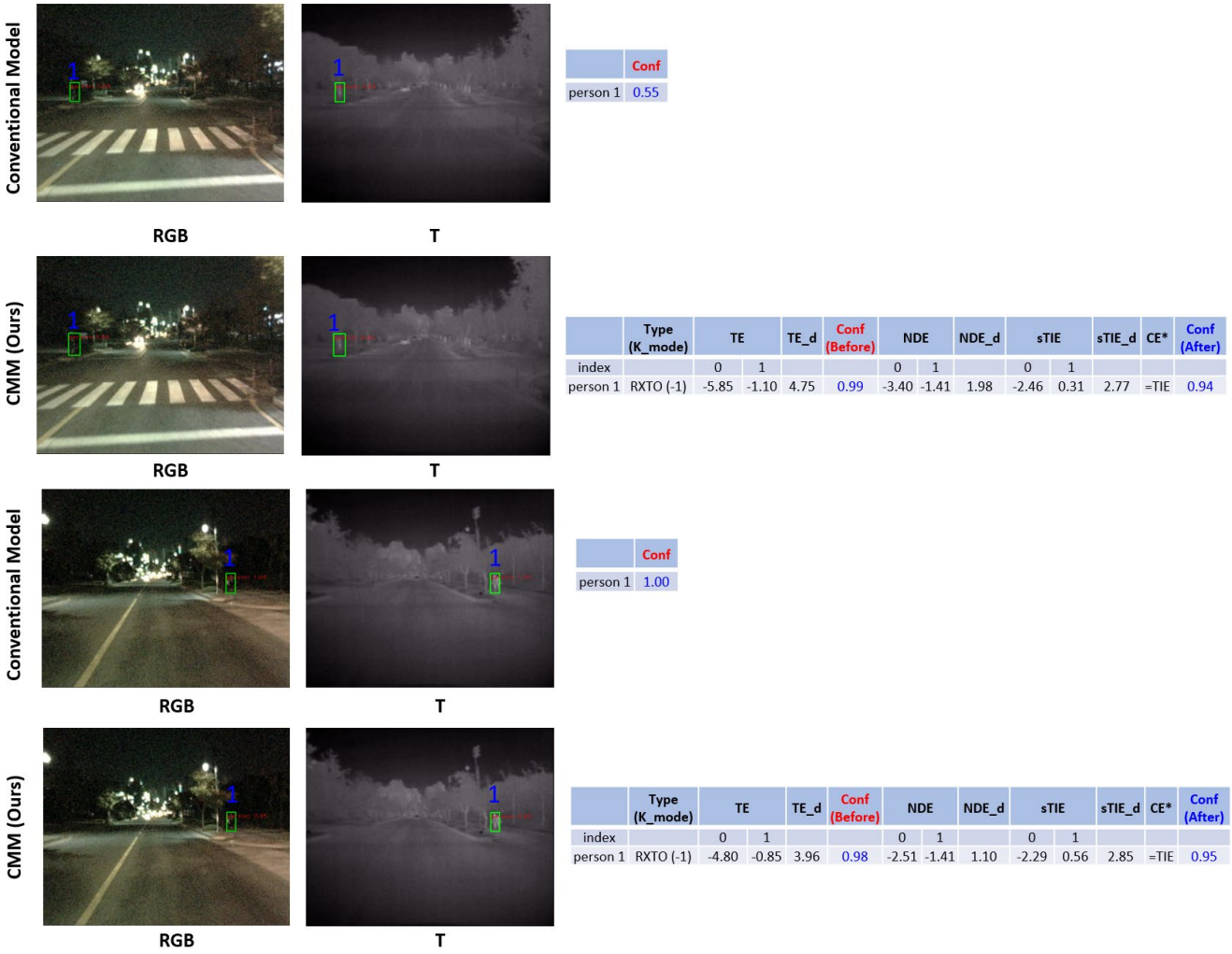


Figure 5. **Visualized examples of multispectral pedestrian detection on RXTO data**. The table on the right indicates each value for TE, NDE, sTIE, and confidence scores. We compare the conventional model and CMM (ours).

Figure 6. Failure cases of CMM due to occlusion.