MMVIP: A Visible-infrared Paired Dataset for Multi-weather Marine Vision

Supplementary Material

1. More MMVIP Datasets

The detailed category distribution of the MMVIP dataset is illustrated in Fig. 1, encompassing a wide variety of scenarios, viewpoints, and weather conditions. Given the significant disparities in data acquisition difficulty under varying meteorological conditions, the number of available samples for certain extreme weather events (e.g., typhoons) remains relatively limited. Regarding data acquisition methods, sea scenarios primarily rely on unmanned aerial vehicles (UAVs) and shipborne electro-optical pods, whereas coastal and port scenarios predominantly utilize fixed electro-optical pods. Notably, our dataset was collected across various seasons, times of day, and geographic locations to minimize incidental bias.

To provide a more intuitive visualization of the dataset's characteristics, Fig. 2 presents additional examples of visible-infrared image pairs captured in diverse, real-world maritime environments. Acquired using various types of electro-optical equipment, these image pairs cover a broad spectrum of illumination conditions, complex and fluctuating weather states, and diverse target scales. These rich qualitative samples not only highlight the visual breadth of the MMVIP dataset but also fully validate its comprehensive coverage and representational capability for authentic, complex marine environments.

2. Supplementary Experiment of Image Fusion

2.1. Additional Fusion Results

To further demonstrate the advantage of multimodal data, we present image fusion results from various methods. These results illustrate the performance gains in different scenes and help evaluate the applicability and challenges of the MMVIP dataset for fusion tasks.

As shown in Fig. 3a, Fig. 3b, and Fig. 3c, the visual comparisons of different fusion methods under night, foggy, and low-light scenarios are presented. In night scenes, visible images typically suffer from low brightness and significant noise, while infrared images highlight target structures but lack detailed textures. Under such conditions, different fusion methods exhibit notable differences in brightness enhancement, target highlighting, and noise suppression. In foggy scenes, visible images generally exhibit low contrast and struggle to present the complete contours of distant targets, whereas fusion methods can leverage the infrared modality to preserve clearer edge structures. In low-light scenarios, some methods tend to suffer from issues such as over-enhancement or loss of details, whereas more stable fusion models achieve a better balance among texture preservation, brightness restoration, and target visibility.

Collectively, the proposed MMVIP dataset establishes a critical benchmark for maritime fusion research. Furthermore, our results expose the core challenges inherent to these complex environments.

2.2. Quantitative Analysis of Fusion Results

This section employs six quantitative metrics to evaluate ten representative fusion methods. The evaluated methods include CAF [11], EMMA [32], Text-IF [28], LUT-Fuse [29], RFfusion [24], TG-ECNet [21], DCEvo [12], GIFNet [3], SAGE [25], and TDFusion [1].

We employ the following metrics for fusion quality assessment: EN [18] (information content), MI [16] (mutual information with source images), SD [17] (image contrast), VIF [7] (visual fidelity), Qabf [26] (salient information preservation), and SSIM [23] (structural similarity). Higher values indicate better performance for all metrics.

As shown in Tab. 1, the quantitative evaluation results of different image fusion methods on the MMVIP dataset are presented. It can be observed that RFfusion [24] achieves the highest values on the EN and SD metrics, indicating that its fused images contain richer detail information. DCEvo [12] performs best in terms of MI, suggesting that it preserves complementary information from the source images more effectively. For the VIF metric, TDFusion [1] attains the highest score, demonstrating strong capability in maintaining visual clarity. In addition, TDFusion also performs well on SSIM, indicating that its fused results are most similar to the source images in terms of structural and perceptual characteristics. In contrast, Text-IF [28] obtains the highest value on Qabf, showing superior performance in balancing edge and contrast information during fusion.

In summary, although existing fusion algorithms have achieved promising results, there remains substantial room for further improvement.

2.3. Comparison of Image Enhancement and Image Fusion

To comprehensively demonstrate the benchmark value and challenges of the MMVIP dataset, we evaluate state-of-the-art models across diverse visual tasks, including low-light enhancement and image dehazing, along with the corresponding fusion results. Specifically, we provide both qualitative visual comparisons and quantitative evaluations using no-reference image quality assessment metrics. This combination of subjective and objective analysis allows for a

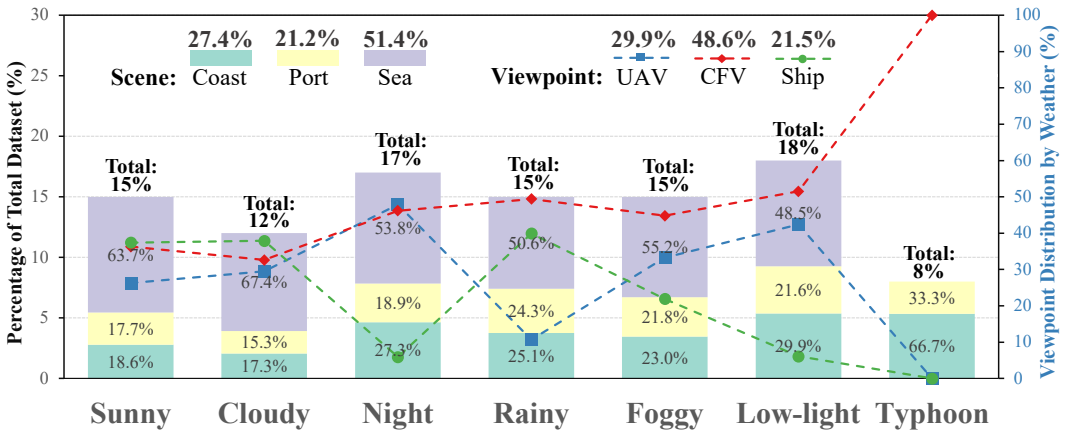

Figure 1. Distribution of scene, viewpoint, and weather conditions in the MMVIP dataset.

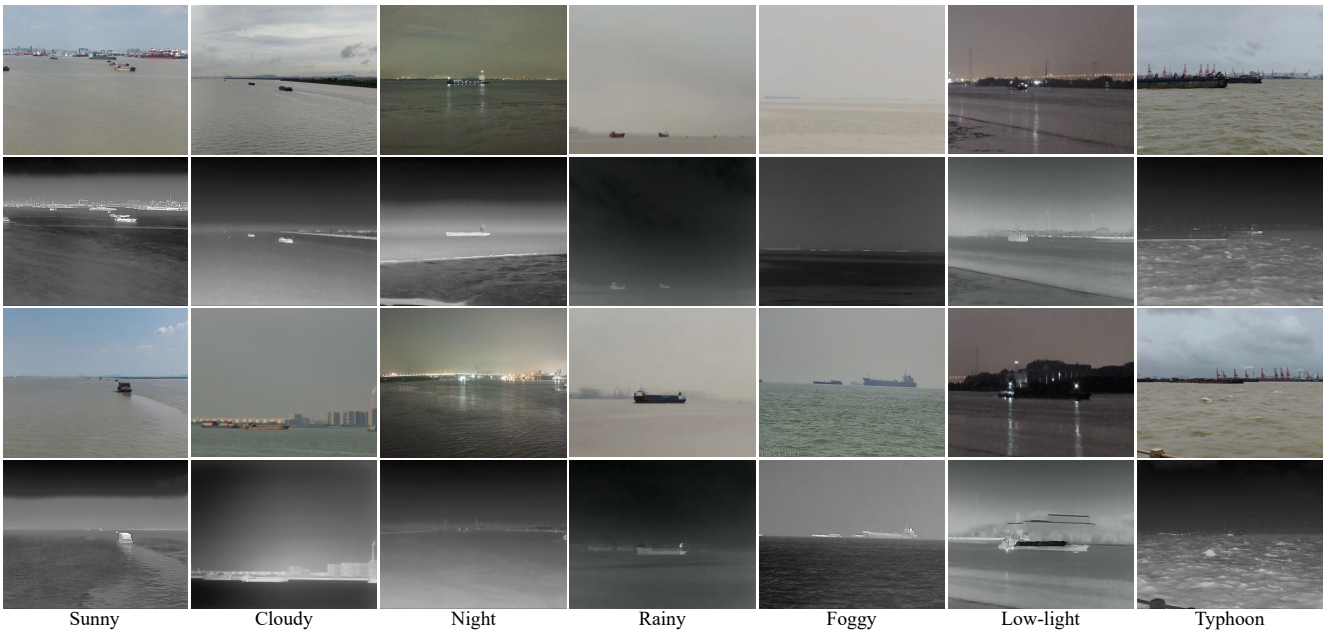

Figure 2. Additional scene examples from the MMVIP dataset.

thorough assessment of model performance under complex maritime scenarios.

2.3.1. Low-light Image Enhancement

For experimental comparison, we select five representative low-light enhancement methods: EnlightenGAN [9], R2RNet [6], DENet [8], DarkIR [4], and CIDNet [27].

As shown in Fig. 4, visual comparisons are presented between low-light enhancement methods and visible-infrared fusion methods under the same scenarios. The first row corresponds to low-light enhancement models, which primarily improve visibility by increasing brightness and contrast but are prone to amplified noise, overexposed details, or texture loss. The second row shows the fusion methods, which leverage the structural advantages of the infrared modality to maintain more stable contours in poorly illuminated regions while reducing common texture artifacts seen in enhancement models. By examining the highlighted and dark regions in the comparison images, it is evident that fusion models exhibit stronger target discernibility and structural consistency under night and low-light conditions.

Tab. 2 summarizes the quantitative comparisons of five low-light enhancement methods and visible-infrared fusion approaches across four no-reference image quality metrics: NIQE [15], BRISQUE [14], CLIP-IQA [22], and LIQE [30]. The results reveal that low-light enhancement methods exhibit certain fluctuations in terms of naturalness and contrast improvement, while several fusion methods

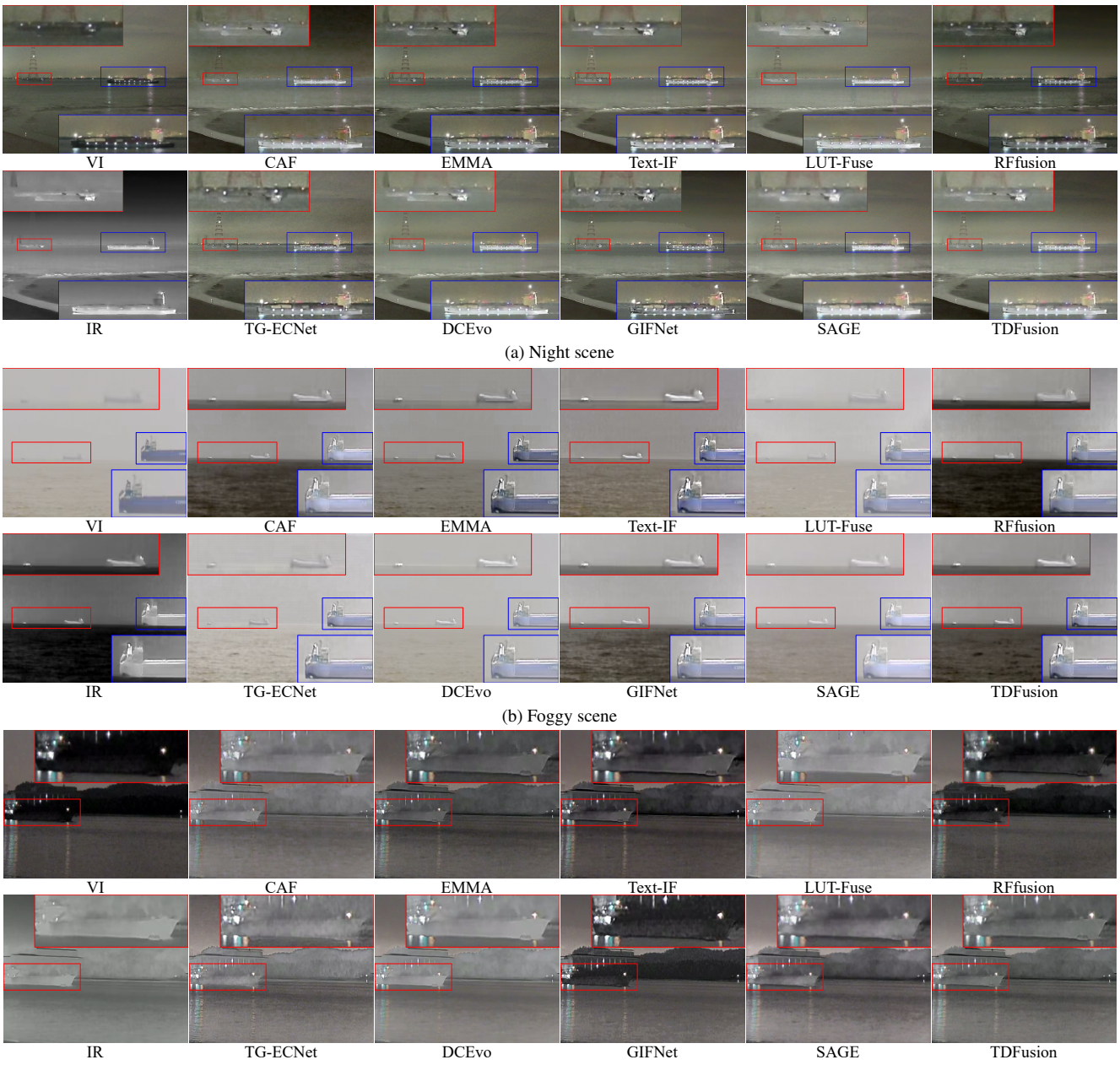

Figure 3. Comparison of fusion results of three typical marine conditions.

achieve superior or near-optimal performance across multiple metrics, indicating better structural preservation and content consistency. Overall, fusion methods provide more stable image quality in complex low-light environments, which aligns with the qualitative comparisons.

2.3.2. Image Dehazing Experiment

For image dehazing, we evaluate five representative methods: DehazeFormer [20], DEA-Net [2], CoA [13], IPC-Dehaze [5], and SFMN [19].

As shown in Fig. 5, we present visual comparisons between single-image dehazing approaches and visible–infrared fusion methods. Under heavy haze, visible images typically suffer from severe contrast degradation and blurred distant targets, and existing dehazing models may further introduce color shifts or over-sharpening artifacts. In contrast, fusion methods leverage the stable structural cues available in the infrared modality, enabling clearer preservation of distant contours and effectively suppressing

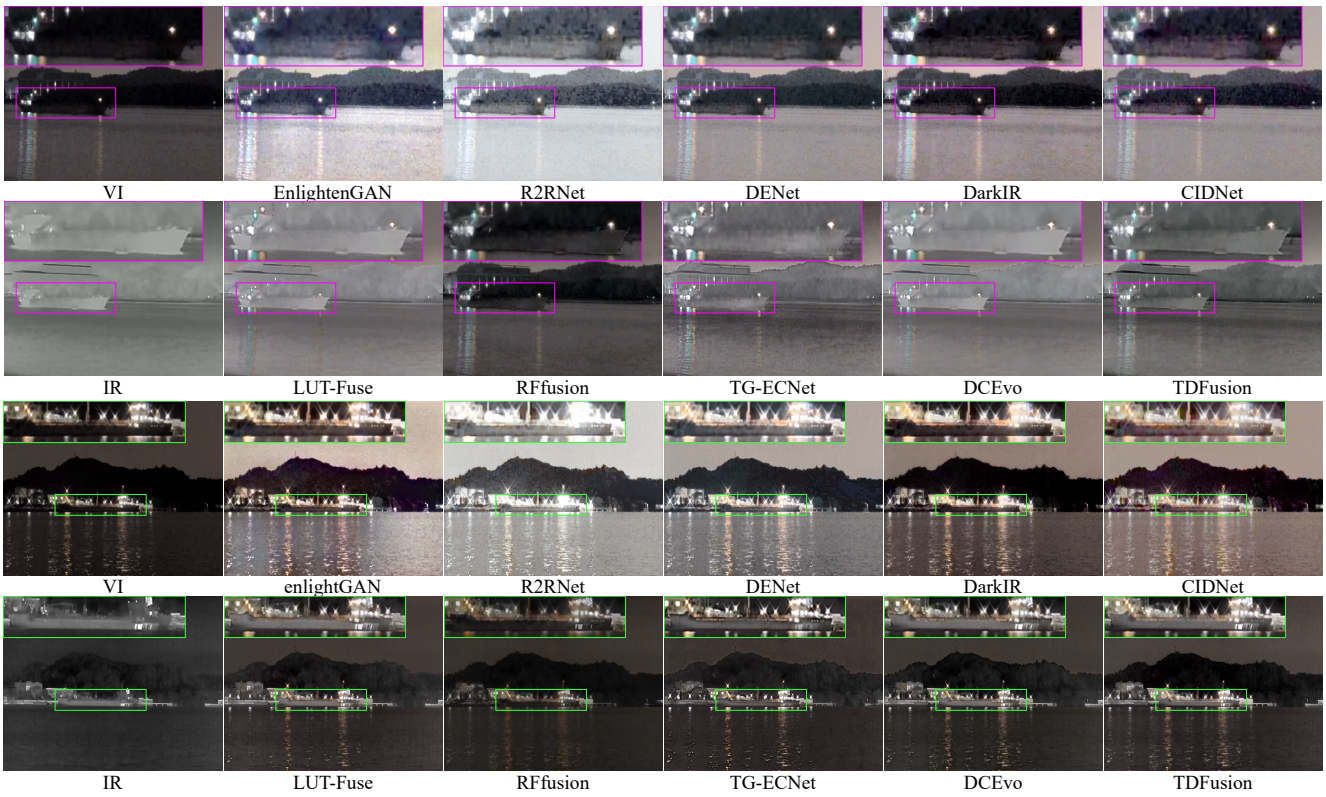

Figure 4. Visual comparison of low-light enhancement and selected image fusion methods. The first and third rows display the enhancement results, while the remaining rows show the fusion results.

Table 1. The quantitative indicators of various fusion methods in the MMVIP dataset, with Red indicating the best result, Blue representing the second-best result, and Green signifying the third-best result.

Methods	EN	MI	SD	VIF	Qabf	SSMI
CAF[11]	6.064	3.060	23.569	1.058	0.455	0.701
EMMA[32]	6.259	4.150	27.602	1.281	0.506	0.677
Text-IF[28]	6.125	3.797	23.839	1.264	0.580	0.824
Lut-Fuse[29]	5.851	4.579	21.620	1.088	0.518	0.775
RFfusion[24]	6.515	3.909	31.601	1.260	0.489	0.826
TG-ECNet[21]	6.471	2.879	28.213	1.041	0.369	0.450
DCEvo[12]	5.876	4.755	21.987	1.190	0.564	0.864
GIFNet[3]	6.238	3.165	26.210	1.181	0.445	0.763
SAGE[25]	5.851	3.059	22.718	1.189	0.536	0.884
TDFusion[1]	6.278	3.149	27.245	1.312	0.567	0.903

artifacts caused by aggressive enhancement. By examining the behavior of different approaches in dense haze regions and along fine structural boundaries, it is evident that fusion-based strategies yield more stable, artifact-free, and visually natural results in challenging foggy scenes.

Tab. 3 reports the performance of various dehazing and visible–infrared fusion methods on four no-reference image quality metrics: NIQE [15], BRISQUE [14], CLIP-IQA [22], and LIQE [30]. Dehazing approaches show differing levels of effectiveness in reducing blur and improving clarity; for instance, IPC-Dehaze [5] performs well on BRISQUE, yet its overall perceptual quality remains constrained by the limitations of single-modality information. In contrast, fusion-based methods exhibit greater stability and consistency across all metrics. Notably, models such as LUT-Fuse [29] and TDFusion [1] achieve state-of-

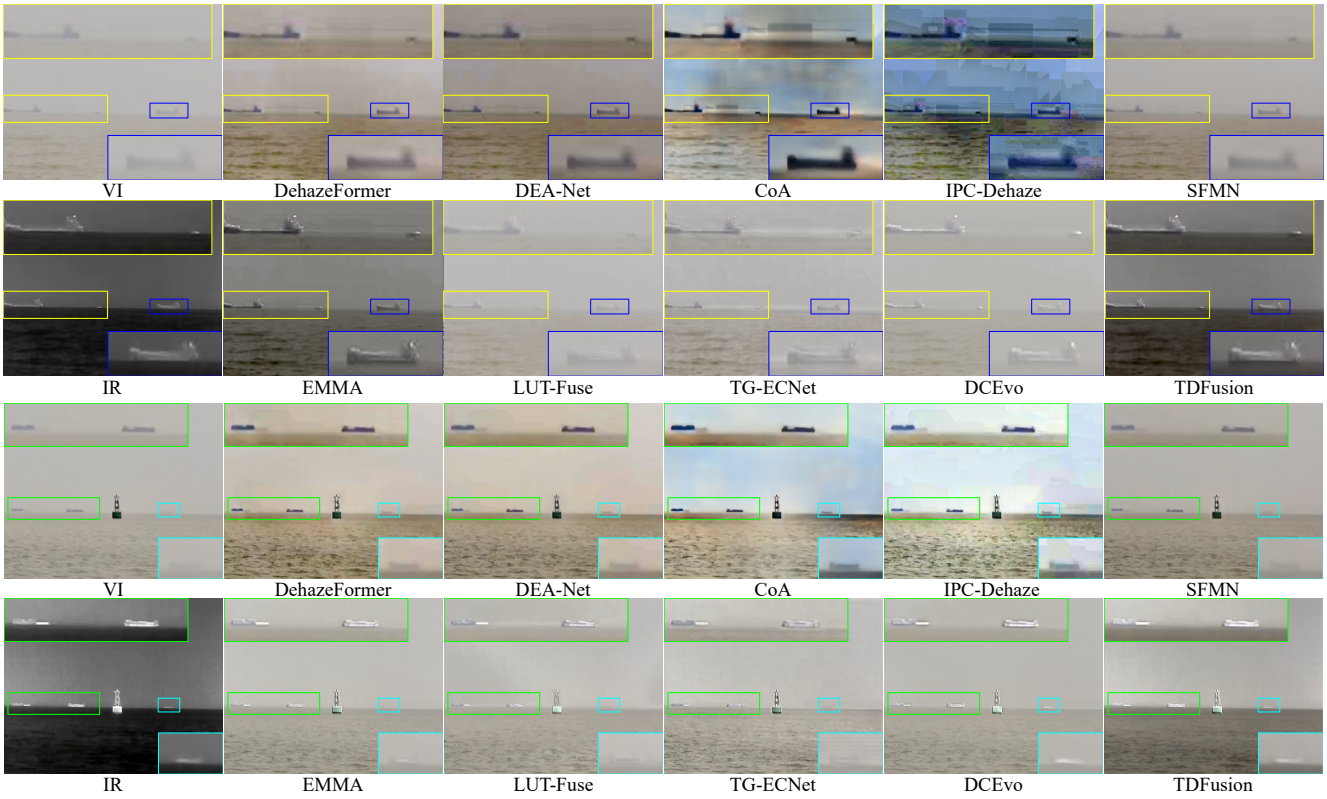

Figure 5. Visual comparison of image dehazing and selected fusion methods. The first and third rows display dehazing outputs, while the remaining rows show the results of various fusion methods.

Table 2. Performance comparison of various low-light enhancement and fusion methods on non-reference IQA metrics (NIQE, BRISQUE, CLIP-IQA, LIQE). Red indicates the best result. Blue indicates the second-best result. Green indicates the third-best result.

Type	Methods	NIQE↓	BRISQUE↓	CLIP-IQA↑	LIQE↑
	Visible	4.191	43.933	0.145	1.025
Enhance	EnlightGAN[9]	4.372	41.405	0.126	1.005
	R2RNet[6]	4.254	49.240	0.324	1.022
	DENet[8]	4.645	45.097	0.210	1.014
	DarkIR[4]	4.179	48.391	0.308	1.025
	CIDNet[27]	4.130	39.280	0.137	1.016
Fusion	CAF[11]	3.643	32.712	0.178	1.082
	EMMA[32]	5.079	22.021	0.180	1.062
	Text-IF[28]	4.118	38.397	0.208	1.083
	LUT-Fuse[29]	2.953	35.018	0.370	1.117
	RFfusion[24]	3.564	36.430	0.217	1.036
	TG-ECNet[21]	5.977	16.577	0.123	1.029
	DCEVo[12]	3.789	37.877	0.198	1.091
	GIFNet[3]	4.122	37.030	0.200	1.057
	SAGE[25]	3.933	43.023	0.251	1.060
	TDFusion[1]	3.910	34.146	0.202	1.061

Table 3. Performance comparison of various dehazing and fusion methods across no-reference image quality assessment metrics (NIQE, BRISQUE, CLIP-IQA, LIQE). Red indicates the best result. Blue indicates the second-best result. Green indicates the third-best result.

Type	Methods	NIQE↓	BRISQUE↓	CLIP-IQA↑	LIQE↑
	Visible	4.797	63.239	0.245	1.293
Dehaze	DehazeFormer[20]	5.016	40.692	0.161	1.083
	DEA-Net[2]	5.206	39.450	0.149	1.091
	CoA[13]	5.125	44.267	0.156	1.067
	IPC-Dehaze[5]	4.406	33.937	0.119	1.090
	SFMN[19]	5.013	63.507	0.301	1.117
Fusion	CAF[11]	4.139	45.588	0.268	1.363
	EMMA[32]	7.222	33.290	0.262	1.484
	Text-IF[28]	4.459	40.389	0.193	1.301
	LUT-Fuse[29]	4.137	56.634	0.373	1.578
	RFfusion[24]	4.052	46.089	0.282	1.199
	TG-ECNet[21]	8.426	43.537	0.196	1.322
	DCEVo[12]	4.379	50.811	0.314	1.591
	GIFNet[3]	4.740	44.225	0.323	1.233
	SAGE[25]	4.450	48.666	0.324	1.341
	TDFusion[1]	4.005	41.996	0.264	1.329

Figure 6. Visualized tracking results of ByteTrack on the test sequences. The rows, from top to bottom, illustrate four different maritime sequences.

the-art performance on CLIP-IQA and LIQE, underscoring their strengths in preserving structural details and enhancing semantic content. These results demonstrate that fusion strategies offer superior visual fidelity and robustness under complex weather conditions.

The above experimental results indicate that under complex maritime conditions, such as low light and fog, image fusion methods outperform single-modality enhancement or dehazing approaches in both visual quality and quantitative metrics. This fully demonstrates the effectiveness of multi-modal collaboration in improving visual perception performance in challenging marine environments. These findings also highlight the significance of the proposed MMVIP dataset. It provides a unified benchmark for related research areas, including maritime low-light enhancement and dehazing, playing a positive role in advancing maritime visual perception studies.

3. Multi-Object Tracking Experiment

For the object tracking task, our MMVIP dataset also serves as an evaluation benchmark. Specifically, we evaluate ByteTrack [31] by employing YOLO11 [10] as the baseline detector. Using the original pre-trained weights as a baseline, we fine-tuned on MMVIP, splitting the video sequences 9:1 for training and testing. The model was trained for 100 epochs with AdamW and a batch size of 32.

Qualitative Comparisons. Fig. 6 illustrates the performance of ByteTrack [31] in complex maritime environments. As shown in the figure, ByteTrack exhibits a clear tendency toward missed detections under low-visibility conditions, especially for small targets such as buoys and small vessels. In the presence of occlusions, its identity association capability further degrades, leading to frequent and pronounced ID switches. The ByteTrack model fine-tuned on the MMVIP dataset achieves moderate performance in maritime multi-object tracking. Its re-identification ability remains limited, particularly in scenarios with frequent occlusions.

The qualitative evaluations indicate that MMVIP serves as a high-quality, challenging benchmark for the training and evaluation of maritime multi-object tracking models, further highlighting its critical role in advancing research within this domain.

References

[1] Haowen Bai, Jiangshe Zhang, Zixiang Zhao, Yichen Wu, Lilun Deng, Yukun Cui, Tao Feng, and Shuang Xu. Task-driven image fusion with learnable fusion loss. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7457–7468, 2025. 1, 4, 5, 6

[2] Zixuan Chen, Zewei He, and Zhe-Ming Lu. Dea-net: Single image dehazing based on detail-enhanced convolution and content-guided attention. *IEEE transactions on image processing*, 33:1002–1015, 2024. 3, 6

[3] Chunyang Cheng, Tianyang Xu, Zhenhua Feng, Xiaojun Wu, Zhangyong Tang, Hui Li, Zeyang Zhang, Sara Atito, Muhammad Awais, and Josef Kittler. One model for all: Low-level task interaction is a key to task-agnostic image fusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28102–28112, 2025. 1, 4, 5, 6

[4] Daniel Feijoo, Juan C Benito, Alvaro Garcia, and Marcos V Conde. Darkir: Robust low-light image restoration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10879–10889, 2025. 2, 5

[5] Jiayi Fu, Siyu Liu, Zikun Liu, Chun-Le Guo, Hyunhee Park, Ruiqi Wu, Guoqing Wang, and Chongyi Li. Iterative predictor-critic code decoding for real-world image dehazing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12700–12709, 2025. 3, 4, 6

[6] Jiang Hai, Zhu Xuan, Ren Yang, Yutong Hao, Fengzhu Zou, Fang Lin, and Songchen Han. R2rnet: Low-light image enhancement via real-low to real-normal network. *Journal of Visual Communication and Image Representation*, 90: 103712, 2023. 2, 5

[7] Yu Han, Yunze Cai, Yin Cao, and Xiaoming Xu. A new image fusion performance metric based on visual information fidelity. *Information fusion*, 14(2):127–135, 2013. 1

[8] Li Huaqiu, Haoqian Wang, et al. Interpretable unsupervised joint denoising and enhancement for real-world low-light scenarios. In *The Thirteenth International Conference on Learning Representations*, 2025. 2, 5

[9] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. Enlightengan: Deep light enhancement without paired supervision. *IEEE transactions on image processing*, 30:2340–2349, 2021. 2, 5

[10] Glenn Jocher and Jing Qiu. Ultralytics yolo11, 2024. 7

[11] Jinyuan Liu, Guanyao Wu, Zhu Liu, Long Ma, Risheng Liu, and Xin Fan. Where elegance meets precision: Towards a compact, automatic, and flexible framework for multi-modality image fusion and applications. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 1110–1118, 2024. 1, 4, 5, 6

[12] Jinyuan Liu, Bowei Zhang, Qingyun Mei, Xingyuan Li, Yang Zou, Zhiying Jiang, Long Ma, Risheng Liu, and Xin Fan. Dcevo: Discriminative cross-dimensional evolutionary learning for infrared and visible image fusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2226–2235, 2025. 1, 4, 5, 6

[13] Long Ma, Yuxin Feng, Yan Zhang, Jinyuan Liu, Weimin Wang, Guang-Yong Chen, Chengpei Xu, and Zhuo Su. Coa: Towards real image dehazing via compression-and-adaptation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 11197–11206, 2025. 3, 6

[14] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12): 4695–4708, 2012. 2, 4

[15] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 2, 4

[16] Guihong Qu, Dali Zhang, and Pingfan Yan. Information measure for performance of image fusion. *Electronics letters*, 38(7):313–315, 2002. 1

[17] Yun-Jiang Rao. In-fibre bragg grating sensors. *Measurement science and technology*, 8(4):355–375, 1997. 1

[18] J Wesley Roberts, Jan A Van Aardt, and Fethi Babikker Ahmed. Assessment of image fusion procedures using entropy, image quality, and multispectral classification. *Journal of Applied Remote Sensing*, 2(1):023522, 2008. 1

[19] Hao Shen, Henghui Ding, Yulun Zhang, Zhong-Qiu Zhao, and Xudong Jiang. Spatial frequency modulation network for efficient image dehazing. *IEEE Transactions on Image Processing*, 2025. 3, 6

[20] Yuda Song, Zhuqing He, Hui Qian, and Xin Du. Vision transformers for single image dehazing. *IEEE Transactions on Image Processing*, 32:1927–1941, 2023. 3, 6

[21] Yiming Sun, Xin Li, Pengfei Zhu, Qinghua Hu, Dongwei Ren, Huiying Xu, and Xinzhong Zhu. Task-gated multi-expert collaboration network for degraded multi-modal image fusion. In *International Conference on Machine Learning*, pages 57571–57586. PMLR, 2025. 1, 4, 5, 6

[22] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2555–2563, 2023. 2, 4

[23] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 1

[24] Zirui Wang, Jiayi Zhang, Tianwei Guan, Yuhan Zhou, Xingyuan Li, Minjing Dong, and Jinyuan Liu. Efficient rectified flow for image fusion. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 1, 4, 5, 6

[25] Guanyao Wu, Haoyu Liu, Hongming Fu, Yichuan Peng, Jinyuan Liu, Xin Fan, and Risheng Liu. Every sam drop counts: Embracing semantic priors for multi-modality image fusion and beyond. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17882–17891, 2025. 1, 4, 5, 6

[26] Costas S Xydeas and Vladimir Petrovic. Objective image fusion performance measure. *Electronics letters*, 36(4):308–309, 2000. 1

[27] Qingsen Yan, Yixu Feng, Cheng Zhang, Guansong Pang, Kangbiao Shi, Peng Wu, Wei Dong, Jinqiu Sun, and Yanning Zhang. Hvi: A new color space for low-light image enhancement. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5678–5687, 2025. 2, 5

[28] Xunpeng Yi, Han Xu, Hao Zhang, Linfeng Tang, and Jiayi Ma. Text-if: Leveraging semantic text guidance for degradation-aware and interactive image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27026–27035, 2024. 1, 4, 5, 6

[29] Xunpeng Yi, Yibing Zhang, Xinyu Xiang, Qinglong Yan, Han Xu, and Jiayi Ma. Lut-fuse: Towards extremely fast infrared and visible image fusion via distillation to learnable look-up tables. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14559–14568, 2025. 1, 4, 5, 6

[30] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14071–14081, 2023. 2, 4

[31] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *European conference on computer vision*, pages 1–21. Springer, 2022. 7

[32] Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Kai Zhang, Shuang Xu, Dongdong Chen, Radu Timofte, and Luc Van Gool. Equivariant multi-modality image fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 25912–25921, 2024. 1, 4, 5, 6