# Supplementary material for "Missing No More: Dictionary-Guided Cross-Modal Image Fusion under Missing Infrared"

Yafei Zhang[1], Meng Ma[1], Huafeng Li[1]*, Yu Liu[2],
[1]Faculty of Information Engineering and Automation, Kunming University of Science and Technology,
[2]Department of Biomedical Engineering, Hefei University of Technology

## 1. More Details of Our Method

We present the process of Joint Shared-dictionary Representation Learning (JSRL) in Algorithm 1. First, the infrared and visible images are fed into the HeadNet to generate their coefficient maps, with the shared dictionary being initialized to zeros. Subsequently, these initial representations are progressively refined through N IV-DLBs, enabling the final coefficients and shared dictionary to accurately reconstruct the input infrared and visible images.

**Algorithm 1** Joint Shared-dictionary Representation Learning

**Input:** Infrared image $\mathbf{I}_{ir}$, visible image $\mathbf{I}_{vis}$, scale factor $\sigma$, the number of IV-DLBs $N$
**Output:** Reconstructed infrared image $\mathbf{I}'_{ir}$, Reconstructed visible image $\mathbf{I}'_{vis}$, shared dictionary $\mathbf{D}$

$\mathbf{S}_{vis,(0)} = \text{HeadNet}(\mathbf{I}_{vis})$, $\mathbf{S}_{ir,(0)} = \text{HeadNet}(\mathbf{I}_{ir})$, $\mathbf{D}_{(0)} = 0$
**for** $n = 1$ to $N$ **do**
  $\{\mu_{1,(n)}, \mu_{2,(n)}, \mu_{3,(n)}, \beta_{1,(n)}, \beta_{2,(n)}, \beta_{3,(n)}\} = \text{HypNet}(\sigma)$
  $\mathbf{S}'_{vis,(n)} = \text{CSB-VIS}\big(\mathbf{I}_{vis}, \mathbf{D}_{(n-1)}, \mathbf{S}_{vis,(n-1)}, \mu_{1,(n)}\big)$
  $\mathbf{S}_{vis,(n)} = \text{CoeNet-VIS}\big(\mathbf{S}'_{vis,(n)}, \beta_{1,(n)}\big)$
  $\mathbf{S}'_{ir,(n)} = \text{CSB-IR}\big(\mathbf{I}_{ir}, \mathbf{D}_{(n-1)}, \mathbf{S}_{ir,(n-1)}, \mu_{2,(n)}\big)$
  $\mathbf{S}_{ir,(n)} = \text{CoeNet-IR}\big(\mathbf{S}'_{ir,(n)}, \beta_{2,(n)}\big)$
  $\mathbf{D}'_{(n)} = \text{DSB}\big(\mathbf{I}_{vis}, \mathbf{I}_{ir}, \mathbf{D}_{(n-1)}, \mathbf{S}_{vis,(n)}, \mathbf{S}_{ir,(n)}, \mu_{3,(n)}\big)$
  $\mathbf{D}_{(n)} = \text{DicNet}\big(\mathbf{D}'_{(n)}, \beta_{3,(n)}\big)$
**end for**
$\mathbf{I}'_{vis} = \mathbf{D}_{(N)} * \mathbf{S}_{vis,(N)}$, $\mathbf{I}'_{ir} = \mathbf{D}_{(N)} * \mathbf{S}_{ir,(N)}$

In addition, we summarize the training pipeline for the VIS-Guided IR Inference (VGII) and Adaptive Fusion via Representation Inference (AFRI) modules in Algorithm 2.

## 2. Hyperparameter Analysis

The proposed method involves two key hyperparameters: the kernel size of the dictionary and the number of IV-DLBs. In our implementation, we set the kernel size of the dictionary to $5 \times 5$ and use a single IV-DLB layer. In this section, we analyze the influence of the hyperparameters on model performance on the FLIR dataset.

**Algorithm 2** VIS-Guided IR Inference and Adaptive Fusion via Representation Inference

**Input:** visible image $\mathbf{I}_{vis}$, shared dictionary $\mathbf{D}$, pretrained LLM: Qwen-7B-Chat
**Output:** Fusion image $\mathbf{I}_f$

$\tilde{\mathbf{S}}_{vis} = \text{REN}(\mathbf{I}_{vis}, \mathbf{D})$
$\mathbf{S}^{(0)}_{p\_ir} = \text{RIN}(\tilde{\mathbf{S}}_{vis})$
$\mathbf{I}^{(0)}_{p\_ir} = \text{TailNet}(\mathbf{S}^{(0)}_{p\_ir}, \mathbf{D})$
$\mathbf{F}_{text} = \text{LLM}\big(\mathbf{I}_{vis}, \mathbf{I}^{(0)}_{p\_ir}\big)$
$\gamma = \Phi_\gamma(\mathbf{F}_{text}), \quad \beta = \Phi_\beta(\mathbf{F}_{text})$
$\mathbf{S}^{(1)}_{p\_ir} = \text{RIN}(\gamma \odot \tilde{\mathbf{S}}_{vis} + \beta)$
$\{\mathbf{W}_{vis}, \mathbf{W}_{p\_ir}\} = \Psi_{gate}(\text{Concat}(\tilde{\mathbf{S}}_{vis}, \mathbf{S}^{(1)}_{p\_ir}))$
$\mathbf{I}_f = \text{TailNet}(\mathbf{W}_{vis}, \tilde{\mathbf{S}}_{vis}, \mathbf{W}_{p\_ir}, \mathbf{S}^{(1)}_{p\_ir})$

Table 1. Influence of the dictionary kernel size on model perfomance. The best performance for each metric is marked with Red.

| Kernel size | AG ↑ | CE ↓ | EI ↑ | EN ↑ | Qcb ↑ | SF ↑ |
|---|---|---|---|---|---|---|
| $3 \times 3$ | 4.156 | 1.357 | 44.882 | 6.545 | 0.430 | 11.195 |
| $5 \times 5$ | 4.518 | 0.596 | 48.784 | 6.639 | 0.435 | 12.554 |
| $7 \times 7$ | 4.499 | 1.298 | 48.560 | 6.627 | 0.356 | 12.300 |

**Influence of the dictionary kernel size.** To evaluate the impact of the dictionary kernel size on model perfomance, we evaluate three different kernel configurations: $3 \times 3$, $5 \times 5$, and $7 \times 7$. As shown in Table 1, increasing the kernel size from $3 \times 3$ to $5 \times 5$ results in notable improvements across all metrics. However, further enlarging the kernel to $7 \times 7$ leads to varying degrees of performance degradation. $5 \times 5$ yields the best overall fusion performance.

**Influence of the number of IV-DLB.** To evaluate the effect of the number of IV-DLBs, we conduct experiments with $N = 1, 2, 3$. As shown in Table 2, increasing $N$ from

*Corresponding author: Huafeng Li (hfchina99@163.com)

1 to 2 leads to only marginal performance gains, while the FLOPs nearly doubles. Further increasing N to 3 results in a degradation of performance. Therefore, we introduce an IV-DLB layer in our experiments to achieve a favorable trade-off between model performance and computational complexity.

Table 2. Influence of the number of IV-DLBs on model perfomance. The best performance for each metric is marked with Red.

| N | AG ↑ | CE ↓ | EI ↑ | EN ↑ | Qcb ↑ | SF ↑ | GFLOPS |
|---|---|---|---|---|---|---|---|
| 1 | 4.518 | 0.596 | 48.784 | 6.639 | 0.435 | 12.554 | 63.023 |
| 2 | 4.529 | 0.565 | 49.096 | 6.598 | 0.432 | 12.671 | 120.795 |
| 3 | 4.485 | 0.679 | 48.641 | 6.705 | 0.432 | 12.265 | 178.567 |

**Influence of the textual prompt.** To evaluate the impact of textual prompt fed to the LLM on the model performance, we conduct experiments using three different types of prompt settings. As shown in the Table 3, the results demonstrate that variations in the prompts affect the fusion performance. The model utilizing the complete prompt achieves the best results. Both our prompt and a simplified version are illustrated in Figure 1.

Table 3. Influence of the textual prompt on model perfomance. The best performance for each metric is marked with Red.

| Prompt | AG ↑ | CE ↓ | EI ↑ | EN ↑ | Qcb ↑ | SF ↑ |
|---|---|---|---|---|---|---|
| No Prompt | 4.183 | 1.016 | 47.166 | 6.215 | 0.401 | 11.694 |
| Simplified Prompt | 4.431 | 0.610 | 48.687 | 6.613 | 0.419 | 12.186 |
| Our Prompt | 4.518 | 0.596 | 48.784 | 6.639 | 0.435 | 12.554 |

*This is a synthetic infrared image generated from the visible image. Please analyze how realistic this infrared image appears, and suggest what modifications are needed to make it more physically accurate. Focus on thermal distribution, contrast, and texture fidelity.*

(a)Our prompt

*This is a synthetic infrared image generated from a visible image. Please evaluate its realism and suggest improvements.*

(b)Simplified prompt

Figure 1. Two different types of text prompts. (a) is our prompt used in the experiment, while (b) is the simplified version.

## 3. Evalution Metrics

We use six widely recognized image fusion evaluation metrics to objectively assess the quality of the fusion results. Among them, the average gradient (AG) [3] measures the edge intensity of the image, indicating that the fused image can effectively restore the boundary of the target object. The contrast entropy (CE) [1] measures the distribution of contrast in an image. The edge intensity (EI) [8] measures the clarity and strength of edges in the image. The image entropy (EN) [6] measures the complexity of details in the image. A higher image entropy indicates that the image contains more information and has more variations. The quality consistency based on fusion (Qcb) [2] measures the quality consistency between the fused image and the reference image. The spatial frequency (SF) [4] measures the spatial frequency of the image to ensure a better visual effect and clearer details in the fused image.

## 4. Complexity Analysis

In this section, we conduct a complexity analysis of the proposed method and compare it with several infrared-visible fusion methods that rely on infrared image generation. Table 4 presents the complexity comparison results and inference time for each image among different methods. Compared with other pixel-level generation fusion methods, our method shows a significantly lower number of learnable parameters, moderate FLOPs and inference time, while maintaining superior performance. This advantage stems from our feature-level infrared modal inference and fusion framework, which eliminates the need for generating infrared images, thereby reducing computational complexity.

## 5. More experimental results

To further evaluate the performance of our method, we conduct additional comparative experiments under missing-IR modility. Specifically, we first employ pixel-level infrared generation methods to reconstruct the missing infrared images and then fused the generated infrared images with visible images. The selected infrared generation methods include PID [7], a GAN-based approach EGGAN [5], and , a diffusion-based approach. The fusion process still adopts the ten mainstream fusion methods used in the main paper. Our method does not generate infrared images. Instead, it directly infers infrared features from the input visible images and produces enhanced images through the adaptive fusion of infrared andvisible features.

As illustrated in Figures 2 and 3, the qualitative results show that our method achieves an optimal balance among detail fidelity, thermal-target enhancement, and overall visual naturalness. The proposed approach effectively preserves the thermal information while fully integrating the structural and textural features of the visible image. In contrast, methods based on infrared generation suffer from modality imbalance, and their fusion results exhibit issues such as blurring, ghosting, and excessively bright or dark regions. The quantitative results in Tables 5 and 6 fur-

Table 4. Complexity analysis of fusion models with missing infrared modality. All parameters listed in the table represent the learnable parameters of the models. The FLOPs and inference time for each model are calculated using input images of size $256 \times 256$. 'E' denotes the EGGAN method, while 'P' denotes the PID method.

| Model | Param (M) | FLOPs (G) | Time (s) | Qcb ↑ |
|---|---|---|---|---|
| E+U2Fusion | 48.064 | 710.385 | 0.103 | 0.368 |
| E+TarDAL | 46.892 | 96.768 | 0.052 | 0.304 |
| E+CDDFuse | 47.784 | 194.176 | 0.059 | 0.429 |
| E+LRRNet | 46.644 | 80.347 | 0.026 | 0.390 |
| E+IVFSWR | 60.098 | 936.753 | 0.176 | 0.422 |
| E+CoCoNet | 55.709 | 118.866 | 0.075 | 0.409 |
| E+EMMA | 48.111 | 85.935 | 0.038 | 0.427 |
| E+TIMFusion | 47.827 | 111.154 | 0.053 | 0.396 |
| E+DCEvo | 48.601 | 272.027 | 0.059 | 0.432 |
| E+SAGE | 46.731 | 212.942 | 0.062 | 0.403 |
| P+U2Fusion | 242.895 | 2,921.78 | 11.066 | 0.388 |
| P+TarDAL | 241.723 | 2,308.16 | 11.015 | 0.282 |
| P+CDDFuse | 242.614 | 2,405.57 | 11.022 | 0.403 |
| P+LRRNet | 241.475 | 2,291.74 | 10.989 | 0.408 |
| P+IVFSWR | 254.929 | 3,148.14 | 11.139 | 0.377 |
| P+CoCoNet | 250.541 | 2,330.26 | 11.038 | 0.433 |
| P+EMMA | 242.776 | 2,297.33 | 11.002 | 0.390 |
| P+TIMFusion | 242.658 | 2,322.54 | 11.026 | 0.404 |
| P+DCEvo | 243.431 | 2,483.42 | 11.032 | 0.412 |
| P+SAGE | 241.562 | 2,424.33 | 11.025 | 0.339 |
| **Ours** | **21.798** | **542.001** | **5.793** | **0.435** |

ther validate that our method delivers superior performance across all evaluation metrics. These resluts clearly confirm that, for infrared–visible image fusion under missing infrared modality, our approach achieves state-of-the-art performance.

## 6. Limitations and Future Work

Although the proposed method is effective, it still has certain limitations. The method primarily addresses the catastrophic effects of missing infrared images but cannot fully replace the role of infrared images in challenging imaging environments. For example, under extreme weather conditions such as heavy fog, visible light imaging devices often fail to capture key targets, making it difficult to reliably infer significant infrared information from the visible images. This limitation constrains the upper bound of the fusion framework. In future work, we will further investigate fusion paradigms under missing infrared conditions, with a focus on enhancing the visual representation of visible images in uncontrollable imaging environments and reducing the uncertainty introduced by variations in imaging conditions and infrared feature inference. Additionally, we plan to incorporate more robust priors and learnable constraints to improve the method's generalization and stability in complex scenarios.

## References


[1] D M Bulanon, T F Burks, and V Alchanatis. Image fusion of visible and thermal images for fruit detection. *Biosystems Engineering*, 103(1):12–22, 2009. 2

[2] Yin Chen and Rick S Blum. A new automated quality assessment algorithm for image fusion. *Image and Vision Computing*, 27(10):1421–1432, 2009. 2

[3] Guangmang Cui, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Detail preserved fusion of visible and infrared images using regional saliency extraction and multi-scale image decomposition. *Optics Communications*, 341:199–209, 2015. 2

[4] Ahmet M Eskicioglu and Paul S Fisher. Image quality measures and their performance. *IEEE Transactions on Communications*, 43(12):2959–2965, 1995. 2

[5] Dong-Guw Lee, Myung-Hwan Jeon, Younggun Cho, and Ayoung Kim. Edge-guided multi-domain rgb-to-tir image translation for training vision tasks with challenging labels. *arXiv preprint arXiv:2301.12689*, 2023. 2

[6] Jiayi Ma, Yong Ma, and Chang Li. Infrared and visible image fusion methods and applications: A survey. *Information fusion*, 45:153–178, 2019. 2

[7] Fangyuan Mao, Jilin Mei, Shun Lu, Fuyang Liu, Liang Chen, Fangzhou Zhao, and Yu Hu. Pid: physics-informed diffusion model for infrared image generation. *Pattern Recognition*, 169:111816, 2026. 2

[8] B Rajalingam and R Priya. Hybrid multimodality medical image fusion technique for feature enhancement in medical diagnosis. *International Journal of Engineering Science Invention*, pages 52–60. 2


Figure 2. Visual comparison of different fusion methods on the FLIR, MSRS, and KAIST datasets, where the infrared images are generated from visible images using the EGGAN method.

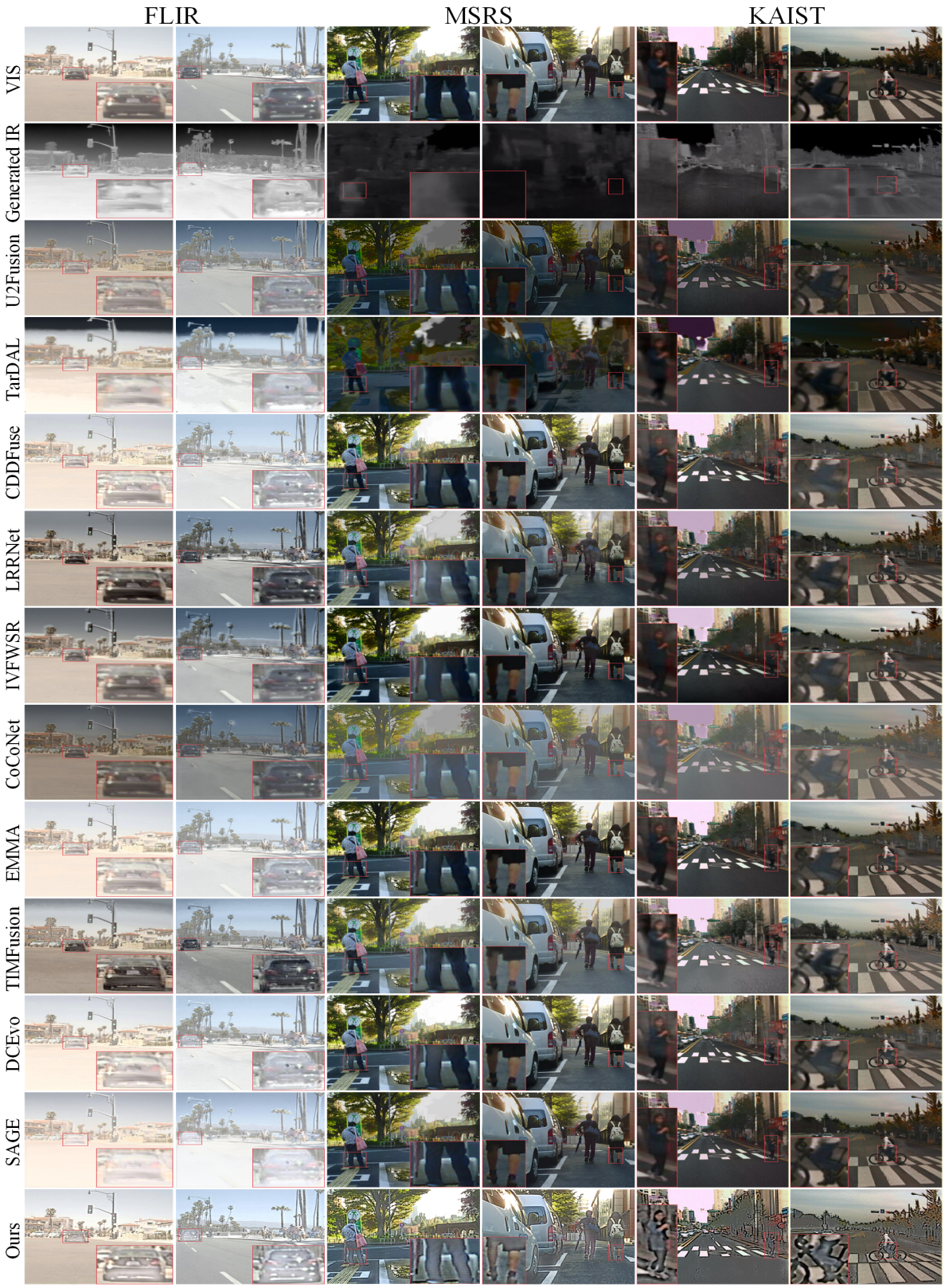


Figure 3. Visual comparison of different fusion methods on the FLIR, MSRS, and KAIST datasets, where the infrared images are generated from visible images using the PID method.

Table 5. Quantitative comparison of different fusion methods on the FLIR, MSRS, and KAIST datasets, where the infrared images are generated from visible images using the EGGAN method. The best and second-best performances are highlighted with Red and Blue backgrounds, respectively.

| **Datasets** | **MSRS** | | | | | | **FLIR** | | | | | | **KAIST** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Methods** | AG ↑ | CE ↓ | EI ↑ | EN ↑ | Qcb ↑ | SF ↑ | AG ↑ | CE ↓ | EI ↑ | EN ↑ | Qcb ↑ | SF ↑ | AG ↑ | CE ↓ | EI ↑ | EN ↑ | Qcb ↑ | SF ↑ |
| U2Fusion | 3.289 | 3.309 | 34.980 | 5.447 | 0.327 | 9.310 | 2.320 | 1.453 | 24.107 | 5.045 | 0.368 | 6.517 | 2.425 | 2.734 | 25.699 | 5.489 | 0.388 | 6.259 |
| TarDAL | 3.205 | 3.458 | 34.550 | 6.640 | 0.363 | 8.602 | 2.341 | 2.756 | 25.123 | 6.536 | 0.304 | 6.898 | 1.916 | 1.728 | 20.399 | 6.197 | 0.382 | 5.807 |
| CDDFuse | 3.644 | 2.839 | 38.606 | 6.513 | 0.318 | 11.581 | 3.004 | 0.515 | 31.810 | 6.406 | 0.429 | 8.746 | 2.944 | 4.315 | 31.419 | 6.225 | 0.374 | 8.835 |
| LRRNet | 3.979 | 4.924 | 42.225 | 6.617 | 0.340 | 12.258 | 3.905 | 1.993 | 41.323 | 6.866 | 0.390 | 11.946 | 3.086 | 2.443 | 33.250 | 6.642 | 0.418 | 8.862 |
| IVFWSR | 3.229 | 4.218 | 34.902 | 6.461 | 0.378 | 8.608 | 3.000 | 3.704 | 32.530 | 6.349 | 0.422 | 7.818 | 2.635 | 2.631 | 28.661 | 6.462 | 0.414 | 6.959 |
| CoCoNet | 3.930 | 2.664 | 41.488 | 6.312 | 0.385 | 11.626 | 3.948 | 3.267 | 41.592 | 6.363 | 0.409 | 11.136 | 2.716 | 3.055 | 28.735 | 6.006 | 0.390 | 7.524 |
| EMMA | 4.466 | 3.236 | 47.469 | 6.608 | 0.366 | 13.207 | 3.028 | 0.691 | 32.297 | 6.467 | 0.427 | 8.689 | 3.178 | 2.969 | 34.234 | 6.405 | 0.414 | 8.941 |
| TIMFusion | 3.887 | 5.072 | 41.836 | 6.839 | 0.356 | 11.165 | 3.892 | 2.938 | 41.172 | 6.571 | 0.396 | 11.900 | 3.479 | 2.473 | 37.688 | 6.894 | 0.446 | 9.279 |
| DCEvo | 4.170 | 4.718 | 44.500 | 6.814 | 0.308 | 13.240 | 3.246 | 0.591 | 34.602 | 6.532 | 0.432 | 9.768 | 3.451 | 4.359 | 37.032 | 6.628 | 0.375 | 9.887 |
| SAGE | 4.150 | 3.057 | 44.114 | 6.777 | 0.327 | 12.560 | 2.724 | 1.160 | 28.303 | 5.966 | 0.403 | 8.138 | 3.376 | 2.986 | 35.924 | 6.608 | 0.430 | 9.534 |
| Ours | 5.037 | 3.142 | 53.168 | 7.188 | 0.454 | 14.898 | 4.518 | 0.596 | 48.784 | 6.639 | 0.435 | 12.554 | 4.414 | 2.227 | 37.115 | 7.073 | 0.405 | 11.031 |

Table 6. Quantitative comparison of different fusion methods on the FLIR, MSRS, and KAIST datasets, where the infrared images are generated from visible images using the PID method. The best and second-best performances are highlighted with Red and Blue backgrounds, respectively.

| **Datasets** | **MSRS** | | | | | | **FLIR** | | | | | | **KAIST** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Methods** | AG ↑ | CE ↓ | EI ↑ | EN ↑ | Qcb ↑ | SF ↑ | AG ↑ | CE ↓ | EI ↑ | EN ↑ | Qcb ↑ | SF ↑ | AG ↑ | CE ↓ | EI ↑ | EN ↑ | Qcb ↑ | SF ↑ |
| U2Fusion | 3.074 | 1.668 | 32.810 | 6.047 | 0.428 | 9.168 | 3.277 | 1.353 | 33.900 | 6.576 | 0.388 | 8.329 | 2.533 | 0.828 | 27.040 | 6.180 | 0.459 | 6.911 |
| TarDAL | 1.938 | 2.342 | 20.524 | 6.030 | 0.378 | 5.654 | 2.516 | 0.643 | 26.795 | 7.055 | 0.282 | 6.860 | 1.937 | 1.268 | 20.667 | 6.302 | 0.427 | 5.766 |
| CDDFuse | 4.821 | 1.651 | 51.500 | 7.191 | 0.466 | 14.142 | 3.837 | 0.726 | 39.804 | 6.517 | 0.403 | 10.254 | 3.349 | 1.361 | 35.909 | 7.064 | 0.488 | 9.859 |
| LRRNet | 3.815 | 2.837 | 40.509 | 7.049 | 0.445 | 11.150 | 3.970 | 1.798 | 42.038 | 6.955 | 0.408 | 11.154 | 3.125 | 1.432 | 33.685 | 7.017 | 0.439 | 8.740 |
| IVFWSR | 3.994 | 1.342 | 43.035 | 7.046 | 0.470 | 11.054 | 2.976 | 1.266 | 31.999 | 7.180 | 0.377 | 7.546 | 2.878 | 0.857 | 31.010 | 6.962 | 0.442 | 8.313 |
| CoCoNet | 3.357 | 0.801 | 35.271 | 6.477 | 0.418 | 10.386 | 2.456 | 1.848 | 25.333 | 6.417 | 0.433 | 6.589 | 2.200 | 1.073 | 23.195 | 6.240 | 0.456 | 6.444 |
| EMMA | 4.904 | 1.353 | 52.540 | 7.187 | 0.440 | 14.253 | 3.150 | 0.745 | 33.517 | 6.280 | 0.390 | 8.830 | 3.334 | 0.813 | 36.033 | 6.979 | 0.495 | 9.655 |
| TIMFusion | 3.952 | 3.928 | 42.063 | 6.936 | 0.374 | 11.708 | 3.967 | 1.541 | 42.121 | 6.982 | 0.404 | 11.572 | 3.430 | 1.877 | 36.806 | 7.004 | 0.434 | 9.509 |
| DCEvo | 4.722 | 1.983 | 50.492 | 7.231 | 0.476 | 13.974 | 3.468 | 0.812 | 36.719 | 6.531 | 0.412 | 9.516 | 3.439 | 1.209 | 36.676 | 6.998 | 0.497 | 10.381 |
| SAGE | 4.447 | 1.867 | 47.126 | 7.005 | 0.436 | 13.246 | 2.690 | 0.697 | 27.959 | 6.379 | 0.339 | 7.818 | 3.234 | 1.035 | 34.370 | 6.805 | 0.416 | 9.574 |
| Ours | 5.037 | 3.142 | 53.168 | 7.188 | 0.454 | 14.898 | 4.518 | 0.596 | 48.784 | 6.639 | 0.435 | 12.554 | 4.414 | 2.227 | 37.115 | 7.073 | 0.405 | 11.031 |