Revisiting Network Inertia: Dynamic Inertia Inhibition Coupled Multidimensional Periodicity for Infrared and Visible Image Fusion

Yufeng Chen[1], Yuan Sun[3], Hao Pan[1], Xujian Zhao[2], Jian Dai[4], Zhenwen Ren[2*], Xingfeng Li[2*]

[1]School of information and Control Engineering, Southwest University of Science and Technology, Mianyang, China
[2]School of Computer Science and Technology, Southwest University of Science and Technology, Mianyang, China
[3]National Key Laboratory of Fundamental Algorithms and Models for Engineering Numerical Simulation, Sichuan University, Chengdu, China
[4]Southwest Automation Research Institute, Mianyang, China

Abstract

Infrared and visible image fusion (IVIF) technology has become a frontier of great interest due to the ability to integrate information from multiple sources. However, the progressive slowdown of weight updates in deep networks (i.e., "network laziness" phenomenon), makes existing methods far from realizing the full characterization potential. To this end, we propose a lightweight fusion method for IVIF, Anti-Inert Dynamic Fusion (AIDFusion), to fully utilize the potential of the network at all levels. Specifically, by progressively regulating the collaborative Learning process of multi-level prediction in the network, Dynamic Inertia Inhibition Learning Strategy (DIILS) is proposed to adaptively and efficiently inhibit inertia accumulation. Subsequently, to deeply explore the representation potential while breaking through the performance threshold, lightweight Multi-dimensional modulation fusion module (MMFM) is specifically proposed to capture comprehensive multi-view and multi-scale features efficiently. Finally, considering the semantic bias between the prediction maps of DIILS and the fusion feature of MMFM, Fourier Analysis Convolution (FAConv) is designed in feature recovery as a bridge between prediction and fusion to accomplish the implicit periodic modeling. Based on the above study, extensive experiments on three public IVIF datasets demonstrate the dual advantages of AIDFusion in terms of fusion performance and computational overhead compared to state-of-the-art baseline methods.

Code — https://github.com/YufengChen1113/AIDFusion

Introduction

With the development of the information age, the integration of multi-source data has become a research priority (Xu et al. 2020; Li et al. 2023c; Luo et al. 2025a; Yuan et al. 2025). Multimodal image fusion techniques are able to capture information from different sources for subsequent tasks such as target detection, image segmentation, and intelligent diagnosis (Xu et al. 2024; Luo et al. 2025b; Li et al. 2025a; Wang et al. 2025). In particular, the infrared and visible image fusion (IVIF) method pays effective attention to the salient regions of infrared images (IR) and the texture details of visible images (VI). Therefore, it is widely used in the fields of intelligent driving, military probing, and so on.

Existing methods can be simply categorized into traditional methods and deep neural network (DNN)-based methods. Traditional methods (Ellmauthaler, Pagliari, and Da Silva 2012; Liu, Liu, and Wang 2015) require manually designed rules, resulting in a lack of objectivity along with poor performance. DNNs have become the primary choice of researchers due to powerful characterization capabilities (Luo et al. 2021; Zhao et al. 2024a; Wan et al. 2024).

The early work, DeepFuse (Ram Prabhakar, Sai Srikar, and Venkatesh Babu 2017), uses CNN to extract features. However, the network's depth limits the improvement of performance. RFN-Nest, DenseFuse, and PMGI (Li, Wu, and Kittler 2021; Li and Wu 2018; Zhang et al. 2020) respectively utilize ResNet (He et al. 2016), DenseNet, and dual-path network to alleviate this problem to a certain extent. However, the emergence of ViT's long-range modeling capabilities has broken the dominance of CNN. SwinFusion (Ma et al. 2022) utilizes the mature Swin Transformer to extract global spatial information. This completes the global interaction of complementary information. Since then, methods such as CDDFuse, PSFusion, and CrossFuse (Zhao et al. 2023; Tang et al. 2023a; Li and Wu 2024) have utilized ViT to design brand-new networks, significantly enhancing the network's fitting capability. Methods such as TarDAL, SegMiF, and MRFS (Liu et al. 2022, 2023; Zhang et al. 2024) integrate fusion task with downstream tasks to obtain more semantic information. What's more, Text-IF, GIFNet, SAGE, etc. (Yi et al. 2024; Zhao et al. 2024c; Cheng et al. 2025; Wu et al. 2025) are combined with large model technology by external knowledge guidance, to achieve higher fusion performance.

However, more complex network models inevitably lead to the "network laziness" phenomenon (Liu, Lin, and Jiang 2024; Geirhos et al. 2020). In Figure 1, this phenomenon also exists in the IVIF domain. As the network's training enters middle stage, the magnitude of most of the weight updates decreases, which greatly affects the effectiveness of the training. This phenomenon is even more pronounced in complex networks, so lightweight networks may be a better way forward (Howard et al. 2019; Zhang et al. 2025b,a; Li et al. 2024; Lou et al. 2025). In fact, apart from the early sim-

*Corresponding authors.

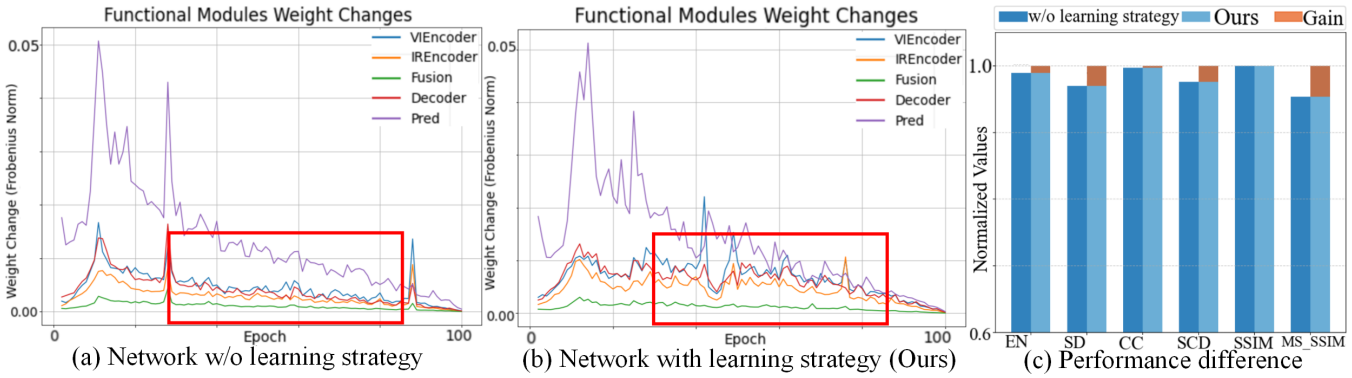

Figure 1: Illustration of "network laziness" and the benefit of our learning strategy on MSRS. (a) vs. (b): We track per-epoch Frobenius-norm weight changes in each functional block (encoder, fusion, decoder, prediction). The baseline (a) exhibits a clear mid-training slowdown, with updates diminishing to near-zero. In contrast, our strategy (b) preserves substantial weight adaptation throughout training, avoiding parameter stagnation. (c): Corresponding quality metrics on MSRS highlight that sustained weight dynamics yield higher performance.

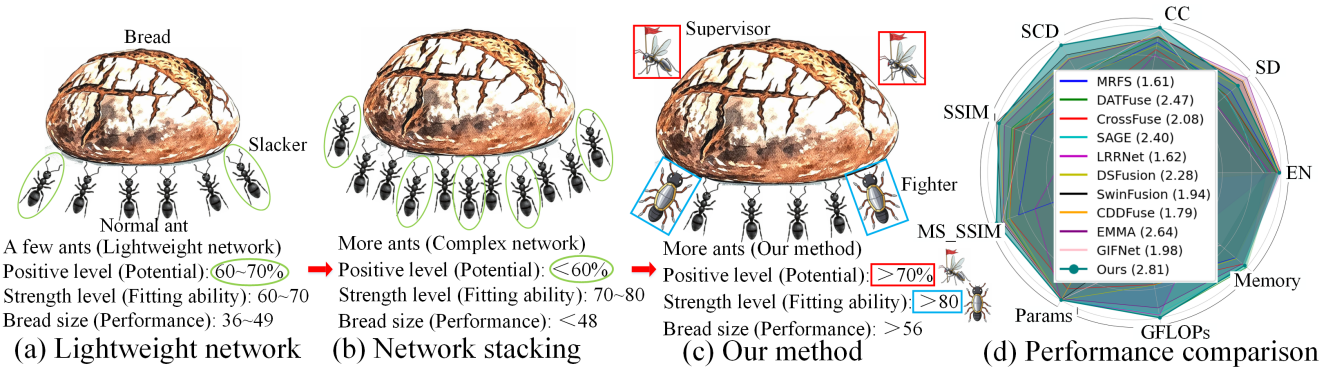

Figure 2: Design motivation and comparative evaluation. (a) Existing case 1: lightweight backbone with minimal parameters yield weak representational capacity and inferior fusion results. (b) Existing case 2: Network stacking with cascading models enhances nonlinearity but introduces heavy computational cost and aggravates mid-training stagnation. (c) Our method: We fully leverage a compact backbone via two bespoke modules (a supervisory learning controller and specialized fusion agents) thereby preserving efficiency while boosting expressiveness. (d) On the MSRS dataset, our approach attains the highest aggregate score over nine standard metrics, demonstrating its superior trade-off between accuracy and complexity.

ple networks, there are not no specially designed lightweight methods. Methods such as SeAFusion, DATFuse, and LRRNet (Tang, Yuan, and Ma 2022; Tang et al. 2023b; Li et al. 2023a; Lu et al. 2024) have long recognized the importance of lightweight. However, even though these methods have specialized designs to enhance fitting ability, the chasm of inherent nonlinear disadvantage is difficult to cross.

Rather than stacking ever-heavier models, this work innovatively explores a new perspective of network "utilization rate", fully exploiting the potential of lightweight networks through novel learning strategy and tightly coupled modules. We first design the Dynamic Inertia Inhibition Learning Strategy (DIILS), which employs Progressive Hierarchical Optimization (PHO) and Adaptive Collaborative Fusion Loss ($\mathcal{L}_{ACF}$) to sustain gradient activity and avert mid-training stagnation. Then, this stable convergence progressively promotes our Multi-dimensional Modulation Fusion Module (MMFM), using Multi-perspective Sparse Attention (MPSA) and a Multi-scale Recurrent Transformer (MSRT) to selectively refine and merge cross-modal features at minimal cost. Finally, Fourier Analysis Convolution (FAConv) embeds periodic basis functions into the decoder, sharpening fine-grained reconstruction. By interweaving these components, our method delivers superior fusion quality on MSRS without prohibitive complexity as shown in Figure 2. Our work contributions are as follows:

- We investigate the "network laziness" phenomenon in IVIF and introduce the Dynamic Inertia Inhibition Learning Strategy, which leverages Progressive Hierarchical Optimization and Adaptive Collaborative Fusion Loss to prevent mid-training stagnation.
- We design a lightweight multi-dimensional modulation fusion module, employing Multi-perspective Sparse At-

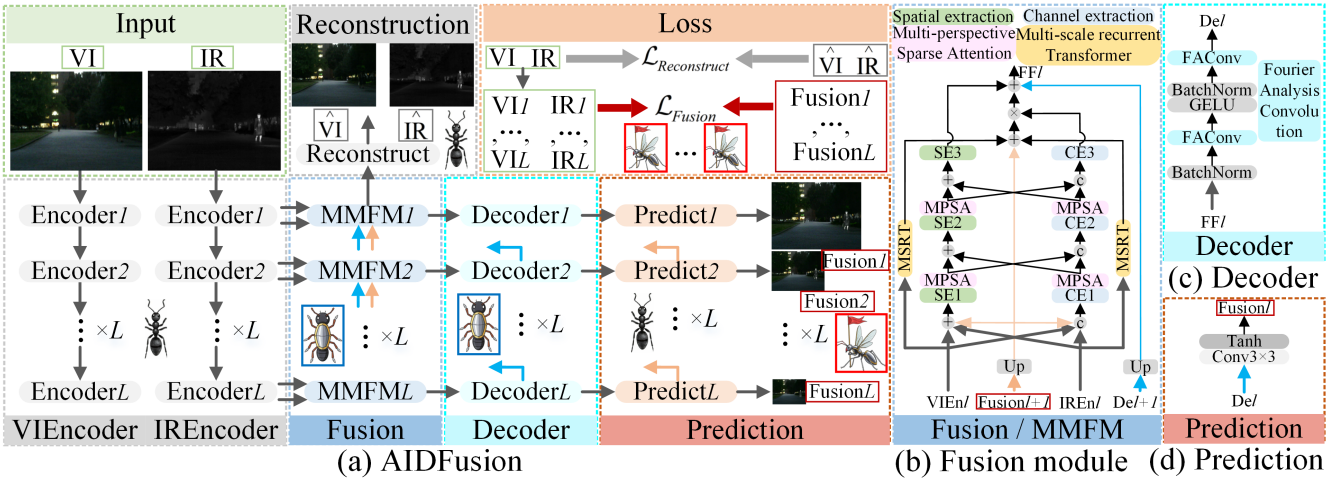

Figure 3: Overall framework diagram of this method. (a) Overall flowchart of AIDFusion. (b) Details of fusion module. (c) Structure of decoder. (d) Structure of prediction module.

tention and Multi-scale Recurrent Transformer to preserve rich cross-modal features with minimal overhead.

- A periodic-modeling unit Fourier Analysis Convolution is proposed that embeds Fourier series priors between the network's intermediate and output layers, accelerating feature recovery and signal prediction.
- We integrate these innovations into AIDFusion, the first utilization-driven IVIF framework, which delivers a superior balance of fusion accuracy and computational efficiency on three public IVIF datasets.

Method

Overall framework

In this paper, a novel IVIF method is designed and the overall framework is shown in Figure 3. (1) For L-layer networks, after inputting VI and IR to the respective encoders (using ResNet), the features of each layer of the network are obtained $\{VIEn1, ..., VIEnL\}$ and $\{IREn1, ..., IREnL\}$. (2) Input the $VIEnl$ and $IREnl$ of layer l, the output $Del + 1$ of Decoder and Prediction output $Fusionl + 1$ of layer $l + 1$ (when l=L, only $VIEnL$ and $IREnL$) into the Fusion module to get the fused features FFl. (3) Input FFl into Decoder for feature recovery to get Del. (4) Input Del into Prediction to get final output $Fusionl$. (5) $Fusionl$ (l = 1, ... , L) are performed with VI and IR for the computation of the fusion Loss ($\mathcal{L}_{Fusion}$). And to prevent training collapses reconstructed VI, IR ($\hat{VI}$, $\hat{IR}$) are performed with VI and IR for the computation of the reconstructed Loss ($\mathcal{L}_{Reconstruct}$).

Dynamic Inertia Inhibition Learning Strategy

This part explains in detail the use of DIILS to mitigate the "network laziness" phenomenon. As the network gradually learns, the weights are more likely to be updated only close to the prediction layer, which results in the network's fitting ability being far from being utilized. Therefore, it becomes a natural choice to increase the number of prediction layers to strengthen the supervision and management. However, a simple increase in the number will inevitably bring about a training collapse caused by too many supervisory signals. Three core issues arise: which supervisory signals are selected; when to introduce supervisory signals; and how to design effective supervisory rules.

PHO: We use a hierarchical supervision paradigm with gradual injection according to epoch. (1) Layering: We use Prediction after each Decoder layer and treat each layer output as a supervised path instead of just one supervised path. This way, most parts of the network are covered during the backpropagation of training. (2) Progressive: Based on the occurrence time of the phenomenon, this method distinguishes the training epochs. When all epochs are set to T, the first half of the learning period is normal. In the second half, for each additional T/10 epochs, one more supervisory signal is added until all supervisory signals are considered. Therefore, the supervisory signals are:

$$\begin{cases} Fusionl, 0 < \text{ epoch } \leq \mathrm{T}/2 \\ Fusion1, Fusion2, \mathrm{T}/2 < \text{ epoch } \leq \mathrm{T}/2 + \mathrm{T}/10 \\ Fusion1, Fusion2, ..., FusionL, \text{Others} \end{cases}. \quad (1)$$

$\mathcal{L}_{\mathbf{ACF}}$: The first two core issues of this part are addressed in the PHO. However, the supervisory signals at different layers have different impacts. Therefore, what kind of supervisory rules can better guide the network to develop in a positive direction is also an important issue. The losses at different layers adaptively adjust to each other to achieve self-supervision. (1) IR and VI are downsampled multiple times using MaxPooling (MP) to get multiple sizes of images IRl and VIl to match the size of $Fusionl$. (2) The IRl, VIl and $Fusionl$ are subjected to the loss of fusion ($\mathcal{L}_{ACFl}$) for each layer calculated as follows:

$$\begin{aligned} \mathcal{L}_{ACFl} = \alpha\mathcal{L}_{corr}(IRl, VIl, Fusionl) \\ +\beta\mathcal{L}_{grad}(IRl, VIl, Fusionl) \\ +\gamma\mathcal{L}_{int}(IRl, VIl, Fusionl), \end{aligned} \quad (2)$$

where $\mathcal{L}_{corr}$, $\mathcal{L}_{grad}$ and $\mathcal{L}_{int}$ are correlation loss, gradient loss and intensity loss. α, β and γ are hyperparameters. The

formula for calculating the $\mathcal{L}_{corr}$ is as follows:

$$\mathcal{L}_{corr} = \frac{1}{corr\left(IRl, Fusionl\right) + corr\left(VIl, Fusionl\right) + \varepsilon},$$
$$s.t. \quad corr\left(X, Y\right) = \frac{\sum\left(X - \overline{X}\right)\left(Y - \overline{Y}\right)}{\sqrt{\sum\left(X - \overline{X}\right)^2 \sum\left(Y - \overline{Y}\right)^2}}, \varepsilon = 10^{-6}, \tag{3}$$

where $corr(\cdot)$ denotes the Pearson's correlation coefficient, $\overline{X}$ and $\overline{Y}$ are the mean values of X and Y, and ε is a minimal constant. The $\mathcal{L}_{grad}$ is calculated as follows:

$$\mathcal{L}_{grad} = \frac{1}{HlWl} \left\| \nabla Fusionl| - \max\left(|\nabla IRl|, |\nabla VIl|\right) \right\|_1, \tag{4}$$

where Hl and Wl denote the height and width of the network output at layer l, ∇ represents the Sobel gradient operation, $|\cdot|$ represents the absolute value operation, $max(\cdot)$ represents the selection of the maximum element value, and $||\cdot||_1$ represents the calculation of the l_1-norm. The formula for calculating $\mathcal{L}_{int}$ is as follows:

$$\begin{aligned}
\mathcal{L}_{int} =& \frac{1}{HlWl}\omega_{soft}(map_{ir} MSE\left(IRl, Fusionl\right)) \\
&+ \frac{1}{HlWl}\omega_{soft}(map_{vi} MSE\left(VIl, Fusionl\right)), \\
s.t. \quad & \omega_{soft} = Sigmoid\left(-\delta \cdot \left(IRl - \overline{IRl}\right)\right), \\
& map_{ir} = \begin{cases} 1, \text{ if } std_{ir} > std_{vi} \\ 0, \text{ otherwise} \end{cases}, map_{vi} = 1 - map_{ir}, \\
& MSE\left(X, Y\right) = \frac{1}{N}\sum\left(X - Y\right)^2,
\end{aligned} \tag{5}$$

where ω_{soft} is the soft-assigned weights, $Sigmoid(\cdot)$ is the Sigmoid activation function, and δ is the control coefficient (set to 1). map_{ir} and map_{vi} are the mask weights of IRl and VIl, and std is the variance. $MSE(\cdot)$ is the computed mean square error. (3) Finally, the weights are assigned by adaptively adjusting the different importance of each layer of loss. The final fusion loss ($\mathcal{L}_{ACF}$) is as follows:

$$\mathcal{L}_{ACF} = \sum_{l=1}^{L} \omega_l \mathcal{L}_{ACFl}, \ \ s.t. \ \omega_l = \begin{cases} 1 + \frac{\mathcal{L}_{ACFl}}{\sum_{l=1}^{L}\mathcal{L}_{ACFl}}, l = 1 \\ \frac{\mathcal{L}_{ACFl}}{\sum_{l=1}^{L}\mathcal{L}_{ACFl}}, l \neq 1 \end{cases}, \tag{6}$$

where ω_l is the weight value of the layer l loss. Therefore, combining the $\mathcal{L}_{Reconstruct}$, the overall loss function ($\mathcal{L}_{Total}$) of this method is as follows:

$$\mathcal{L}_{Total} = \lambda_1 \mathcal{L}_{Reconstruct} + \lambda_2 \mathcal{L}_{ACF}, \tag{7}$$

here λ_1 and λ_2 are the control factors.

Multi-dimensional Modulation Fusion Module

The above describes how DIILS stimulates the network's potential. However, the simple design of traditional lightweight methods allows for inherent limitations in characterization capabilities. It is difficult to break through the natural performance bottleneck. In view of this, it is crucial to efficiently capture multi-dimensional high-level features under cost constraints. In the following, it is explained how MMFM accomplishes feature interaction, enhancement, and fusion through the specially designed MPSA, MSRT.

As shown in Figure 3(b), MMFM is divided into the following 3 parts: (1) Cross-modal interaction: summing

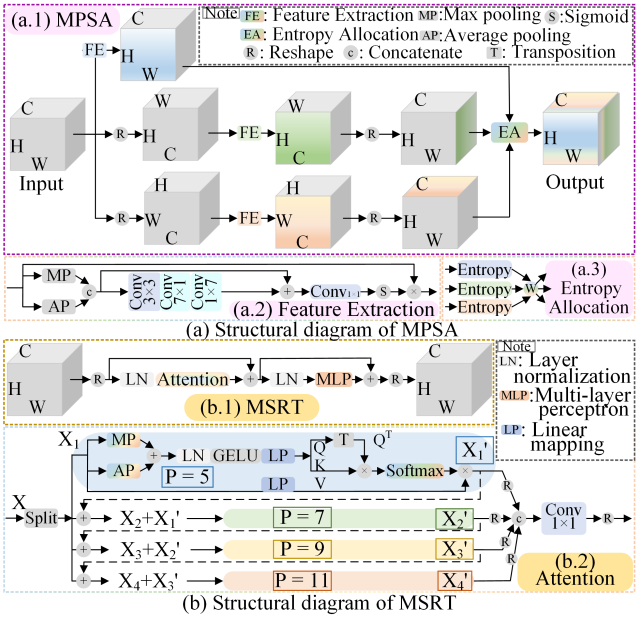

Figure 4: Framework diagram of MMFM. (a) Structure diagram of Multi-perspective Sparse Attention. (b) Structure diagram of Multi-scale recurrent Transformer.

and channel splicing operations are performed on $VIEnl$, $IREnl$, and the up-sampled $Fusionl + 1$. The two outputs are subsequently fed into spatial extraction (SE), channel extraction (CE) and the MPSA for multi-level and multi-angle feature extraction many times, which makes the features of VI and IR fully interactive. (2) Self-enhancement: $VIEnl$, $IREnl$ complete further feature enhancement through the designed MSRT for final output. (3) Full fusion: the cross-modal interaction with the self-enhanced results are finally fused and supplemented with $Del + 1$ and $Fusionl + 1$ to obtain the fusion result FFl.

MPSA: Previous approaches have designed many excellent attention modules (Woo et al. 2018; Luo et al. 2024; Yang, Huang, and Peng 2024), most of which focus from the spatial (H, W) or channel dimensions (C). However, previous fusion methods ignored the interactions between H (or W) and C, which can capture critical information from more perspectives. Based on this, MPSA focuses on the feature of multiple perspectives, as shown in Figure 4(a). (1) After the input is Reshaped, Feature Extraction (FE) is performed from the two interacting perspectives of HW, HC, and WC. Then the information entropy is calculated by Entropy Assignment (EA) to obtain the corresponding sparse weights. The final sum is obtained as the key feature enhancement map from multiple viewpoints. The computational process is as follows:

$$\begin{aligned}
Out = EA\left(FE\left(I\right) + R_{(H,W,C)}\left(FE\left(R_{(H,C,W)}\left(I\right)\right)\right.\right. \\
\left.\left. + R_{(H,W,C)}\left(FE\left(R_{(W,C,H)}\left(I\right)\right)\right)\right),
\end{aligned} \tag{8}$$

here $R(\cdot)$ means to Reshape the input map. (2) FE: The inputs are concatenated along the third dimension (e.g., C of $[H, W, C]$) after Max Pooling (MP) and Average Pooling (AP). This will greatly reduce the computational overhead

while allowing subsequent operations to focus more on the first two dimensions. Then, extraction is performed by 3×3 convolution, strip convolution with convolution kernel of 7, and 1×1 convolution. Finally, the output is obtained by multiplying with the input after $Sigmoid$. (3) Here we integrate the output of multiple viewpoint enhancements by calculating the information entropy. Unlike previous approaches, MPSA gives less weight to viewpoint maps with higher entropy, and retains more critical feature information rather than redundant information. The computational process is as follows:

$$Y = \sum \omega_i X_i, \quad s.t. \quad \omega_i = 1 - \sum_{j \neq i} Entropy(X_j), \tag{9}$$

here $Entropy(\cdot)$ denotes the information entropy calculation. X_i, ω_i are the inputs and weights of the ith perspective.

MSRT: In contrast to MPSA, MSRT focuses on the self-attention of the feature, which targets extracting important features at more scales. As shown in Figure 4(b), MSRT has the structure of Transformer and emphasizes Attention part. (1) This part inputs X (shape: $[HW, C]$), and then Splits into 4 parts along the C to $[X_1, ..., X_4]$ (Chen et al. 2023). (2) MP and AP are performed on the HW dimension of X_1 to limit the scale size (P), while greatly reducing the subsequent computational overhead (HW becomes P^2). After completing the mapping Q, K, and V are obtained, and unlike the general ViT, our method transposes (T) Q. Therefore the obtained attention weight shape is $[C/4, C/4]$ (the original ViT is $[HW, HW]$) again reducing the computational overhead. Next, the weight is multiplied with V to obtain the output X_1'. (3) Corresponding again to the lightweight design of this paper, the MSRT performs feature multiplexing, and the corresponding operation in (2) is performed after adding X_1' to X_2. And so on, all features (with different P) are fully captured due to the loop structure. (4) The extracted multi-scale features are integrated by 1×1 convolution and output. The computational flow of Attention is as follows:

$$\begin{aligned}
&X_{s=[1,2,3,4]} = Split(X), \quad X_i = \begin{cases} X_s, & i = 1 \\ X_s + X'_{i-1}, & i > 1 \end{cases}, \\
&Q_i, K_i = LP\left(GELU\left(LN\left(MP(X_i) + AP(X_i)\right)\right)\right), \\
&V_i = LP(X_i), \quad X'_i = V_i(Softmax(Q_i^T K_i)), \\
&Y = Conv_{1\times1}(concat(X'_{i=[1,2,3,4]})),
\end{aligned} \tag{10}$$

where X and Y are the input-output graphs, X_s is the feature graph after $Split$, X_i is the feature graph for which a certain scale of attention computation is performed, and X_i' is the feature graph after a certain scale of attention computation.

Fourier Analysis Convolution

Through the above study, learning strategy and fusion module help the network to effectively alleviate "network laziness" phenomenon and at the same time, carry out a comprehensive reshuffling of features. However, there are semantic differences between fusion layer and prediction layer, which usually require a certain period of learning to build up the potential patterns. Inspired by recent works (Liu et al. 2025; Dong et al. 2024; Li et al. 2023b, 2025b), we propose a novel alternative product to ordinary convolution for periodic implicit modeling. As shown in Figure 3(c), FAConv is used in the Decoder with the specific structure in Figure 5.

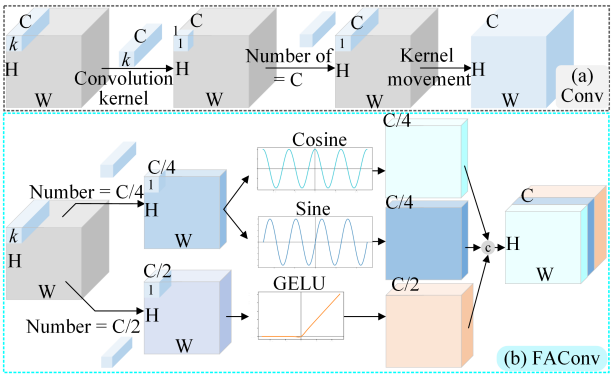

Figure 5: Comparison of normal convolution and FAConv.

Ordinary convolution uses a convolutional kernel to compute the corresponding region of the input image to get the values of the output location elements, and then movement operation makes the whole feature map be updated, as shown in Figure 5(a). However, due to the lack of prior knowledge, the convolutional kernel does not have a clear learning direction in the learning process, and it is more of a "trial and error process". Therefore, as shown in Figure 5(b): (1) Our FAConv convolves the input feature maps twice independently to obtain two intermediate maps with channels $C/4$ and $C/2$, respectively. (2) On the one hand, $GELU$ is performed for the graph of $C/2$ for rectification, and on the other hand, periodic modeling is performed for the graph of $C/4$, i.e., using $Cosine$ and $Sine$ for the search of periodic laws. Different from ordinary convolution, FAConv has a clear learning goal to find a periodic law to complete the mapping and updating of features. (3) The periodic mapping and rectified feature maps are concatenated to get the final output. This can help the feature recovery of the overall network with the fast learning of the output layer. The overall computation of FAConv is as follows:

$$\begin{aligned}
&X_1 = \omega_{C/4} X, \quad X_2 = \omega_{C/2} X, \\
&Y = Concat\left(Cosine\left(X_1\right), Sine\left(X_1\right), GELU\left(X_2\right)\right),
\end{aligned} \tag{11}$$

in which $Concat(\cdot)$ is the concatenate operation and X and Y are the input and output, respectively. $\omega_{C/4}$ and $\omega_{C/2}$ are the weights of the convolution whose output channels are $C/4$ and$C/2$, respectively.

Experiment

Experimental setup

Datasets and evaluation metrics: We conduct a large number of comparative experiments on three publicly available image fusion datasets, namely MSRS (Tang et al. 2022), M^3FD (Liu et al. 2022), and FMB (Liu et al. 2023). To verify the performance of the fusion method, common EN, SD, CC, SCD, SSIM and MS_SSIM are employed for evaluation.

Implementation details: Our model is built using Pytorch and accelerated through an RTX 4090 GPU. We choose SGD algorithm as the optimizer, in which the initial learning rate is set to 0.001. Finally, 100 epochs are performed on the training set of MSRS (batch_size is set to 12 and the image size is resized to 352×352).

Method	EN	SD	CC	SCD	SSIM	MS_SSIM	Params/M	GFLOPs	Memory/G
SwinFusion (Ma et al. 2022)	6.731	47.533	0.594	1.718	**1.020**	1.039	0.898	118.53	18.92
LRRNet (Li et al. 2023a)	6.335	35.835	0.517	0.860	0.486	0.778	0.048	5.71	8.12
CDDFuse (Zhao et al. 2023)	6.798	47.995	0.596	1.666	1.002	1.017	1.19	220.92	16.93
DATFuse (Tang et al. 2023b)	6.576	40.477	0.592	1.439	0.907	1.000	**0.010**	2.21	**0.47**
DSFusion (Liu et al. 2024)	6.682	45.482	0.555	1.396	0.899	0.974	0.428	51.25	2.17
CrossFuse (Li and Wu 2024)	6.630	40.477	0.538	1.089	0.885	0.974	12.00	51.88	1.11
MRFS (Zhang et al. 2024)	6.617	42.595	0.570	1.374	0.763	0.905	134.82	55.99	2.33
EMMA (Zhao et al. 2024b)	**6.826**	**49.505**	0.591	1.652	0.962	1.027	1.52	16.75	1.17
SAGE (Wu et al. 2025)	6.137	40.059	0.595	1.475	0.899	0.974	0.136	8.20	0.55
GIFNet (Cheng et al. 2025)	6.010	35.547	0.624	1.413	0.826	1.030	0.725	91.21	4.74
AIDFusion (Ours)	6.736	45.540	**0.634**	**1.860**	1.002	**1.077**	0.247	**1.85**	0.83

Table 1: Quantitative comparison of AIDFusion with other baseline methods on the MSRS dataset. Optimal and sub-optimal values are highlighted by **bolding** and underlining.

Method	M^3FD						FMB					
	EN	SD	CC	SCD	SSIM	MS_SSIM	EN	SD	CC	SCD	SSIM	MS_SSIM
SwinFusion	6.611	30.838	0.587	1.456	1.014	0.997	6.680	35.504	0.591	1.597	0.994	1.024
LRRNet	6.331	23.972	0.620	1.381	0.775	0.850	6.307	26.728	0.594	1.355	0.742	0.830
CDDFuse	6.747	32.859	0.611	1.584	**1.029**	1.032	6.788	37.272	0.600	1.661	**1.006**	1.054
DATFuse	6.308	24.104	0.556	1.147	0.900	0.886	6.317	26.197	0.568	1.290	0.921	0.961
DSFusion	6.487	28.172	0.492	0.874	0.911	0.866	6.535	32.164	0.505	1.080	0.903	0.967
CrossFuse	6.402	24.746	0.498	0.840	0.908	0.878	6.502	31.028	0.506	0.956	0.913	0.915
MRFS	6.836	35.307	0.529	1.136	0.938	0.896	6.796	**39.194**	0.529	1.254	0.898	0.952
EMMA	6.750	33.030	0.572	1.399	0.933	0.992	6.786	37.305	0.574	1.532	0.895	1.033
SAGE	6.737	33.147	0.635	1.660	1.002	1.040	6.814	36.879	0.617	1.730	0.975	1.081
GIFNet	6.896	**35.407**	0.634	1.717	0.902	1.023	**6.876**	37.670	0.622	1.749	0.856	1.039
AIDFusion (Ours)	**6.916**	34.615	**0.652**	**1.803**	0.985	**1.071**	6.851	35.756	**0.631**	**1.815**	0.936	**1.092**

Table 2: Performance comparison of AIDFusion with other baseline methods on M^3FD, FMB datasets. Optimal and sub-optimal values are highlighted by **bolding** and underlining.

Comparison Experiment

In order to verify the excellent performance of the present method, we compare the method with other widely used baseline methods for IVIF on the above three public datasets. These include ten classical methods, such as: SwinFusion (Ma et al. 2022), LRRNet (Li et al. 2023a), CDDFuse (Zhao et al. 2023), DATFuse (Tang et al. 2023b), DSFusion (Liu et al. 2024), CrossFuse (Li and Wu 2024), MRFS (Zhang et al. 2024), EMMA (Zhao et al. 2024b), SAGE (Wu et al. 2025), and GIFNet (Cheng et al. 2025). All comparison methods use their published weights.

Quantitative Comparison. We report the results on MSRS dataset as shown in Table 1. We can get the following 3 observations: (1) Methods such as LRRNet, DATFuse, and SAGE have poor overall fusion performance although they have less computational overhead. (2) Methods such as SwinFusion, CDDFuse, and EMMA improve the fusion performance but ignore the lightweighting. (3) AIDFusion obtains the best performance in CC, SCD and MS_SSIM, and the second best in SSIM. Specifically, AIDFusion improves 0.182, 2.642, 0.054, 0.424, 0.126, and 0.099 in EN, SD, CC, SCD, SSIM, and MS_SSIM, as compared to the mean value of all methods. Taken together, our method, accomplished through interweaving of DIILS, MMFM, and FAConv, is able to comprehensively focus on structure, informativeness, and features aspects, with very little computational overhead.

Qualitative Comparison. We demonstrate the fusion image of AIDFusion with other advanced methods, as shown in Figure 1. The following 3 cases exist: (1) Blue box: it can be seen that both IR and VI in the region have rich information, and the fusion method is needed to select the important information of both for integration in a reasonable way. (2) Red box: IR has key information missing from VI, and the fusion method is needed to supplement the features used in IR. (3) Green ellipse: VI has rich information (IR is almost close to noise), and the fusion method is needed to filter out the influence of IR. Through careful observation, only AIDFusion is able to cope with the above 3 situations at the same time, effectively verifying the superiority of our method.

Generalization Comparison. We directly test the model trained on MSRS on two other IR-VI datasets (M^3FD and FMB) to verify the generalization of this method. The experimental results are shown in Table 2, AIDFusion obtains the best performance on both datasets for CC, SCD, and MS_SSIM, and for EN it achieves the optimal on M^3FD and

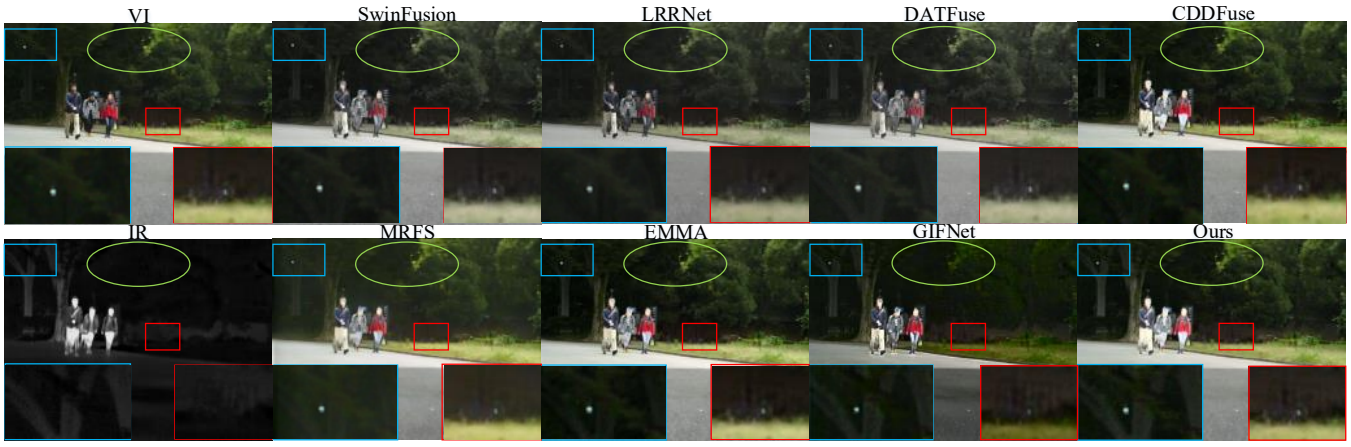

Figure 6: Qualitative comparison of AIDFusion with other baseline methods on MSRS dataset.

Method	EN	SD	SCD	SSIM
AIDFusion (Ours)	**6.736**	**45.540**	**1.860**	**1.002**
w/o DIILS:①	6.665	44.157	1.815	**1.002**
w/o MMFM:②	6.603	43.687	1.788	0.998
w/o FAConv:③	6.667	43.762	1.819	0.99
w/o ①&②	6.51	42.636	1.695	0.701
w/o ①&③	6.603	42.812	1.775	0.986
w/o ②&③	6.542	42.236	1.752	0.986
w/o ①&②&③	6.38	39.184	1.654	0.899

Table 3: Ablation experiments of AIDFusion.

sub-optimal on FMB, and AIDFusion's SD and SSIM are also better than most methods. Specifically, on the M^3FD dataset compared to all methods' mean value, AIDFusion achieves 0.296, 4.288, 0.076, 0.464, 0.050, and 0.118 enhancements for EN, SD, CC, SCD, SSIM, and MS_SSIM. On the FMB dataset, AIDFusion achieves 0.200, 1.570, 0.058, 0.375, 0.021, and 0.099 improvements for EN, SD, CC, SCD, SSIM, and MS_SSIM. In summary, our method under motivation through DIILS reinforcement has excellent generalization to unknown samples.

Computational overhead analysis. Table 1 reports Params, GFLOPs, and Memory of each method (input images size are 352×352). AIDFusion not only has comprehensive fusion performance (Table 1, 2, Figure 6), but also has low computational overhead (Rank: 4/11, 1/11, 3/11 in terms of Params, GFLOPs, Memory). When SAGE does not use distillation scheme, our Params, GFLOPs, Memory rankings are 3, 1, 2. This is due to the fact that our method takes a different approach and analyzes the design from a new perspective of "utilization rate".

Ablation Study

In order to show the contribution of each part of the present method, we perform the following ablation experiments on the MSRS dataset, as shown in Table 3. (1) From the second to fifth rows of the table, all the performances are degraded

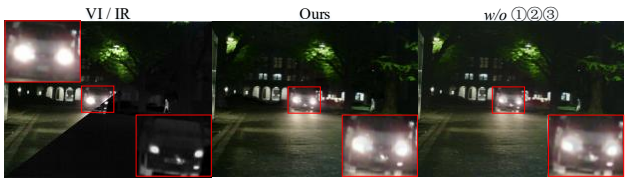

Figure 7: Visual presentation of the AIDFusion ablation experiment on MSRS dataset.

to different degrees when the learning strategy①, fusion module②, and FAConv③are removed from this method respectively. This proves that all the designs are effective for the final method. (2) Further, when this method removes two of them at the same time, the metrics further decrease. In particular, the SSIM metric plummets from 1.002 to 0.701 when both the DIILS①and the MMFM②are removed (both of which capture multiscale information). (3) When all three of the above are removed, the metrics that drop further come to a less-than-ideal situation. This proves that the design of our method is complementary and indispensable. Figure 7 shows that after removing the designed DIILS, MMFM, and FAConv, the network is more of a "lazy" copy of VI and IR, lacking discriminability (e.g. vehicle lights).

Conclusion

In this work, we propose a lightweight IVIF method, AIDFusion. Particularly, we find and consider "network laziness" phenomenon in IVIF, and propose an effective mitigation method, DIILS, to effectively exploit the network potential. Secondly, a lightweight multi-dimensional feature fusion module, MMFM, is proposed to enhance upper potential limit by extracting higher-order information with limited computation. Finally, a periodic modeling convolution, FAConv, is redesigned that can potentially build a bridge between feature fusion and prediction. Based on this, extensive experiments on three publicly available datasets demonstrate AIDFusion's superiority of performance and computational overhead.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant no. 62576298), the Sichuan Science and Technology Program (Grant no. 2025ZNSFSC0474, 2024ZDZX0004), and the Mianyang Science and Technology Program (Grant no. 2025ZYDF096).

References

Chen, J.; Kao, S.-h.; He, H.; Zhuo, W.; Wen, S.; Lee, C.-H.; and Chan, S.-H. G. 2023. Run, don't walk: chasing higher FLOPS for faster neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12021–12031.

Cheng, C.; Xu, T.; Feng, Z.; Wu, X.; Tang, Z.; Li, H.; Zhang, Z.; Atito, S.; Awais, M.; and Kittler, J. 2025. 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*, 28102–28112.

Dong, Y.; Li, G.; Tao, Y.; Jiang, X.; Zhang, K.; Li, J.; Deng, J.; Su, J.; Zhang, J.; and Xu, J. 2024. Fan: Fourier analysis networks. *arXiv preprint arXiv:2410.02675*.

Ellmauthaler, A.; Pagliari, C. L.; and Da Silva, E. A. 2012. Multiscale image fusion using the undecimated wavelet transform with spectral factorization and nonorthogonal filter banks. *IEEE Transactions on image processing*, 22(3): 1005–1017.

Geirhos, R.; Jacobsen, J.-H.; Michaelis, C.; Zemel, R.; Brendel, W.; Bethge, M.; and Wichmann, F. A. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11): 665–673.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Howard, A.; Sandler, M.; Chu, G.; Chen, L.-C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. 2019. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1314–1324.

Li, H.; and Wu, X.-J. 2018. DenseFuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing*, 28(5): 2614–2623.

Li, H.; and Wu, X.-J. 2024. CrossFuse: A novel cross attention mechanism based infrared and visible image fusion approach. *Information Fusion*, 103: 102147.

Li, H.; Wu, X.-J.; and Kittler, J. 2021. RFN-Nest: An end-to-end residual fusion network for infrared and visible images. *Information Fusion*, 73: 72–86.

Li, H.; Xu, T.; Wu, X.-J.; Lu, J.; and Kittler, J. 2023a. Lrrnet: A novel representation learning guided fusion network for infrared and visible images. *IEEE transactions on pattern analysis and machine intelligence*, 45(9): 11040–11052.

Li, H.; Yang, Z.; Zhang, Y.; Jia, W.; Yu, Z.; and Liu, Y. 2025a. MulFS-CAP: Multimodal fusion-supervised cross-modality alignment perception for unregistered infrared-visible image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Li, X.; Pan, Y. P.; Sun, Y.; Sun, Q.; Sun, Y.; W. Tsang, I.; and Ren, Z. 2025b. Incomplete Multi-view Clustering with Paired and Balanced Dynamic Anchor Learning. *IEEE Transactions on Multimedia*, 7087–7098.

Li, X.; Pan, Y. P.; Sun, Y.; Sun, Q. S.; Tsang, I. W.; and Ren, Z. 2024. Fast Unpaired Multi-view Clustering.

Li, X.; Ren, Z.; Sun, Q.; and Xu, Z. 2023b. Auto-weighted tensor schatten p-norm for robust multi-view graph clustering. *Pattern Recognition*, 134: 109083.

Li, X.; Sun, Y.; Sun, Q.; Ren, Z.; and Sun, Y. 2023c. Cross-view graph matching guided anchor alignment for incomplete multi-view clustering. *Information Fusion*, 100: 101941.

Liu, J.; Fan, X.; Huang, Z.; Wu, G.; Liu, R.; Zhong, W.; and Luo, Z. 2022. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5802–5811.

Liu, J.; Lin, Z.; and Jiang, L. 2024. Laziness, barren plateau, and noises in machine learning. *Machine Learning: Science and Technology*, 5(1): 015058.

Liu, J.; Liu, Z.; Wu, G.; Ma, L.; Liu, R.; Zhong, W.; Luo, Z.; and Fan, X. 2023. Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8115–8124.

Liu, K.; Li, M.; Chen, C.; Rao, C.; Zuo, E.; Wang, Y.; Yan, Z.; Wang, B.; Chen, C.; and Lv, X. 2024. DSFusion: Infrared and visible image fusion method combining detail and scene information. *Pattern Recognition*, 154: 110633.

Liu, Y.; Liu, S.; and Wang, Z. 2015. A general framework for image fusion based on multi-scale transform and sparse representation. *Information fusion*, 24: 147–164.

Liu, Z.; Wang, Y.; Vaidya, S.; Ruehle, F.; Halverson, J.; Soljacic, M.; Hou, T. Y.; and Tegmark, M. 2025. KAN: Kolmogorov–Arnold Networks. In *The Thirteenth International Conference on Learning Representations*.

Lou, Z.; Xue, H.; Wang, Y.; Zhang, C.; Yang, X.; and Hu, S. 2025. Parameter-Free Deep Multi-Modal Clustering With Reliable Contrastive Learning. *IEEE Transactions on Image Processing*.

Lu, J.; Wu, Z.; Chen, Z.; Cai, Z.; and Wang, S. 2024. Towards multi-view consistent graph diffusion. In *Proceedings of the 32nd ACM international conference on multimedia*, 186–195.

Luo, X.; Chen, P.; Liu, C.; Jin, X.; Wen, J.; Liu, Y.; and Wang, J. 2025a. Enhancing Multimodal Protein Function Prediction Through Dual-Branch Dynamic Selection with Reconstructive Pre-Training. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, 7598–7606.

Luo, X.; Pu, Z.; Xu, Y.; Wong, W. K.; Su, J.; Dou, X.; Ye, B.; Hu, J.; and Mou, L. 2021. MVDRNet: Multi-view diabetic retinopathy detection by combining DCNNs and attention mechanisms. *Pattern Recognition*, 120: 108104.

Luo, X.; Xu, Q.; Wang, Z.; Huang, C.; Liu, C.; Jin, X.; and Zhang, J. 2024. A Lesion-Fusion Neural Network for Multi-View Diabetic Retinopathy Grading. *IEEE Journal of Biomedical and Health Informatics*.

Luo, X.; Xu, Q.; Wu, H.; Liu, C.; Lai, Z.; and Shen, L. 2025b. Like an Ophthalmologist: Dynamic Selection Driven Multi-View Learning for Diabetic Retinopathy Grading. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 19224–19232.

Ma, J.; Tang, L.; Fan, F.; Huang, J.; Mei, X.; and Ma, Y. 2022. SwinFusion: Cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA Journal of Automatica Sinica*, 9(7): 1200–1217.

Ram Prabhakar, K.; Sai Srikar, V.; and Venkatesh Babu, R. 2017. Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs. In *Proceedings of the IEEE international conference on computer vision*, 4714–4722.

Tang, L.; Deng, Y.; Ma, Y.; Huang, J.; and Ma, J. 2022. SuperFusion: A versatile image registration and fusion network with semantic awareness. *IEEE/CAA Journal of Automatica Sinica*, 9(12): 2121–2137.

Tang, L.; Yuan, J.; and Ma, J. 2022. Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network. *Information Fusion*, 82: 28–42.

Tang, L.; Zhang, H.; Xu, H.; and Ma, J. 2023a. Rethinking the necessity of image fusion in high-level vision tasks: A practical infrared and visible image fusion network based on progressive semantic injection and scene fidelity. *Information Fusion*, 99: 101870.

Tang, W.; He, F.; Liu, Y.; Duan, Y.; and Si, T. 2023b. DATFuse: Infrared and visible image fusion via dual attention transformer. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(7): 3159–3172.

Wan, X.; Liu, J.; Yu, H.; Qu, Q.; Li, A.; Liu, X.; Liang, K.; Dong, Z.; and Zhu, E. 2024. Contrastive continual multiview clustering with filtered structural fusion. *IEEE Transactions on Neural Networks and Learning Systems*.

Wang, D.; Huang, C.; Pan, H.; Sun, Y.; Dai, J.; Li, Y.; and Ren, Z. 2025. AMLCA: Additive multi-layer convolution-guided cross-attention network for visible and infrared image fusion. *Pattern Recognition*, 163: 111468.

Woo, S.; Park, J.; Lee, J.-Y.; and Kweon, I. S. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, 3–19.

Wu, G.; Liu, H.; Fu, H.; Peng, Y.; Liu, J.; Fan, X.; and Liu, R. 2025. Every SAM Drop Counts: Embracing Semantic Priors for Multi-Modality Image Fusion and Beyond. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 17882–17891.

Xu, H.; Ma, J.; Jiang, J.; Guo, X.; and Ling, H. 2020. U2Fusion: A unified unsupervised image fusion network. *IEEE transactions on pattern analysis and machine intelligence*, 44(1): 502–518.

Xu, Q.; Luo, X.; Huang, C.; Liu, C.; Wen, J.; Wang, J.; and Xu, Y. 2024. HACDR-Net: heterogeneous-aware convolutional network for diabetic retinopathy multi-lesion segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6342–6350.

Yang, M.; Huang, Z.; and Peng, X. 2024. Robust object re-identification with coupled noisy labels. *International Journal of Computer Vision*, 132(7): 2511–2529.

Yi, X.; Xu, H.; Zhang, H.; Tang, L.; and Ma, J. 2024. 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*, 27026–27035.

Yuan, H.; Sun, Y.; Zhou, F.; Wen, J.; Yuan, S.; You, X.; and Ren, Z. 2025. Prototype Matching Learning for Incomplete Multi-view Clustering. *IEEE Transactions on Image Processing*.

Zhang, H.; Xu, H.; Xiao, Y.; Guo, X.; and Ma, J. 2020. Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 12797–12804.

Zhang, H.; Zuo, X.; Jiang, J.; Guo, C.; and Ma, J. 2024. Mrfs: Mutually reinforcing image fusion and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26974–26983.

Zhang, J.; Zhao, Z.; Li, C.; and Yu, Y. 2025a. Lightweight yet Fine-grained: A Graph Capsule Convolutional Network with Subspace Alignment for Shared-account Sequential Recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 13242–13250.

Zhang, X.; Wang, Q.; Wang, P.; and Wang, W. 2025b. A Lightweight Sparse Interaction Network for Time Series Forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 13304–13312.

Zhao, L.; Xie, Q.; Li, Z.; Wu, S.; and Yang, Y. 2024a. Dynamic Graph Guided Progressive Partial View-Aligned Clustering. *IEEE Transactions on Neural Networks and Learning Systems*.

Zhao, Z.; Bai, H.; Zhang, J.; Zhang, Y.; Xu, S.; Lin, Z.; Timofte, R.; and Van Gool, L. 2023. Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5906–5916.

Zhao, Z.; Bai, H.; Zhang, J.; Zhang, Y.; Zhang, K.; Xu, S.; Chen, D.; Timofte, R.; and Van Gool, L. 2024b. Equivariant multi-modality image fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 25912–25921.

Zhao, Z.; Deng, L.; Bai, H.; Cui, Y.; Zhang, Z.; Zhang, Y.; Qin, H.; Chen, D.; Zhang, J.; Wang, P.; et al. 2024c. Image Fusion via Vision-Language Model. In *International Conference on Machine Learning*, 60749–60765.