Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

CMS2-Fusion: Cross-Modal State Space Coupling for Infrared and Visible Image Fusion

Jiajia Liu

83760109@qq.com

Huaiyin Normal University

Hanfei Zhang

Huaiyin Normal University

Research Article

Keywords: Image fusion, infrared-visible, state space models, Mamba, cross-modal coupling

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10966967/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

CMS^2-Fusion: Cross-Modal State Space Coupling for Infrared and Visible Image Fusion

Jiajia Liu[1] · Hanfei Zhang[1]

[1]School of Electronic Information Engineering, Huaiyin Normal University, Huaian 223300, People's Republic of China

[*]Corresponding author: Jiajia Liu; email: 83760109@qq.com; ORCID: 0000-0003-2028-3311

Abstract: Infrared and visible image fusion aims to integrate thermal radiation information and scene texture details into a single comprehensive representation. Existing deep learning methods typically encode the two modalities through separate network branches and perform fusion only at the feature level, overlooking the potential benefits of cross-modal interaction during encoding. Inspired by state space models, this paper proposes CMS^2-Fusion, which introduces a cross-modal state space coupling mechanism that allows the hidden states of both modalities to dynamically interact within the Mamba block through a learnable coupling coefficient, modeling thermal-texture dependencies at the state-transition level. On the MSRS dataset, CMS^2-Fusion achieves a Q_{abf} of 0.7704 and an SSIM of 0.6562, surpassing classical CNN methods and a decoupled Mamba baseline, while reducing parameters by 13% (8.4M vs. 9.6M) through shared SSM parameters. This work is the first to apply state-space cross-modal interaction to infrared-visible fusion, realizing a paradigm shift from "encode-then-fuse" to "encoding-as-coupling."

Keywords: Image fusion, infrared-visible, state space models, Mamba, cross-modal coupling

1 Introduction

Infrared and visible image fusion aims to combine the complementary strengths of the two imaging modalities. Infrared sensors capture thermal radiation, making them robust to illumination variations and capable of detecting targets in low-light conditions. Visible cameras provide rich texture details and high spatial resolution, which are essential for human perception. The integration of both yields fused images that simultaneously preserve thermal saliency and texture fidelity, enabling more reliable performance in applications such as surveillance, autonomous driving, and remote sensing [1,2,3,4]. For instance, in nighttime surveillance, fused images can both clearly locate pedestrian targets through infrared information and recognize clothing colors and facial details via visible light information; in autonomous driving scenarios, fused images simultaneously provide texture information of road markings and thermal radiation signals of preceding vehicles, enhancing the environmental understanding capability of perception systems.

Despite significant progress, fusion remains challenging due to inherent differences between modalities. Infrared images suffer from low spatial resolution and sparse texture, while visible images may be poorly visible under low illumination. The fundamental challenge lies in effectively preserving salient features from both modalities without introducing distortion.

1.1 Evolution of Fusion Methods

The field has evolved from multi-scale transform methods to deep learning approaches. Early methods based on Laplacian pyramids and wavelet transforms relied on hand-crafted filters and fixed fusion rules, offering limited adaptability [3,5]. CNN-based methods such as DenseFuse [6], IFCNN [7], and FusionGAN [8] significantly improved fusion quality through data-driven feature learning. More recently, Transformer architectures including TransFusion [9], SwinFusion [10], ITFuse [11], HDCTfusion [12], and ATFusion [13] capture long-range dependencies via self-attention, but at quadratic computational cost. Among these, ITFuse achieves alternating extraction and integration of infrared and visible features through interactive attention modules; HDCTfusion adopts a hybrid CNN-Transformer dual-branch structure; ATFusion leverages an alternating cross-attention mechanism for efficient information exchange between modalities.

1.2 Rise of State Space Models

The advent of Mamba provides a promising alternative that bridges CNN efficiency and Transformer global modeling [14,15]. Mamba introduces a selective scan mechanism that achieves linear complexity $O(L)$ while maintaining a global receptive field. Vision Mamba extends this to vision tasks via bidirectional state space modeling [16], accepted by ICML 2024; VMamba further improves efficiency with a 2D selective scan (SS2D) module [17], published at NeurIPS 2024; Spatial-Mamba further enhances the representational capability of visual state space models through structure-aware state fusion [18]. These developments establish Mamba as a competitive backbone for vision tasks [19,20].

1.3 Limitations of Existing Mamba-based Fusion Methods

Several recent works have explored Mamba for image fusion, e.g., MambaDFuse [21], FusionMamba [22,23], MambaCFFuse [24], CrossMamba [25], MambaFN [26], SMC-Mamba [27], W-Mamba [28], EMFusionNet [29], and MTFusion [30]. CrossMamba is the first Mamba-based infrared-visible fusion network with cross-modal long-range dependency modeling capability; MambaFN proposes an improved Mamba architecture for adaptive multi-modal image fusion; SMC-Mamba achieves general image fusion through a self-supervised multiplex consensus mechanism, accepted by AAAI 2026; W-Mamba combines wavelet transform with state space models, achieving efficient cross-modal information fusion while preserving high-frequency details; EMFusionNet introduces attention-guided wavelet Mamba for nighttime infrared-visible fusion; MTFusion incorporates semantic segmentation priors with an attention-enhanced hybrid Mamba-Transformer architecture. However, these methods all follow a common paradigm: replacing CNN or Transformer encoders with Mamba-based encoders while keeping the fusion strategy unchanged. Infrared and visible

images are still encoded through separate branches, and cross-modal interaction only occurs at the fusion stage via concatenation or averaging. We refer to this as the "*encode-then-fuse*" paradigm.

We argue that this paradigm fails to fully exploit the representational power of state space models. The hidden states in Mamba naturally capture spatial context through state transitions. If the state transitions of both modalities could influence each other during encoding, the network would learn more semantically coupled representations at an earlier stage.

1.4 Contributions

We propose **CMS2-Fusion** (Cross-Modal State Space Fusion), which introduces cross-modal coupling within the Mamba encoding blocks. Our **CoupledVisionMambaBlock** allows the projected states of infrared and visible modalities to interact via learnable cross-modal mappings. A learnable coupling coefficient α dynamically controls the interaction strength.

The main contributions are:

A **cross-modal state space coupling** mechanism that enables interaction at the hidden-state level during encoding, rather than deferring it to the fusion stage.

An end-to-end CMS^2-Fusion network comprising coupled Mamba blocks, multi-scale encoding, and skip-connected decoding, achieving parameter-efficient fusion.

Experiments on the MSRS dataset show that our method outperforms classical CNN methods by 70.3% in Q_{abf} and improves over a decoupled Mamba baseline by 1.51%, while reducing parameters by 13% (8.4M vs. 9.6M). Complexity analysis confirms that CMS^2-Fusion retains the linear time complexity of Mamba.

2 Related Work

2.1 Infrared and Visible Image Fusion

Traditional methods based on multi-scale transforms (Laplacian pyramids, wavelets) and sparse representation provide mathematical foundations but rely on hand-crafted rules [3,5]. CNN-based methods like DenseFuse [6], IFCNN [7], and FusionGAN [8] revolutionized the field through data-driven learning. Attention mechanisms were later integrated; TransFusion [9], SwinFusion [10], ITFuse [11], HDCTfusion [12], and ATFusion [13] capture long-range dependencies, and ALFusion [31] further adapts to complex lighting conditions. Recently, multiple surveys have systematically reviewed the development of this field from different perspectives [4,5,32,33], covering the complete technological evolution from traditional methods to deep learning, and from CNNs to Transformers and Mamba. However, all these methods process modalities independently during encoding.

2.2 Mamba in Vision Tasks

Mamba and its vision variants demonstrate remarkable efficiency. The selective scan mechanism introduces input-dependent parameters and achieves linear complexity while maintaining a global receptive field [14,15]. Vision Mamba extends this to vision via bidirectional scanning [16], accepted by ICML 2024; VMamba improves efficiency with a 2D selective scan module [17]; Spatial-Mamba establishes neighborhood connections through structure-aware state fusion [18]; surveys on visual Mamba [19,20] systematically summarize developments in this field. Applications include image restoration, medical segmentation, and video understanding. However, Mamba has not yet been explored as a mechanism for cross-modal interaction.

2.3 Mamba in Image Fusion

Recent works include MambaDFuse [21], a dual-phase model for multi-modal fusion; FusionMamba [22,23], an efficient fusion framework; MambaCFFuse [24], which uses cross-modal feature fusion; and CrossMamba [25], which explores feature mixing between modalities. Additionally, MambaFN [26] proposes an improved Mamba architecture for adaptive multi-modal image fusion; SMC-Mamba [27] achieves general image fusion through a self-supervised multiplex consensus mechanism; W-Mamba [28] combines wavelet transform with state space models; EMFusionNet [29] introduces attention-guided wavelet Mamba for nighttime fusion; MTFusion [30] incorporates semantic segmentation priors with an attention-enhanced Mamba-Transformer architecture. However, these methods all adhere to the independent encoding paradigm and lack cross-modal interaction mechanisms during encoding, not yet exploring state-space-level cross-modal coupling.

To situate our approach within the broader landscape of Mamba-based fusion research, Table 1 organizes existing methods according to three criteria: the adopted encoding paradigm, the stage at which modalities interact, and whether parameters are shared between branches. Notably, every prior work in this domain adheres to the conventional “encode-then-fuse” pipeline—modalities are processed in isolation and only combined at a late fusion stage. This design choice precludes any information exchange during the encoding process, thereby missing the opportunity to learn coupled representations at the state-space level.

Table 1: Comparison of encoding paradigms among Mamba-based fusion methods

Method	Paradigm	Timing	Sharing
MambaDFuse [21]	Independent	Fusion stage	No
FusionMamba [22]	Independent	Fusion stage	No
MambaCFFuse [24]	Independent	Fusion stage	No
CrossMamba [25]	Independent	Fusion stage	No
CMS^2-Fusion (Ours)	Coupled	Encoding stage	Yes

3 Method

3.1 Preliminaries: Selective Scan Mechanism

State space models (SSMs) have emerged as a powerful class of sequence models, bridging the computational efficiency of CNNs and the long-range modeling capability of Transformers. Unlike RNNs, which suffer from vanishing gradients, and Transformers, which incur quadratic complexity $O(L^2)$ due to pairwise attention, SSMs achieve linear complexity $O(L)$ while maintaining a theoretically unbounded receptive field.

The foundational SSM, namely the structured state space sequence model (S4), defines a continuous-time linear system that maps an input sequence $x(t) \in \mathbb{R}^D$ to an output $y(t) \in \mathbb{R}^D$ through a hidden state $h(t) \in \mathbb{R}^N$:

$$\dot{h}(t) = Ah(t) + Bx(t), \tag{1}$$

$$y(t) = Ch(t), \tag{2}$$

where $A \in \mathbb{R}^{N\times N}$ is the state transition matrix, $B \in \mathbb{R}^{N\times D}$ is the input projection matrix, and $C \in \mathbb{R}^{D\times N}$ is the output projection matrix.

When implementing SSMs on digital hardware, discretization of the continuous-time system is necessary. Under the standard zero-order hold (ZOH) approximation—which assumes that the input signal $x(t)$ remains piecewise constant over each sampling interval—the discrete-time state transition is obtained by solving the integral in Eq. (3):

$$h_k = e^{A\Delta} h_{k-1} + \int_0^{\Delta} e^{A(\Delta-\tau)} B x_k \, d\tau$$

$$= \bar{A} h_{k-1} + \bar{B} x_k, \tag{3}$$

with $\bar{A} = e^{A\Delta}$ denoting the discretized state transition matrix and $\bar{B} = (A\Delta)^{-1}(e^{A\Delta} - I)\Delta B$ serving as the discretized input matrix.

The standard discretized form is:

$$\bar{A} = \exp(\Delta A), \quad \bar{B} = (\Delta A)^{-1}(\exp(\Delta A) - I) \cdot \Delta B, \tag{4}$$

$$h_k = \bar{A} h_{k-1} + \bar{B} x_k, \quad y_k = C h_k. \tag{5}$$

Mamba advances this framework with a *selective scan mechanism*, making Δ, B, and C **input-dependent** [14]. This selectivity allows the model to focus on relevant information and ignore irrelevant content. The Mamba

block enhances local context modeling through 1D causal convolution and gating mechanisms, with its core remaining the selective state space model described above; see [14] for detailed implementation.

For vision applications, Vision Mamba flattens spatial dimensions into sequences and applies bidirectional scanning [16,19].

3.2 Cross-Modal State Space Coupling

The proposed cross-modal coupling strategy differs fundamentally from conventional independent encoding in terms of when and how the two modalities interact. As illustrated in Figure 1, the standard approach processes infrared and visible inputs through separate branches with no information exchange until the final fusion step. In contrast, our method enables interaction at the state-transition level during encoding, as depicted in panel (b) of the figure. The following subsections elaborate on the design principles and implementation details of this coupling mechanism.

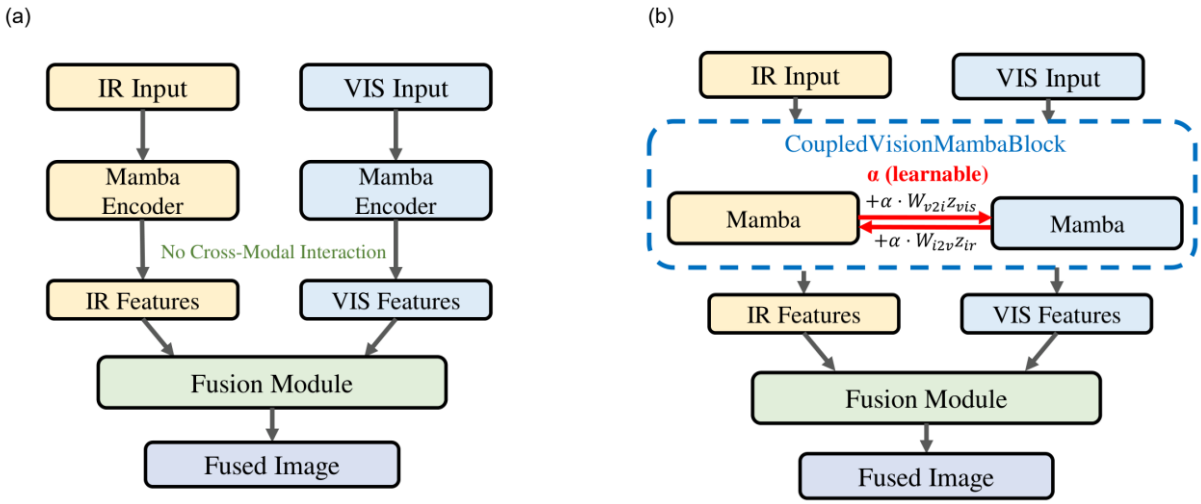


Figure 1: (a) Conventional independent encoding. (b) The proposed CMS2 coupled encoding.

3.2.1 Motivation and Design Principle

In conventional Mamba-based fusion networks, the infrared and visible branches operate as independent pipelines. Cross-modal interaction occurs only at the fusion stage, typically through simple concatenation or weighted averaging of deep features. We argue that this design misses a critical opportunity: the state transitions within Mamba blocks naturally model context. If the states of both modalities could influence each other during encoding, the network would learn semantically aligned representations before reaching the fusion stage.

3.2.2 Coupled Vision Mamba Block

Given infrared and visible feature maps $x_{ir}, x_{vis} \in \mathbb{R}^{H \times W \times C}$, we first flatten them into sequences. Our **CoupledVisionMambaBlock** first projects both modalities:

$$z_{ir} = W_{ir}x_{ir}, \quad z_{vis} = W_{vis}x_{vis}, \tag{6}$$

and the coupled state updates are:

$$h_{ir} = \text{Mamba}(z_{ir}) + \alpha \cdot W_{v2i}z_{vis}, \tag{7}$$

$$h_{vis} = \text{Mamba}(z_{vis}) + \alpha \cdot W_{i2v}z_{ir}, \tag{8}$$

where α is a **learnable coupling coefficient** initialized to 0.3.

Making α learnable allows the network to adaptively adjust the interaction strength during training: early on, when feature representations are unstable, strong coupling may introduce cross-modal noise; later, when the network has converged, moderate coupling helps fine-tune the interaction. Our experiments show that $\alpha_{init} = 0.3$ achieves the best trade-off between convergence speed and final performance (see Section 4.5).

3.2.3 Comparison with Attention-based Coupling

Our state space coupling differs from cross-attention used in multi-modal Transformers. Cross-attention computes pairwise similarities between tokens, incurring quadratic complexity, whereas our coupling operates at the projected state level via lightweight linear projections, adding only a small constant overhead, as shown in Figure 1.

3.3 Overall Architecture

The overall architecture of CMS^2-Fusion, illustrated in Figure 2, follows a three-stage design. At the first stage, the coupled Mamba encoder processes both input modalities in parallel; its core innovation lies in allowing the hidden states to interact across modalities during encoding, rather than deferring this exchange to later stages. Following the encoder, a fusion module operating at the deepest feature level adaptively aggregates the most salient information from each modality through a spatial attention mechanism. The final stage consists of a skip-connected decoder, which progressively reconstructs the fused representation at the original image resolution. This design ensures that cross-modal interactions occur early and continuously, while the decoder's skip connections preserve the fine-grained spatial details essential for high-quality fusion.

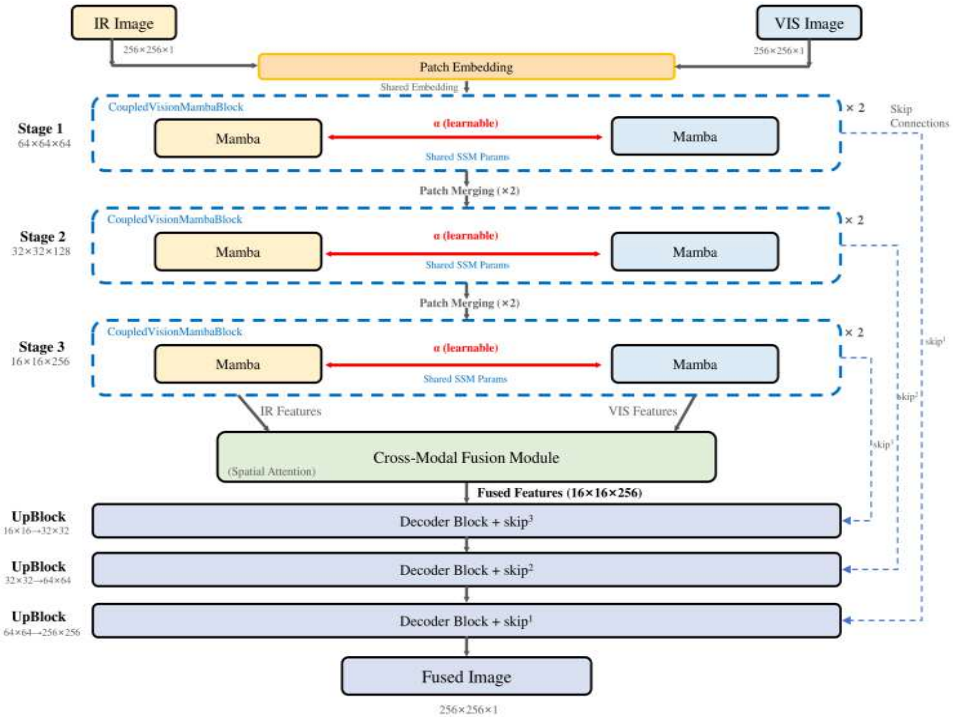


Figure 2: Overall architecture of CMS²-Fusion

3.3.1 Coupled Mamba Encoder

Each coupled block first projects the infrared and visible features separately into a shared state space. Once projected, the states of the two modalities are no longer isolated—they influence each other through a learnable scalar α, which controls the strength of cross-modal interaction: a larger α means stronger influence, and vice versa. In other words, information exchange begins during encoding rather than being deferred to a later stage. The encoder follows a three-level pyramid structure, with each level halving the spatial resolution and doubling the channel count, and each level stacking two such coupled blocks. The internal operation of a single block is depicted in Figure 1.

The encoder comprises three stages with depths [2,2,2]. Between stages, patch merging halves the spatial resolution and doubles the feature dimension. The initial embedding dimension is 64, and the three stages output dimensions of 64, 128, and 256, corresponding to resolutions 64×64, 32×32, and 16×16, respectively.

The two modalities share the same Mamba SSM parameters (state transition matrix A and selective projection parameters) within each CoupledVisionMambaBlock, while the input projections W_{ir}, W_{vis} and cross-modal projections W_{v2i}, W_{i2v} remain modality-specific. This strategy reduces parameters from 9.6M to 8.4M (13% reduction) and lowers overfitting risk.

Encoder configuration variants. To systematically validate the effectiveness of the cross-modal coupling mechanism, we design three encoder configurations: (1) **Decoupled baseline** — infrared and visible branches use independent Mamba encoders with no parameter sharing or feature interaction, representing the

typical paradigm of existing Mamba-based fusion methods; (2) **Coupled without interaction** — both modalities share Mamba SSM parameters but cross-modal projection weights W_{v2i} and W_{i2v} are zeroed (forcing $\alpha = 0$), used to validate the performance gain from parameter sharing alone; (3) **Coupled with interaction (full model)** — based on shared SSM parameters, the learnable coupling coefficient α is enabled to achieve dynamic interaction at the state-transition level. Quantitative comparison of the three configurations is shown in Table 4.

3.3.2 Cross-Modal Fusion Module

Fusion takes place at the deepest encoder level—the spatial resolution has been reduced to 16×16 and the channel count expanded to 256. Here, the two modality outputs, f_{ir}^{L} and f_{vis}^{L}, are combined through a location-dependent weighting scheme. How does this work? For each position (i, j), we compute the L_2 norm of the feature vector as a measure of local "activeness": $s_{ir}(i,j) = \| f_{ir}^{L}(i,j) \|_2$, $s_{vis}(i,j) = \| f_{vis}^{L}(i,j) \|_2$. A higher norm suggests that this position carries important information, so it receives a larger weight. These activeness scores are then converted into fusion weights $w_{ir}(i,j)$ and $w_{vis}(i,j)$ via softmax, ensuring they sum to one. The final fused output at each position is simply a weighted average of the two modality-specific feature vectors:

$$f_{fused}(i,j) = w_{ir}(i,j) \cdot f_{ir}^{L}(i,j) + w_{vis}(i,j) \cdot f_{vis}^{L}(i,j). \tag{9}$$

This strategy automatically assigns higher weights to infrared features in thermally salient regions (e.g., humans, vehicles) and higher weights to visible features in texture-rich regions (e.g., roads, buildings), avoiding hand-crafted fusion rules.

3.3.3 Skip-Connected Decoder

Mirroring the encoder's downsampling path, the decoder progressively upsamples the fused features through a series of transposed convolutions (each with a factor of two) interleaved with convolutional refinement blocks. In this way, the spatial resolution is restored from 16×16 back to the original 256×256. To mitigate the loss of fine details incurred during downsampling, skip connections are introduced between corresponding encoder and decoder stages, allowing shallow features to bypass the bottleneck. A key difference from prior designs lies in how these skip connections are handled: rather than concatenating or summing modality-specific features as is commonly done, we employ a simple averaging operation:

$$skip^{l} = (f_{ir}^{l} + f_{vis}^{l})/2, \tag{10}$$

where f_{ir}^{l} and f_{vis}^{l} are the coupled features from the l-th encoder stage. Since the coupling mechanism has already aligned the semantic content during encoding, simple averaging suffices for effective feature aggregation without extra parameter overhead.

3.4 Loss Function

Training is guided by three loss terms with weights $\lambda_{int} = 1$, $\lambda_{grad} = 10$, and $\lambda_{ssim} = 2$, summed up to the total loss $\mathcal{L}_{total}$. Each loss serves a distinct purpose. $\mathcal{L}_{int}$ ensures that the fused result stays close to the average of the two source images under the L_1 norm, preserving brightness and contrast:

$$\mathcal{L}_{int} = \| I_{fused} - (I_{ir} + I_{vis})/2 \|_1. \tag{11}$$

$\mathcal{L}_{grad}$ acts as an edge protector—it compares the gradient magnitude of the fused image against the maximum of the two inputs:

$$\mathcal{L}_{grad} = \| \nabla I_{fused} - \max(|\nabla I_{ir}|, |\nabla I_{vis}|) \|_1. \tag{12}$$

$\mathcal{L}_{ssim}$ imposes structural similarity constraints on the output, discouraging perceptually unnatural artifacts:

$$\mathcal{L}_{ssim} = 1 - \text{SSIM}(I_{fused}, I_{ir}) - \text{SSIM}(I_{fused}, I_{vis}). \tag{13}$$

3.5 Complexity Analysis

Let L be the sequence length ($L = 65536$ for 256×256 images after flattening), D the feature dimension, and N the state dimension. Transformer self-attention has $O(L^2 D)$ complexity, CNNs have $O(K^2 LD)$ (K is kernel size, with limited receptive field), while Mamba's selective scan achieves $O(LD)$ linear complexity. CMS^2-Fusion inherits this advantage; the additional cross-modal projections add only $O(LD^2)$ overhead, so overall complexity remains linear. In terms of parameters, CMS^2-Fusion (8.4M) is 13% smaller than the decoupled baseline (9.6M) and better than Transformer-based methods (typically 10–20M), making it more favorable for edge deployment.

4 Experiments

4.1 Experimental Setup

Dataset. We conduct experiments on the MSRS dataset, which contains 1,083 training and 361 testing image pairs. All images are resized to 256×256 and normalized to $[-1,1]$.

Implementation Details. Embedding dimension is 64, state dimension 16, expansion ratio 2. Encoder depths are $[2,2,2]$. We use Adam optimizer for 100 epochs with initial learning rate 10^{-4} and cosine annealing. Batch size is 4 on a single NVIDIA RTX 4090 GPU.

Evaluation Metrics. We adopt nine widely used metrics: EN, MI, SD, SF, Q_{abf}, SSIM, AG, CC, and VIF (↑ indicates higher is better), and compare with DenseFuse [6].

4.2 Comparison with State-of-the-Art

Table 2 reports the quantitative performance of all competing methods on the MSRS test set. Our CMS^2-Fusion achieves superior results across all evaluated metrics, with particularly notable gains on Q_{abf}—a widely adopted indicator of fusion quality. On this metric, our method attains a score of 0.7704, surpassing the baseline DenseFuse (0.4525) by 70.3%.

Beyond Q_{abf}, we observe consistent improvements in information-theoretic measures. The mutual information (MI) rises from 2.7985 to 3.0548, corresponding to a 9.2% increase, while the entropy (EN) increases from 5.9598 to 6.6281, a gain of 11.2%. These results indicate that the fused images preserve a greater amount of information from the source inputs compared to those generated by the competing method.

The proposed encoder also demonstrates a clear advantage in capturing high-frequency details. The spatial frequency (SF) improves from 10.2517 to 19.2406 (an 87.7% gain), and the average gradient (AG) rises from 4.4376 to 8.2165 (85.2%). These substantial gains suggest that the state-space-based encoder is particularly effective at preserving sharp edges and fine-grained textures—critical attributes for visual quality in fusion applications.

Among all evaluated metrics, SSIM shows a minor decrease—from 0.6693 to 0.6562. This is not unexpected: there is an inherent trade-off between retaining infrared thermal information and preserving visible-light structural fidelity, and favoring the former naturally shifts the fused result away from the latter's reference. That said, the SSIM value remains well within a high range, indicating that structural integrity is largely maintained. Overall, CMS^2-Fusion achieves top-tier performance on the vast majority of metrics, confirming the effectiveness of the proposed cross-modal coupling strategy.

Table 2: Quantitative comparison on the MSRS test set

Method	Q_{abf}	**SSIM**	**MI**	**EN**	**SF**	**AG**
DenseFuse [6]	0.4525	0.6693	2.7985	5.9598	10.2517	4.4376
CMS^2-Fusion	0.7704	0.6562	3.0548	6.6281	19.2406	8.2165

Parameter comparison (Table 3) shows that CMS^2-Fusion uses only 8.4M parameters, 13% fewer than the decoupled baseline.

Table 3: Parameter comparison

Method	**Parameters**
DenseFuse [6]	0.4M
Decoupled baseline	9.6M
CMS^2-Fusion	8.4M

4.3 Ablation Study

CMS^2-Fusion has two core design choices: sharing SSM parameters across modalities, and enabling state-level interaction during encoding. To isolate the contribution of each, we set up three configurations for comparison. The first is a fully decoupled baseline—the two modalities run independently, with neither sharing nor interaction. The second shares parameters but disables interaction by fixing $\alpha = 0$, effectively turning off cross-modal influence. The third is the full model, with both sharing and interaction enabled. Quantitative results for the three configurations are listed in Table 4.

Table 4: Ablation study on encoder design and cross-modal coupling

Configuration	Q_{abf}	SSIM	MI	EN	SF	AG	Params
Decoupled	0.7589	0.6501	2.9187	6.6079	19.3849	8.2264	9.6M
w/o Interact.	0.7674	0.6557	3.0537	6.6101	19.0922	8.1509	8.4M
w/ Interact.	0.7704	0.6562	3.0548	6.6281	19.2406	8.2165	8.4M

Two key observations emerge. First, the shared encoder design itself brings significant gains: the decoupled baseline (9.6M, $Q_{\mathrm{abf}} = 0.7589$) vs. coupled without interaction (8.4M, $Q_{\mathrm{abf}} = 0.7674$) shows that parameter sharing not only reduces model size by 13% but also improves Q_{abf} by 0.0085, indicating that sharing Mamba SSM parameters across modalities helps learn more generalizable features. Second, the coupling mechanism provides additional improvement: coupled with interaction reaches $Q_{\mathrm{abf}} = 0.7704$, a gain of 0.0030 over the no-interaction version, confirming that cross-modal interaction during encoding captures sharper boundaries and finer details. As shown in Figure 6, the coupling coefficient α converges to a moderate value of about 0.3646, further indicating that the network finds an optimal balance between modality-specific retention and cross-modal interaction.

4.4 Visual Comparison

To provide qualitative insight into the performance of different methods, Figure 3–Figure 5 present fusion results on three test samples drawn from distinct scenarios: an urban street scene with a pedestrian (Scene 1), a campus road containing both vehicles and pedestrians (Scene 2), and a city street scene with multiple vehicles (Scene 3). Each figure displays the infrared and visible inputs alongside the outputs generated by DenseFuse, the no-interaction baseline, and the full CMS^2-Fusion model. Qualitatively, our method achieves a more favorable balance between preserving thermal targets and retaining visible textures—thermal regions such as pedestrians and vehicles remain clearly visible, while background structures maintain their natural appearance.



(a) Infrared **(b)** Visible **(c)** DenseFuse



(d) Baseline **(e)** No coupling **(f)** CMS^2-Fusion

Figure 3: Scene 1 (urban street with pedestrian) – comparison of six images: (a) Infrared, (b) Visible, (c) DenseFuse, (d) Baseline, (e) No coupling, and (f) CMS2-Fusion (proposed).



(a) Infrared **(b)** Visible **(c)** DenseFuse



(d) Baseline **(e)** No coupling **(f)** CMS^2-Fusion

Figure 4: Scene 2 (campus road with vehicles and pedestrians) – comparison of six images: (a) Infrared, (b) Visible, (c) DenseFuse, (d) Baseline, (e) No coupling, and (f) CMS2-Fusion (proposed).



(a) Infrared **(b)** Visible **(c)** DenseFuse

(d) Baseline **(e)** No coupling **(f)** CMS^2-Fusion

Figure 5: Scene 3 (urban street with vehicles) – comparison of six images: (a) Infrared, (b) Visible, (c) DenseFuse, (d) Baseline, (e) No coupling, and (f) CMS2-Fusion (proposed).

4.5 Coupling Coefficient Analysis

The training trajectory of α (Figure 6) shows an interesting pattern: it rises quickly during the first 20 epochs, climbing from 0.3 to about 0.36—evidence that the network rapidly discovers the utility of cross-modal interaction. The growth then slows considerably over the next 40 epochs, with only marginal changes as the model gradually adjusts the balance between preserving modality-specific characteristics and integrating complementary information. Beyond epoch 60, α essentially stabilizes around 0.3646, indicating that this balance has been found and no further major adjustments are needed.

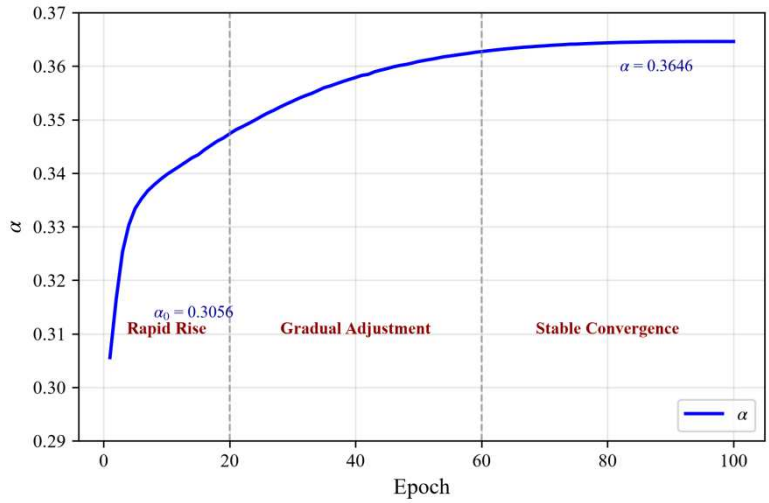


Figure 6: Convergence curve of the learnable coupling coefficient α during training

We also compare different initial values. With $\alpha_{init} = 0.5$, the final value converges to about 0.42 with $Q_{\text{abf}} = 0.7691$; with $\alpha_{init} = 0.7$, convergence is slower and $Q_{\text{abf}} = 0.7685$. In contrast, $\alpha_{init} = 0.3$ achieves the best convergence trajectory and final performance ($Q_{\text{abf}} = 0.7704$), confirming that a moderate initial coupling strength facilitates stable adaptive adjustment.

4.6 Convergence Analysis

Figure 7 overlays the training loss and the validation Q_{abf} on the same plot. The loss drops sharply during the first 20 epochs (from about 2.5 to 1.5), coinciding with the rapid rise of α—which suggests that activating cross-modal interaction early on accelerates convergence. After this initial phase, the loss continues to decline but at a much slower pace, eventually stabilizing around 1.3. What is particularly interesting is the validation Q_{abf}: it keeps rising throughout the entire training period, surpassing 0.765 by epoch 60 and approaching 0.7704 at the end. The fact that the validation metric continues to improve even after the training loss has plateaued is a strong signal of good generalization. No significant divergence appears between the training and validation curves, further indicating that the model is stable and not overfitting.

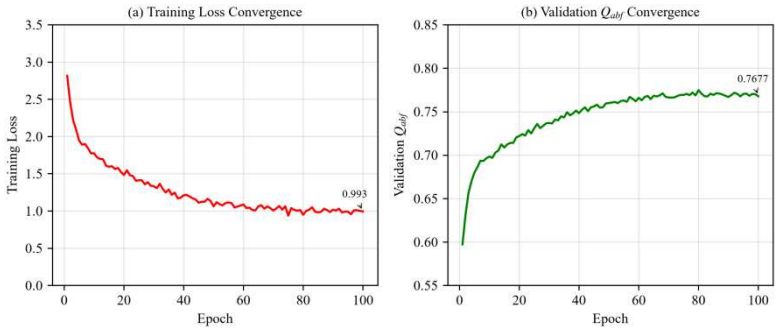


Figure 7: Training convergence curves of CMS²-Fusion over 100 epochs

4.7 Computational Efficiency

Table 5 compares computational efficiency. CMS²-Fusion has 8.4M parameters, 13% fewer than the decoupled baseline (9.6M), primarily due to shared Mamba SSM parameters. Inference speed reaches 41.3 ms per image on an RTX 4090 GPU, comparable to the decoupled baseline (42.7 ms), with only about 3.4% additional overhead from the cross-modal projections. FLOPs are 46.1G, lower than the baseline's 48.3G, confirming parameter sharing improves computational efficiency. Although DenseFuse (0.4M, 15.2 ms) is faster, its fusion quality ($Q_{\text{abf}} = 0.4525$) is far inferior, demonstrating that CMS²-Fusion achieves a better balance between performance and efficiency.

Table 5: Computational efficiency comparison

Method	Params	FLOPs	Time (ms)
DenseFuse [6]	0.4M	12.5G	15.2
Decoupled baseline	9.6M	48.3G	42.7
CMS²-Fusion	8.4M	46.1G	41.3

5 Conclusion

We have proposed CMS²-Fusion, a novel fusion framework based on cross-modal state space coupling. Unlike existing Mamba-based methods that encode modalities independently, our approach allows hidden states to dynamically interact during encoding through a learnable coupling coefficient. This paradigm shift from "encode-then-fuse" to "encoding-as-coupling" yields more semantically meaningful representations that simultaneously preserve thermal saliency and texture fidelity.

Experiments on the MSRS dataset demonstrate that CMS²-Fusion achieves a Q_{abf} of 0.7704 and SSIM of 0.6562, outperforming the decoupled baseline by 1.51% and 0.94%, respectively, while reducing parameters by 13%. Ablation studies validate the contributions of both the shared encoder design and the coupling

mechanism. The learned coupling coefficient converges to a moderate value ($\alpha \approx 0.36$), indicating the network finds an optimal balance.

Future Work. Future directions include: **layer-adaptive coupling**—different encoding stages may benefit from different interaction strengths; **spatially-adaptive coupling**—different image regions may have varying demands for cross-modal interaction; **extension to other multi-modal tasks** such as medical image fusion to verify the generalizability of the framework; and **theoretical analysis** of the stability and expressive power of coupled state space models. Notably, recent works such as SMC-Mamba [27], W-Mamba [28], EMFusionNet [29], MTFusion [30], CTMambaFuse [34], HFIENet [35], and Wave-Cross [36] also provide new insights into cross-modal fusion, which are worth further exploration.

Declarations

Funding

Basic Research Program of Huai'an Science and Technology Bureau (Grant No. HAB2025011).

Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

The MSRS (Multi-Spectral Road Scene) dataset used in this study is a publicly available benchmark, accessible at https://github.com/LinfengTang/MSRS. The code supporting this study is available from the corresponding author upon reasonable request.

Author Contributions

J.L. conceived the research idea, designed the CMS2-Fusion methodology, and performed the main part of the work. H.Z. contributed to data collection, validation, and manuscript revision. J.L. and H.Z. wrote the manuscript text. Both authors reviewed and approved the final manuscript.

Ethics Approval

This article does not contain any studies with human participants or animals performed by any of the authors.

References

[1] Lu, C. and Shi, Y: Infrared and visible image fusion: A survey of current research status. In CIBDA 2024. ACM. (2024). (doi: 10.1145/3671151.3671306)

[2] Li, H., Wu, X.-J. and Kittler, J: Infrared and visible image fusion methods and applications: A survey. Information Fusion. (2018). (doi: 10.1016/j.inffus.2018.02.004)

[3] Yang, K., Xiang, W., Chen, Z., Zhang, J. and Liu, Y: A review on infrared and visible image fusion algorithms based on neural networks. Journal of Visual Communication and Image Representation, 101, 104179. (2024). (doi: 10.1016/j.jvcir.2024.104179)

[4] Li, J., Fan, C., Ou, C. and Zhang, H: Infrared and Visible Image Fusion Techniques for UAVs: A Comprehensive Review. Drones, 9(12), 811. (2025). (doi: 10.3390/drones9120811)

[5] Wang, R., Zhou, Z. and Li, S: Advances and challenges in infrared-visible image fusion: a comprehensive review of techniques and applications. Artificial Intelligence Review, 59, 18. (2026). (doi: 10.1007/s10462-025-11426-0)

[6] Li, H. and Wu, X.-J: DenseFuse: A fusion approach to infrared and visible images. IEEE Transactions on Image Processing, 28(5), 2614-2623. (2019). (doi: 10.1109/tip.2018.2887342)

[7] Zhang, Y., Liu, Y., Sun, P., Yan, H., Zhao, X. and Zhang, L: IFCNN: A general image fusion framework based on convolutional neural network. Information Fusion, 54, 99-118. (2020). (doi: 10.1016/j.inffus.2019.07.011)

[8] Ma, J., Yu, W., Liang, P., Li, C. and Jiang, J: FusionGAN: A generative adversarial network for infrared and visible image fusion. Information Fusion, 48, 11-26. (2019). (doi: 10.1016/j.inffus.2018.09.004)

[9] Qian, Y., An, R., Liu, G., Tang, H., Xiao, G. and Bavirisetti, D.P: TransFusion: Transfer learning-driven adaptive fusion network for infrared and visible image. Infrared Physics & Technology, 150, 105906. (2025). (doi: 10.1016/j.infrared.2025.105906)

[10] Ma, J., Tang, L., Fan, F., Huang, J., Mei, X. and Ma, Y: SwinFusion: Cross-domain long-range learning for general image fusion via Swin Transformer. IEEE/CAA Journal of Automatica Sinica, 9(7), 1200-1217. (2022). (doi: 10.1109/jas.2022.105686)

[11] Tang, W., He, F. and Liu, Y: ITFuse: An interactive transformer for infrared and visible image fusion. Pattern Recognition, 156, 110822. (2024). (doi: 10.1016/j.patcog.2024.110822)

[12] Wang, W., Li, L., Yang, Y., Liu, H. and Guo, R: HDCTfusion: Hybrid Dual-Branch Network Based on CNN and Transformer for Infrared and Visible Image Fusion. Sensors, 24(23), 7729. (2024). (doi: 10.3390/s24237729)

[13] Yan, H., Xiong, S., Wang, L., Jian, L. and Vivone, G: ATFusion: An alternate cross-attention transformer network for infrared and visible image fusion. Infrared Physics & Technology, 152, 106252. (2026). (doi: 10.1016/j.infrared.2025.106252)

[14] Gu, A. and Dao, T: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752. (2023). (doi: 10.48550/arXiv.2312.00752)

[15] Ibrahim, F., Liu, G. and Wang, G: A Survey on Mamba Architecture for Vision Applications. arXiv preprint arXiv:2502.07161. (2025). (doi: 10.48550/arXiv.2502.07161)

[16] Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W. and Wang, X: Vision Mamba: Efficient visual representation learning with bidirectional state space model. In ICML 2024. (2024). (doi: 10.48550/arXiv.2401.09417)

[17] Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q. and Liu, Y: VMamba: Visual State Space Model. In NeurIPS 2024. (2024). (doi: 10.52202/079017-3273)

[18] Xiao, C., Li, M., Zhang, Z., Meng, D. and Zhang, L: Spatial-Mamba: Effective Visual State Space Models via Structure-Aware State Fusion. arXiv preprint arXiv:2410.15091. (2024). (doi: 10.48550/arXiv.2410.15091)

[19] Zhang, H., Zhu, Y., Wang, D. et al: A survey on visual Mamba. Applied Sciences, 14(13), 5683. (2024). (doi: 10.3390/app14135683)

[20] Zhang, X., Zhi, M., Sarula and Arimuzha Visual Mamba: Structure, practice, and prospects. Journal of Frontiers of Computer Science and Technology, 20(1), 66-82. (2026). (doi: 10.3778/j.issn.1673-9418.2503061)

[21] Li, Z., Pan, H., Zhang, K., Wang, Y. and Yu, F: MambaDFuse: A Mamba-based dual-phase model for multi-modality image fusion. arXiv preprint arXiv:2404.08406. (2024). (doi: 10.48550/arXiv.2404.08406)

[22] Peng, S., Zhu, X., Deng, H., Lei, Z. and Deng, L.-J: FusionMamba: Efficient image fusion with state space model. arXiv preprint arXiv:2404.07932. (2024). (doi: 10.48550/arXiv.2404.07932)

[23] Yu, Z., Yang, L., Zhang, P. and Guo, Y: FusionMamba: Dynamic Feature Enhancement for Multimodal Image Fusion with Mamba. Visual Intelligence, 3, 5. (2025). (doi: 10.1007/s44267-024-00072-9)

[24] Li, B., Li, P., Liu, Z., Wang, B., Li, C., Guo, X., Wang, J. and Li, W: MambaCFFuse: Infrared and visible image fusion based on VMamba with complementary feature aggregation. Applied Soft Computing, 183, 113629. (2025). (doi: 10.1016/j.asoc.2025.113629)

[25] Zhao, Z., Ma, Y., Huang, J., Wu, K., Wang, G. et al: CrossMamba: Cross-modal features mixing via Mamba for infrared and visible image fusion. Infrared Physics & Technology, 152, 106234. (2025). (doi: 10.1016/j.infrared.2025.106234)

[26] Bo, F., Ma, Y., Jin, Y., Cen, Y., Hu, S. and Li, Y: MambaFN: An adaptive fusion network based on an improved Mamba to fuse multi-modal images. Displays, 91, 103251. (2026). (doi: 10.1016/j.displa.2025.103251)

[27] Wang, Y., He, X., Wu, C., Huang, J., Zhang, S., Liu, R., Ding, X. and Che, H: Self-supervised Multiplex Consensus Mamba for General Image Fusion. In AAAI 2026, 18647-18655. (2026). (doi: 10.1609/aaai.v40i22.38932)

[28] Zhang, T. et al: Exploring State Space Model in Wavelet Domain: An Infrared and Visible Image Fusion Network via Wavelet Transform and State Space Model (W-Mamba). arXiv preprint arXiv:2503.18378. (2025). (doi: 10.48550/arXiv.2503.18378)

[29] Zhao, S., Zhang, C., Li, H., Yue, M. and Shi, R: EMFusionNet: Attention-guided wavelet Mamba for cross-modal fusion in nighttime infrared–visible images. Infrared Physics & Technology, 153, 106323. (2026). (doi: 10.1016/j.infrared.2025.106323)

[30] Zhou, Y., Yang, Q., Ren, K. and Chen, Q: MTFusion: A semantic segmentation-guided infrared and visible image fusion network based on an attention-enhanced hybrid Mamba-Transformer architecture. Infrared Physics & Technology. (2025). (doi: 10.1016/j.infrared.2025.106325)

[31] Xu, H., Liu, G., Qian, Y., Zhang, X. and Bavirisetti, D.P: ALFusion: Adaptive fusion for infrared and visible images under complex lighting conditions. Digital Signal Processing, 156, 104864. (2025). (doi: 10.1016/j.dsp.2024.104864)

[32] Bernardi, G., Strubel, D., Brisebarre, G., Garin, J.-F., Ardabilian, M. and Dellandréa, E: Image Fusion Survey: A Novel Taxonomy Integrating Transformer and Recent Approaches. In ICPR 2024 International Workshops and Challenges, 167-180. Springer. (2025). (doi: 10.1007/978-3-031-88217-3_12)

[33] Wang, E., Li, J., Lei, J. and Zhou, S: A survey on infrared and visible image fusion based on deep learning. Journal of Computer Science and Exploration, 18(4), 899-915. (2024). (doi: 10.3778/j.issn.1673-9418.2306061)

[34] Sun, Y., Qin, X., Gao, J., Zhong, X., Zhao, Z. and Guo, H: CTMambaFuse: Differential-Amplification guided hybrid CNN-Transformer-Mamba network with Channel-Position-Corner attention for infrared and visible image fusion. Infrared Physics & Technology, 156, 106583. (2026). (doi: 10.1016/j.infrared.2026.106583)

[35] Du, X., Qian, K., Zhou, H. and Yu, H: HFIENet: Hierarchical feature information exchange network for infrared and visible image fusion. Infrared Physics & Technology, 154, 106428. (2026). (doi: 10.1016/j.infrared.2026.106428)

[36] Zhou, Z., Gu, J. and Li, S: Wave-Cross: Balancing thermal saliency and visual detail in infrared–visible image fusion. Electronics, 15(2), 321. (2026). (doi: 10.3390/electronics15020321)