Information Fusion 117 (2025) 102839

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full length article

Conti-Fuse: A novel continuous decomposition-based fusion framework for infrared and visible images

Hui Li [a,*], Haolong Ma [a], Chunyang Cheng [a], Zhongwei Shen [b], Xiaoning Song [a], Xiao-Jun Wu [a]

[a] School of Artificial Intelligence and Computer Science, Jiangnan University, 214122, Wuxi, China
[b] School of Electronic & Information Engineering, Suzhou University of Science and Technology, 215009, Suzhou, China

ARTICLE INFO

Keywords:
Image decomposition
Image fusion
Multimodality
Common feature

ABSTRACT

For better explore the relations of inter-modal and inner-modal, even in deep learning fusion framework, the concept of decomposition plays a crucial role. However, the previous decomposition strategies (base & detail or low-frequency & high-frequency) are too rough to present the common features and the unique features of source modalities, which leads to a decline in the quality of the fused images. The existing strategies treat these relations as a binary system, which may not be suitable for the complex generation task (e.g. image fusion). To address this issue, a continuous decomposition-based fusion framework (Conti-Fuse) is proposed. Conti-Fuse treats the decomposition results as few samples along the feature variation trajectory of the source images, extending this concept to a more general state to achieve continuous decomposition. This novel continuous decomposition strategy enhances the representation of complementary information of inter-modal by increasing the number of decomposition samples, thus reducing the loss of critical information. To facilitate this process, the continuous decomposition module (CDM) is introduced to decompose the input into a series continuous components. The core module of CDM, State Transformer (ST), is utilized to efficiently capture the complementary information from source modalities. Furthermore, a novel decomposition loss function is also designed which ensures the smooth progression of the decomposition process while maintaining linear growth in time complexity with respect to the number of decomposition samples. Extensive experiments demonstrate that our proposed Conti-Fuse achieves superior performance compared to the state-of-the-art fusion methods.

1. Introduction

As a fundamental field of image processing, image fusion seeks to create informative and visually appealing images by extracting the most significant information from various source images [1–4]. One of the notable challenges in image processing is Infrared and Visible Image Fusion (IVIF), which entails integrating complementary information from distinct modalities [5–7]. In the IVIF task, the input comprises both infrared and visible images. Visible images are distinguished by their abundant texture information, which aligns more closely with human visual perception. However, they are susceptible to lighting variations, occlusion, and other factors, resulting in the loss of vital information. In contrast, infrared images excel in highlighting targets in extreme conditions (e.g., low light) by capturing thermal radiation but are prone to noise. Consequently, in IVIF tasks, the fused image must mitigate the shortcomings of both modalities to achieve superior visual quality [8].

Image decomposition, as a crucial technique, is frequently applied in image fusion tasks. Depending on the spatial domain in which the decomposition occurs, image decomposition methods can be broadly classified into the Shallow Feature space-based Image Decomposition Methods (SFID) [10,11] and the Deep Feature space-based Image Decomposition Methods (DFID) [9,12,13]. In early traditional image fusion methods, SFIDs are the most prevalent. Among these, multi-scale transform (MST)-based decomposition methods [14,15] are particularly popular. These methods employ manually design decomposition operations, such as discrete Fourier transforms, to decompose the original image into coefficients at multiple scales. Subsequently, fusion strategies and corresponding inverse transformations are applied to the coefficients to obtain the fused image. These methods achieve excellent results in early image fusion methods. However, since they only perform decomposition at the shallow feature space using manually designed strategies, these SFIDs lack adaptability to the source images, resulting in poor generalization capabilities.

To overcome the limitations of SFID, many approaches utilize deep neural networks to decompose at deeper feature space [16–18], representing source images. These methods typically consist of Encoders,

* Corresponding author.
E-mail address: lihui.cv@jiangnan.edu.cn (H. Li).

https://doi.org/10.1016/j.inffus.2024.102839
Received 6 September 2024; Received in revised form 24 November 2024; Accepted 26 November 2024
Available online 2 December 2024

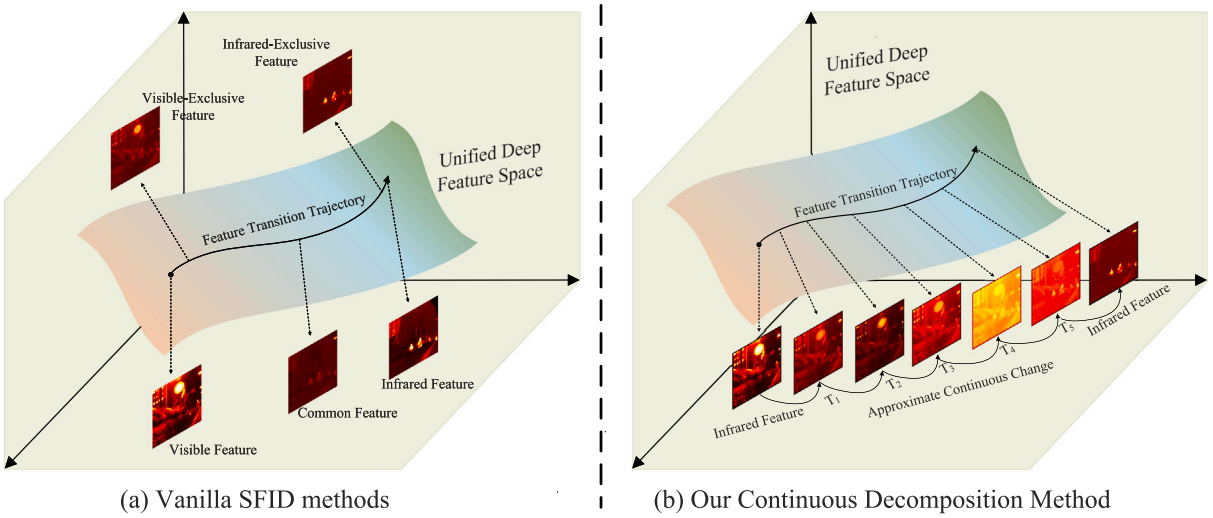

Fig. 1. A schematic of the unified deep feature space between the common SFID (e.g. DeFusion [9]) methods and our proposed continuous decomposition method.

Decomposition Modules, Fusion Modules, and Decoders. The Encoder is responsible for extracting basic features from the source images and mapping them into the Unified Deep Feature Space (UDFS) for representation. The UDFS refers to a unified feature space where two different input modalities are mapped. This enables the representation of information from different modalities within a unified space, laying the foundation for the subsequent image decomposition process. Subsequently, the Decomposition Module decomposes the features of the source images within this unified feature space into several decomposed features. Finally, these decomposed features are fused by the Fusion Module and mapped back to the pixel space by the Decoder to obtain the fused image.

Taking DeFusion [9] as an example, it uses convolutional networks as the Encoder to decompose the two source images into a shared feature and two modality-specific features within a feature space. These three sets of decomposed features are then fused and reconstructed by a learnable convolutional network to produce the fused result. However, these methods roughly decompose the original images into few non-overlapping features, leading to the loss of some critical information from the source images (such as detailed information).

To address these shortcomings, we propose a continuous decomposition-based fusion framework, referred to as Conti-Fuse. As illustrated in Fig. 1(a), we consider the decomposed features of two source images as sample points along a continuous change trajectory in the unified deep feature space from one source image feature to another. Taking DeFusion [9] as an example, its common features can be viewed as sample points in the middle of its trajectory, while the two unique features can be approximated as sample points near the two ends of its trajectory. Based on this concept, as depicted in Fig. 1(b), the proposed Conti-Fuse generalizes the decomposition in the unified deep feature space beyond the past simple decomposition method such as the common and unique features in DeFusion [9]. By performing multiple decompositions along the continuous transition trajectory in the feature space, we obtain richer and more diverse decomposed features, referred to as transition states, which more finely represent the critical information of the two source images.

Inspired by MST-based methods [19,20], we apply our continuous decomposition method to a multi-scale framework and propose the Continuous Decomposition Module (CDM) to decompose features and obtain transition states. Additionally, to fully capture the complementary information between transition states, we introduce the State Transformer to enhance the complementarity between transition states. Finally, to guide the entire decomposition process, we design a novel continuous decomposition loss function and the corresponding computational strategy, termed as Support Decomposition Strategy (SDS). SDS employs the Monte Carlo method to perform random sampling of the continuous decomposition loss, approximating the true decomposition loss with sufficient loss samples, thereby reducing the computational complexity of the continuous decomposition loss from quadratic to linear, which accelerates the training speed of the model.

The contributions of our work are summarized as follows:

- A novel decomposition strategy is introduced, which achieves enriched decomposition features by densely sampling along the variation trajectories of deep features across the two modalities. This method effectively reduces the loss of crucial information in fused images.
- An efficient decomposition loss is designed to facilitate continuous decomposition. By leveraging the Monte Carlo method, this loss function accelerates computation, thereby enhancing the scalability of the proposed approach.
- Extensive qualitative and quantitative experiments were conducted, demonstrating excellent performance of our approach compared to other state-of-the-art fusion methods.

2. Related work

2.1. Shallow feature space-based image decomposition (SFID) methods

SFID methods are commonly used in image fusion [21–23,23]. Among these methods, multi-scale transformation-based decomposition techniques are particularly popular [14,15,24]. Typically, multi-scale transformation methods involve three steps [25]: (1) Decomposing the original image into coefficients at various scales using a specific transformation, (2) Integrating the coefficients of different modalities through carefully designed fusion rules, (3) Generating the fused image from the aggregated coefficients using the inverse transformation.

For instance, Pajares et al. [14] proposed a fusion model based on discrete wavelet transform decomposition. This model decomposed two source images into multi-scale coefficients using discrete wavelet transform, followed by manually designed merging strategies and inverse discrete wavelet transform to obtain the fused image. These models reliance on manually designed decomposition operations and fusion strategies lead to poor robustness of the models.

Compared to SFID methods, our method does not rely on manually designed decomposition operations and fusion strategies, providing better robustness.

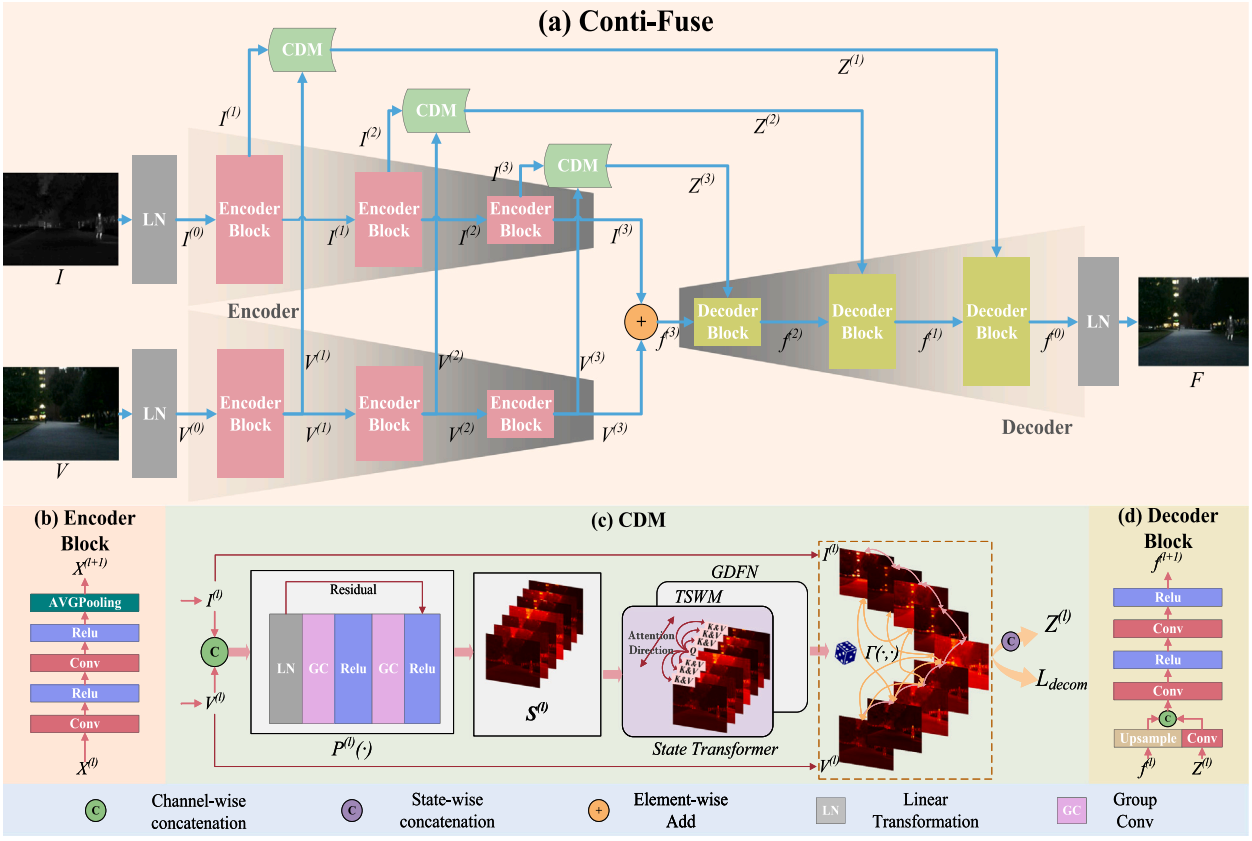

Fig. 2. The architecture of Conti-Fuse. (a) The pipeline of proposed method. (b, c, d) The internal structure diagrams for Encoder Block, CDM and Decoder Block in the l-th layer, respectively. The input of Encoder Block ($X^{(l)}$) can be visible feature or infrared feature. 'Channel-wise concatenation' and 'State-wise concatenation' refer to concatenation along the channel and state dimensions of the tensors, respectively; 'Linear Transformation' refers to a 1×1 convolution, and 'Group Conv' refers to grouped convolution.

2.2. Deep feature space-based image decomposition (DFID) methods

With the development of deep learning, many image fusion methods leverage the powerful representation capabilities of deep learning for image decomposition [26–29]. These methods typically include an Encoder and a Decoder. The Encoder maps the two source images from the original pixel space to a unified deep feature space for representation. The Decoder maps the fused features, rich in deep semantic information, back to the original pixel space to obtain the fused image [9]. The decomposition and fusion processes of features of source images are carried out in the unified deep feature space to ensure the model learns robust representations of them.

This representation can be further uniformly described as follows [30]: In a high-dimensional unified feature space, the decomposed features are considered as several sampling points on a feature transition trajectory between the source images. For example, in DeFusion [9], the three decomposed features (infrared unique features, common features, and visible unique features) are considered as three feature sampling points on the variation trajectory from infrared to visible features. Its recent successor, DeFusion++[31], follows the same decomposition strategy of common and unique features. However, existing DFID crudely decompose the source images into non-overlapping multiple features, which are sparse sampling points. This leads to insufficient representation of source images, losing much critical information.

Compared to existing DFID methods, our method offers a more general decomposition strategy with richer decomposed features, capable of preserving key information in the source images, thereby improving image quality.

3. Proposed method

In this section, firstly, the workflow of our proposed model and the detailed design of each module are introduced. Then, we provide the formulas for calculating the proposed loss function and explain the underlying design principles.

3.1. Overview

Conti-Fuse is mainly composed of three types of modules: Encoder, Decoder and The Continuous Decomposition Module (CDM). Encoder and Decoder are designed to extract shallow features from the source images and reconstructing the fused image, respectively. CDM is designed for interaction between two modalities and for generating transition states. In addition to the above three modules, there are two linear transformations (LN) at the input and output positions to adjust the number of feature channels.

As shown in Fig. 2(a), the multi-scale structure is adopted in the proposed framework. The Encoder and Decoder have several Blocks and are marked $EN^{(l)}$ and $DE^{(l)}(l \in \{1, 2, \ldots, N\}, N = 3)$, respectively. For more general discussion, we set the number of transition states to K and the number of layers to N in this section.

3.2. Encoder block

As shown in Fig. 2(b), the Encoder Block is composed of two convolutional layers with 3×3 kernels, two ReLU activation functions, and one average pooling (2×2). These settings are shared for both infrared and visible branches. The Encoder Block is utilized to extract

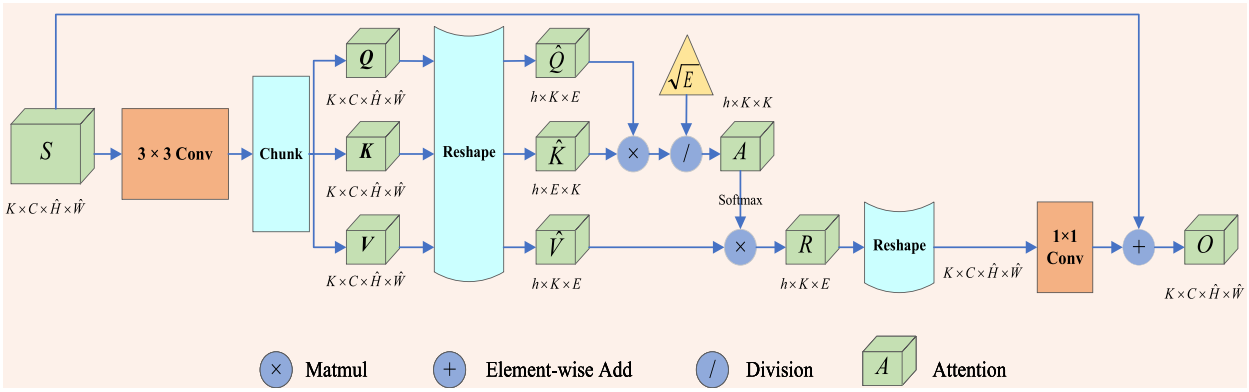

Fig. 3. Illustration of Transition State Wise MHSA (TSWM).

basic shallow features while mapping source images into a Unified Deep Feature Space (UDFS).

In this section, we introduce the following notation: $I, V \in \mathbb{R}^{H\times W}$ denote the input infrared image and visible image. Thus, the processing of input by the lth layer Encoder Block can be expressed as follows,

$$\begin{aligned} &I^{(l)} = EN^{(l)}(I^{(l-1)}), V^{(l)} = EN^{(l)}(V^{(l-1)}) \\ &\quad s.t. \quad l \in \{1, 2, \dots, N\} \end{aligned} \tag{1}$$

where $I^{(0)}$ and $V^{(0)}$ are obtained by the input linear transformation (LN, 1×1 Conv) of the sources image I and V, respectively.

3.3. Continuous decomposition module

The Continuous Decomposition Module (CDM) is proposed to decompose the input features into a continuous transition states, approximately. To do this, it is necessary to first extract information within each transition states, and then capture the complementary information from multiple modalities.

As illustrated in Fig. 2(c), a linear transformation and several group convolutions are employed. The linear transformation are utilized to generate preliminary transition states. Group convolutions further extract finer information within each transition state.

Concretely, 3×3 convolutional layer is employed in which the number of groups is set to K. Then, we obtain K transition states denoted as $S \in \mathbb{R}^{K\times C\times \hat{H}\times \hat{W}}$, where C represents the number of channels, and $\hat{H} \times \hat{W}$ represents the size of the feature maps of transition states. The sub-module in CDM is denoted as $P(\cdot)$ which is consisted by a linear transformation, two group convolutional layers and two ReLU activation functions.

3.3.1. State transformer

To leverage the complementary information as much as possible, a novel feature extractor is designed which is named State Transformer, $ST(\cdot)$. The core module of ST is the **T**ransition **S**tate **W**ise **M**HSA (TSWM), which leverages the multi-head self-attention mechanism to capture complementary relationships between transition states.

As shown in Fig. 3, the TSWM first generate **Q**, **K**, and **V** from S, which is accomplished through a linear transformation. Then, we reshape and split **Q**, **K** and **V** into multiple attention heads, yielding $\hat{\mathbf{Q}}, \hat{\mathbf{K}}, \hat{\mathbf{V}} \in R^{h\times K\times E}$, where $E \times h = C \times \hat{H} \times \hat{W}$ and h denotes the number of attention heads.

Next, standard multi-head and self-attention is applied along the transition state wise. State Transformer can be formulated as follows,

$$\begin{aligned} &\mathbf{Q} = \Phi_Q(S), \mathbf{K} = \Phi_K(S), \mathbf{V} = \Phi_V(S) \\ &Atten = softmax(\hat{\mathbf{Q}} \cdot \hat{\mathbf{K}}^T / \sqrt{E}) \\ &U = \phi_p(\hat{\mathbf{V}} \cdot Atten) + S \\ &T = GDFN(U) + U \end{aligned} \tag{2}$$

where $\Phi(\cdot)$, $\phi(\cdot)$ represent 3×3 and 1×1 convolutional layers, respectively. T denotes the result of State Attention.

To better handle complementary information between modalities in Eq. (2), we utilize the Gated-Dconv Feed-forward Network (GDFN) [32] proposed in Restormer as the Feed-Forward layer, enabling a gated approach to effectively focus on processing complementary information.

In summary, for the lth CDM module $CDM^{(l)}(\cdot,\cdot)$, we can express its transformation in the following formula,

$$\begin{aligned} &S^{(l)} = P^{(l)}([V^{(l)}; I^{(l)}|c]), T^{(l)} = ST^{(l)}(S^{(l)}) \\ &\quad Z^{(l)} = [V^{(l)}; T^{(l)}; I^{(l)}|s] \\ &\quad s.t. \quad l \in \{1, 2, \dots, N\} \end{aligned} \tag{3}$$

where $[\cdot|c]$ and $[\cdot|s]$ denote concatenation operators along with the channel wise and the state wise, respectively. The symbols mentioned above are added indexes with l to indicate they belong to the lth CDM. For instance, $ST^{(l)}$ represents the State Transformer in the lth CDM.

3.4. Decoder block

The Decoder Block is used for feature fusion and image reconstruction. Specifically, different layers of the Decoder Block perform feature fusion and image reconstruction at their respective scales. The Decoder Block conducts elementary feature fusion of the output of CDM through a 3×3 convolutional layer. Subsequently, as illustrated in Fig. 2(d), this result is concatenated with the upsampled output of the previous layer of Decoder Block along with the channel wise. Then, through two 3×3 convolutional layer and two ReLU activation functions, further fusion and reconstruction at this scale are performed.

In summary, the lth Decoder Block can be expressed as follows,

$$\begin{aligned} &f^{(N)} = I^{(N)} + V^{(N)} \\ &f^{(l-1)} = DE^{(l)}(f^{(l)}, Z^{(l)}) \\ &s.t. \quad l \in \{1, 2, \dots, N\} \end{aligned} \tag{4}$$

where we perform element-wise addition on the outputs of the Nth Encoder Block to obtain $f^{(N)}$, which serves as the input to the Nth Decoder Block. The fused image F can be generated through a linear transformation from $f^{(0)}$.

3.5. Loss function

The loss function L_{all} consists of three parts and can be written as follows:

$$L_{all} = L_{decom} + \alpha_1 L_{int} + \alpha_2 L_{grad} \tag{5}$$

where L_{decom} is the proposed decomposition loss, $L_{int} = \frac{1}{HW}\|F - \max(I, V)\|_F^2$ and $L_{grad} = \frac{1}{HW}\|\,|\nabla F| - \max(|\nabla I|, |\nabla V|)\|_F^2$ represent

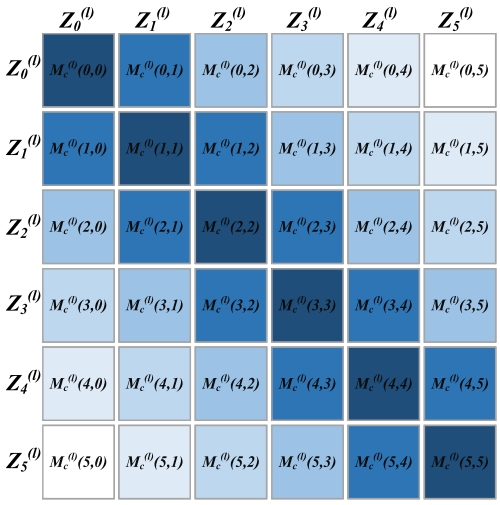

Fig. 4. An example of $M_c^{(l)}$ in the lth layer when $K = 4$. The color depth represents the constraint of distance, with darker colors indicating them closer to 1.

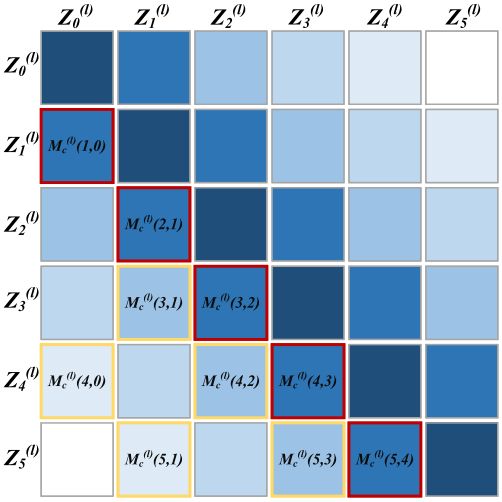

Fig. 5. An example of SDS in the lth layer when $K = 4$. The yellow boxes represent randomly sampled constraints, while the red boxes represent those calculated consistently each time.

the intensity loss and the gradient loss used in SwinFusion [33], respectively.[1] ∇ denotes the Sobel operator. α_1 and α_2 are trade-off parameters.

To achieve continuous decomposition, it is necessary to constrain the transition states within a UDFS. Let $\Gamma(\cdot,\cdot)$ denote the distance metric function in the UDFS. For the sake of simplifying subsequent calculations, we set that the greater the distance between two features, the smaller the value computed by Γ, the formula is given as follows,

$$\begin{aligned} \Gamma(X,Y) &= \frac{1}{C}\sum_{k=1}^{C} pers(X_k, Y_k) \\ pers(A,B) &= \frac{\sum_{i,j}(A_{i,j}-\bar{A})(B_{i,j}-\bar{B})}{\sqrt{\sum_{i,j}(A_{i,j}-\bar{A})^2}\sqrt{\sum_{i,j}(B_{i,j}-\bar{B})^2}} \end{aligned} \tag{6}$$

where, $X, Y \in \mathbb{R}^{C\times H\times W}$ represent two arbitrary features, with X_k and Y_k denoting the corresponding feature maps in the kth channel. Additionally, $\bar{A}$ and $\bar{B}$ represent the mean values of feature maps A and B, respectively.

[1] For more details please refer to SwinFusion [33]

We set the distance between two features with no difference as 1. Initially, we compute the distance between two source features $V^{(l)}$ and $I^{(l)}$ at lth layer in this space using Γ, denoted as $\mu = \Gamma(V^{(l)}, I^{(l)})$, where $\mu \ll 1$.

As illustrated in Fig. 4, for the output $Z^{(l)}$ of the lth layer CDM, we calculate the pairwise distances along the direction of state to obtain a symmetric distance matrix M_c. For the distance matrix $M_c^{(l)}$ of the lth layer:

$$\begin{aligned} &M_c^{(l)}(i,j) = \Gamma(Z_i^{(l)}, Z_j^{(l)}) \\ &s.t. \quad i,j \in \{0,1,\ldots,K+1\} \end{aligned} \tag{7}$$

where i and j are matrix indices, and $Z_i^{(l)}$ and $Z_j^{(l)}$ represent the ith and jth features at the lth layer. According to Eq. (3), $Z_0^{(l)}$ and $Z_{K+1}^{(l)}$ represent the visible and infrared features at the lth layer, respectively. The remaining $Z_i^{(l)}$ for $i \in \{1,2,\ldots,K\}$ correspond to the transitional states. Since the matrix is symmetric, we analyze only its lower triangular part.

By constraining the distance matrix M_c, we can impose overall constraints on the decomposition process. In M_c, the values on the main diagonal equal to 1, and we approximate the value at the lower left corner $M_c^{(l)}(K+1,0)$ to be equal to the distance μ between the source images.

For the remaining distances, we let them decay from 1 to μ along the direction from the main diagonal to the lower left corner. Thus, we construct a target matrix M_t,

$$\begin{aligned} &M_t^{(l)}(i,j) = \Omega(|i-j|, \mu, K+1) \\ &s.t. \quad i,j \in \{0,1,\ldots,K+1\} \end{aligned} \tag{8}$$

where $\Omega(\cdot,\cdot,\cdot)$ is a designed function used to compute distances during the decay process.

In this paper, the decay function denoted as Gaussian decay function and the formula are given as follows,

$$\begin{aligned} &\Omega_g(p,\mu,s) = \exp\left(-\frac{p^2}{2*\sigma}\right) \\ &s.t. \quad \sigma = -\frac{(s-1)^2}{2\ln(\mu)} \end{aligned} \tag{9}$$

where Ω_g denote Gaussian decay which reduces the distance following a Gaussian function along the sub-diagonal direction. We also explore alternative methods for constructing the decay function, with detailed experiments provided in Section 5.3.

In summary, the decomposition loss L_{decom} can be formulated as follows,

$$\begin{aligned} L_{decom} = &\frac{1}{N(K^2+3K)}\sum_{l=1}^{N}\|M_c^{(l)} - M_t^{(l)}\|_F^2 \\ s.t. \quad &M_c^{(l)}(0,K+1) = M_t^{(l)}(0,K+1) = 0 \\ &M_c^{(l)}(K+1,0) = M_t^{(l)}(K+1,0) = 0 \end{aligned} \tag{10}$$

Note that $M_c^{(l)}(0,K+1)$ and $M_c^{(l)}(K+1,0)$ are not constrained. Because they represent the distances between source features ($V^{(l)}$ and $I^{(l)}$ at lth layer). The constraints within s.t. are established to ensure that the corresponding values at the lower left corner and the upper right corner of the two matrices are equal, thereby relaxing the constraints on the features of the two source images.

Furthermore, the values on the main diagonal are equal. Thus, in both M_c and M_t, a total of $(K+2)^2 - (K+2) - 2 = K^2 + 3K$ distances require constraints. In practice, we only compute the lower triangular part of the matrix and ignore the main diagonal and $M_c^{(l)}(K+1,0)$ in lth layer.

3.6. Support decomposition strategy

The time complexity of the decomposition loss L_{decom} is $O(K^2)$ which means that as the number of transition states K increases, the computational cost of calculating the decomposition loss grows quadratically.

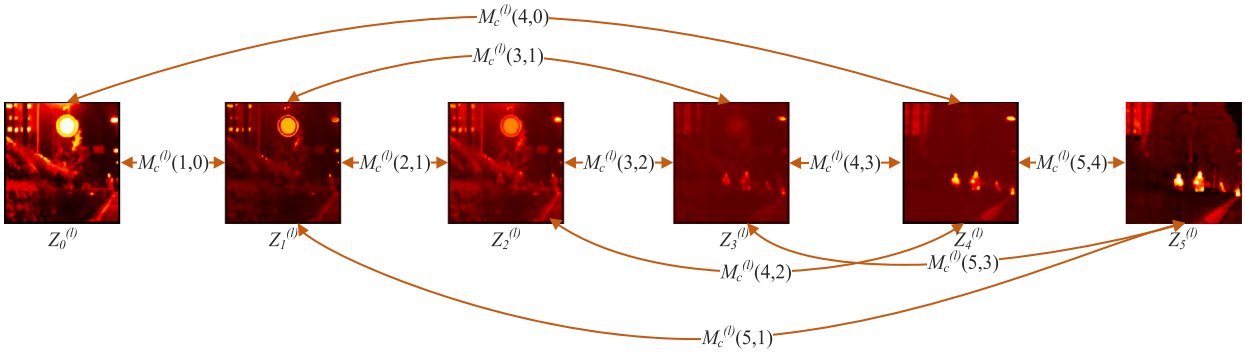

Fig. 6. Illustrative example of Fig. 5, where arrows between features represent distance constraints.

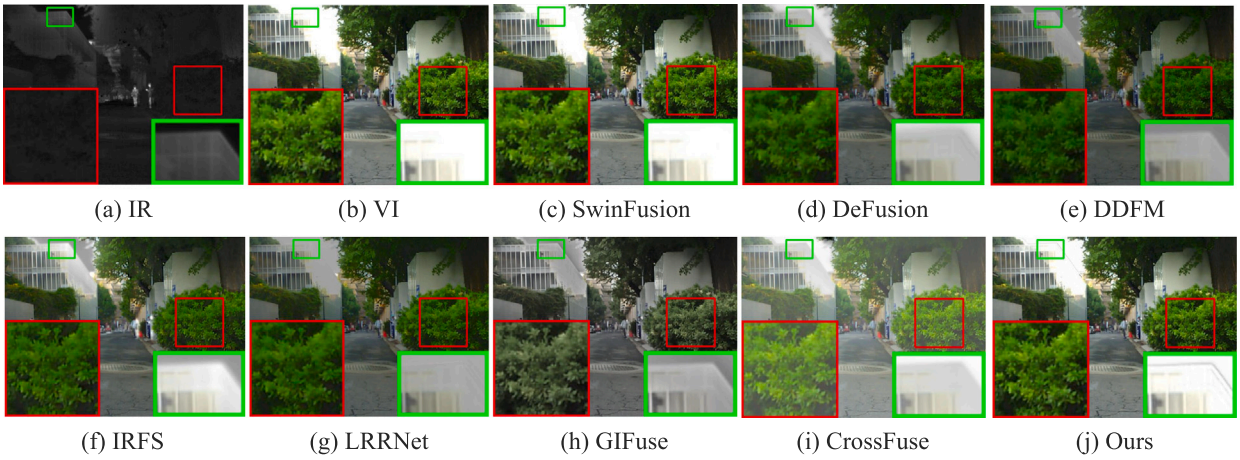

Fig. 7. Qualitative comparison of image "00327" in MSRS.

Table 1
Quantitative comparison on MSRS(361 image pairs). The bolded and underlined values represent the best and the second best results.

Methods	Year	MI	SF	AG	VIF	Qabf	LIQE	TOPIQ
SwinFusion [33]	2022	<u>4.529</u>	<u>11.089</u>	<u>3.566</u>	<u>0.99</u>	<u>0.654</u>	1.048	0.231
DeFusion [9]	2022	3.345	8.606	2.781	0.763	0.530	1.025	0.214
DDFM [34]	2023	2.728	7.388	2.522	0.743	0.474	1.025	0.217
IRFS [35]	2023	2.151	9.888	3.155	0.735	0.477	1.020	0.224
LRRNet [12]	2023	1.108	4.162	0.966	0.194	0.103	1.036	0.229
GIFuse [36]	2024	2.409	10.425	3.310	0.857	0.625	**1.081**	<u>0.239</u>
CrossFuse [37]	2024	3.124	9.621	3.006	0.836	0.560	1.051	0.232
Conti-Fuse	Ours	**5.457**	**11.478**	**3.718**	**1.040**	**0.710**	<u>1.067</u>	**0.245**

To enhance the scalability of the proposed fusion model and ensure proper decomposition, we propose the Support Decomposition Strategy (SDS) to compute L_{decom}. Instead of constraining the distances between all features for each decomposition loss calculation, we only constrain a subset.

Specifically, we constrain the distances between adjacent features and add some randomly sampled distance constraints. For instance, as illustrated in Figs. 5 and 6, when the number of transition states is set to 4, we constrain the distances between all adjacent features (indicated in red) and randomly sample some distances from the remaining ones for constraint (indicated in yellow). Statistically, with sufficient training epochs and ample training data, this partial distance constraint approach approximates the desired decomposition method, thereby reducing the complexity of calculating the decomposition loss.

For a more general case with K transition states, at the lth layer, we define the sampling strategy as follows,

$$\begin{aligned} SDS(\beta, K) = &\{(i+1, i)\}_{i=0}^{K} \cup \\ &\{(u_i, v_i) | (u_i, v_i) \leftarrow \beta\}_{i=0}^{K} \\ s.t. \quad &u_i \neq K+1, v_i \neq 0, u_i + 1 \neq v_i, \\ &u_i > v_i, u_i, v_i \in \{0, 1, \ldots, K+1\} \end{aligned} \tag{11}$$

where $SDS(\cdot, \cdot)$ represents the set of constraints sampled at the lth layer, $\{(u_i, v_i)\}_{i=0}^{K}$ is a set of non-repetitive ordered pairs sampled randomly and (u_i, v_i) denotes the ith sampled unique ordered pair. β is a random seed. Clearly, $|SDS(\beta, K)| = 2K + 2$.

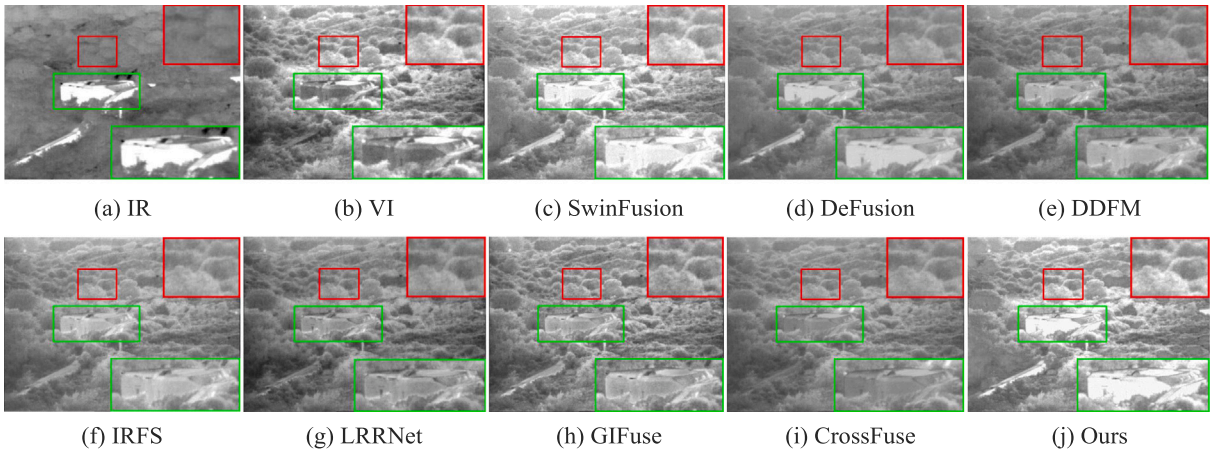

Fig. 8. Qualitative comparison of image "9" in TNO.

Table 2
Quantitative comparison on TNO(21 image pairs). The bolded and underlined values represent the best and the second best results.

Methods	Year	MI	SF	AG	VIF	Qabf	LIQE	TOPIQ
SwinFusion [33]	2022	3.208	<u>10.283</u>	**3.928**	0.708	<u>0.522</u>	1.009	0.225
DeFusion [9]	2022	2.530	5.652	2.298	0.509	0.332	1.013	0.213
DDFM [34]	2023	1.605	7.852	3.183	0.258	0.227	1.012	0.201
IRFS [35]	2023	2.094	8.347	2.950	0.563	0.416	1.013	0.223
LRRNet [12]	2023	<u>3.945</u>	6.471	2.601	0.727	0.314	1.019	0.223
GIFuse [36]	2024	1.502	9.916	3.706	0.35	0.345	<u>1.022</u>	<u>0.246</u>
CrossFuse [37]	2024	3.124	9.936	3.701	<u>0.751</u>	0.453	1.018	0.251
Conti-Fuse	Ours	**4.539**	**10.479**	<u>3.773</u>	**0.801**	**0.553**	**1.031**	**0.289**

Based on the sampled constraints, we can rewrite the decomposition loss as follows,

$$L_{decom} = \frac{1}{N(2K+2)} \sum_{l=1}^{N} \sum_{(i,j)\in SDS(*,K)} [M_c^{(l)}(i,j) - M_t^{(l)}(i,j)]^2 \tag{12}$$

where $*$ denotes the random seed determined by the operating system in our training processing, which is used to ensure that the sampling results are as diverse as possible.

Evidently, by using SDS, we control the time complexity of the decomposition loss to $O(K)$, thereby significantly enhance the scalability of our proposed model.

4. Experiments

4.1. Setup

4.1.1. Implementation details

The number of blocks (Encoder and Decoder) and transition states in our model is set to $N = 3$ and $K = 7$. The model width is configured to 8, which corresponds to the number of channels obtained from the linear layer mapping the input source image. Each layer in the CDM contains one State Transformer, and the number of heads in the TSWM is set to 4. We employ average pooling for downsampling and bilinear interpolation for upsampling. For model training, training images are randomly cropped to 192×192, with random flipping being the only data augmentation technique used. The batch size and number of epochs are set to 20 and 250, respectively. To mitigate potential instability during training, we implement gradient clipping to prevent the occurrence of gradient explosion. AdamW [38] is utilized as the optimizer, and WarmupCosine serves as the learning rate adjustment strategy. We gradually increase the learning rate from 10^{-5} to 6×10^{-5} during the first 50 epochs, and subsequently, it is gradually decayed to 5×10^{-6} over the remaining epochs. The proposed Gaussian decay function is employed as the decay strategy (Eq. (9)) to compute the decomposition loss, with hyperparameters α_1 and α_2 both set to 15. Our code is implemented using the PyTorch framework, and all experiments are conducted on a NVIDIA GeForce RTX 3090 Ti.

4.1.2. Datasets and metrics

Our model is trained on the training set of MSRS [39]. For the test set, the testing images on M3FD [40] and TNO are used in CrossFuse [37]. The testing images on MSRS [39] are provided by the dataset.

Furthermore, we evaluate the quality of fused images from both reference and non-reference perspectives using seven objective metrics,

- Mutual Information (MI) is employed to assess the amount of information retained in the fused image from the two source images.
- Spatial Frequency (SF) [41] is used to evaluate the sharpness of the fused image.
- Average Gradient (AG) measures the richness of texture details in the fused image.
- Visual Information Fidelity (VIF) [42] quantifies the preservation of visual information between the fused image and the two source images.
- Qabf [43] assesses the representation of salient information in the fused image.
- LIQE [44] employs the image-language model to evaluate image quality, with higher values indicating better quality.
- TOPIQ [45] utilizes the attention mechanism to assess the levels of distortion and noise in an image, with higher values indicating better quality.

Higher values of these five metrics indicate better quality of the fused image.

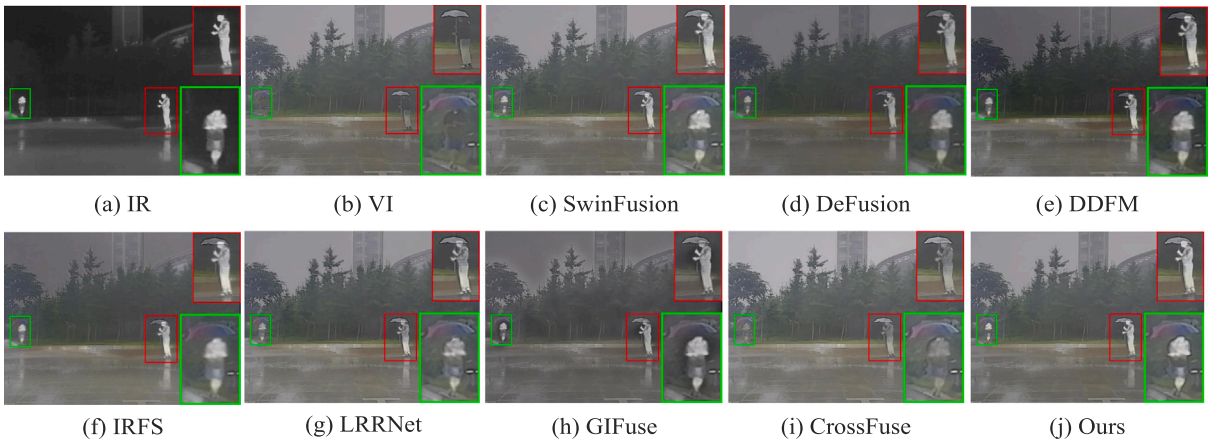

Fig. 9. Qualitative comparison of image "00274" in M3FD.

Table 3
Quantitative comparison on M3FD(300 image pairs). The bolded and underlined values represent the best and the second best results.

Methods	Year	MI	SF	AG	VIF	Qabf	LIQE	TOPIQ
SwinFusion [33]	2022	3.907	10.869	3.363	0.897	0.621	1.399	0.363
DeFusion [9]	2022	2.767	6.107	1.976	0.611	0.356	1.233	0.325
DDFM [34]	2023	2.551	7.161	2.219	0.586	0.374	1.227	0.336
IRFS [35]	2023	2.572	8.469	2.578	0.733	0.510	1.316	0.357
LRRNet [12]	2023	4.801	5.384	1.557	0.682	0.227	1.416	0.354
GIFuse [36]	2024	2.186	11.203	3.442	**0.913**	**0.658**	1.422	**0.385**
CrossFuse [37]	2024	3.476	9.851	2.964	0.815	0.564	1.391	0.376
Conti-Fuse	Ours	**4.919**	**11.850**	**3.475**	0.902	0.638	**1.423**	0.377

Table 4
Quantitative of the multi-modality semantic segmentation task on MSRS [39]. The bolded and underlined values represent the best and the second best results.

	IR	VI	DDFM	GIFuse	IRFS	DeFusion	LRRNet	CrossFuse	SwinFusion	Conti-Fuse
mIoU	62.58	59.78	65.96	66.78	67.00	66.59	63.99	67.35	65.88	**67.80**

4.2. Comparison with other methods

The proposed model is compared with seven state-of-the-art fusion methods, including one diffusion-based method (DDFM [34]), two decomposition-based methods (DeFusion [9] and LRRNet [12]), one downstream task-integrated method (IRFS [35]), one unified fusion method (GIFuse [36]), and two Transformer-based methods (CrossFuse [37] and SwinFusion [33]). The implementations of these approaches are publicly available.

4.2.1. Qualitative comparison

Fig. 7, Fig. 8, and Fig. 9 present the qualitative comparison results on MSRS, TNO and M3FD, respectively. It can be observed that our method significantly highlights salient targets and retains more detailed information.

In Fig. 7, our method produces more vibrant colors and sharper edges of buildings. In contrast, DDFM, GIFuse, IRFS, DeFusion, and LRRNet exhibit color shifts, CrossFuse results in blurry images, and SwinFusion loses the edge details of buildings. In Fig. 8, our method shows more prominent salient targets (highlighted in the green box) compared to all methods except SwinFusion.

Additionally, compared to SwinFusion, our method preserves more texture details and provides sharper edges of salient targets. Similarly, in Fig. 9, our method demonstrates more pronounced salient targets and clearer edges of the figures (green box) compared to other methods. Other methods, such as SwinFusion, DDFM, and IRFS, produce blurrier edges of salient targets. This indicates that our method retains more critical information, demonstrating its effectiveness.

4.2.2. Quantitative comparison

Table 1, Table 2, and Table 3 present the results of the qualitative comparisons on MSRS, TNO and M3FD, respectively. Compared to other methods, our approach consistently achieves superior results across all three datasets. The higher AG and SF values indicate that our method retains more texture information and produces clearer images.

Additionally, our method demonstrates higher VIF, LIQE and Qabf scores, suggesting that the fused images align better with human visual perception and contain more salient information. Finally, the higher MI score indicates that our method preserves more information from the source images.

It is noteworthy that our method is trained only on the MSRS dataset and is not fine-tuned on other datasets, which demonstrates its strong generalization capability.

4.2.3. Multi-modality semantic segmentation

The multi-modality semantic segmentation task is applied as a high-level visual task to further validate the effectiveness of our approach. Specifically, the fused results from all methods are divided into identical training, testing, and validation sets, and the pretrained B2 model of Segformer [46] is fine-tuned on the fused results of each method. Finally, qualitative and quantitative evaluations are conducted on the respective test sets.

All methods are fine-tuned using identical experimental parameters, including the same random seed. We utilize AdamW as the optimizer, set the learning rate to 10^{-5}, and configure the batch size to 12, without employing any data augmentation techniques. Each method is fine-tuned for 50 epochs.

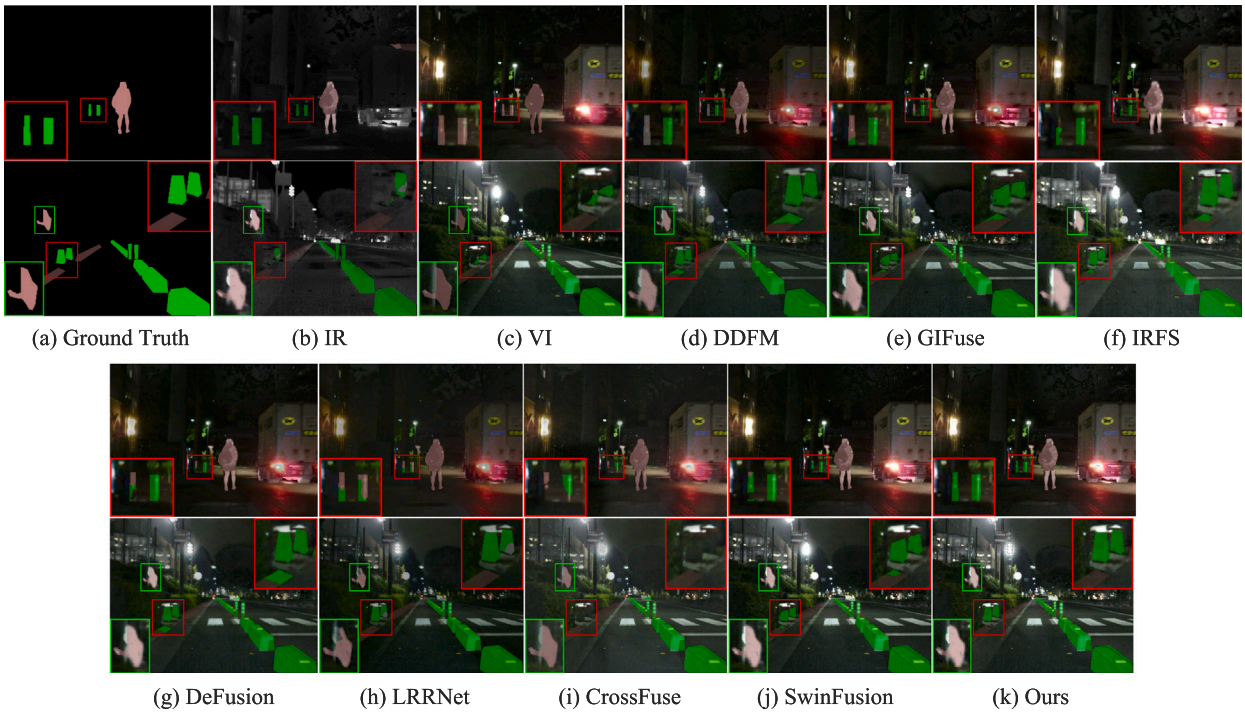

Fig. 10. Qualitative evaluation of the multi-modality semantic segmentation task on the MSRS [39] dataset.

The multi-modality semantic segmentation experiments are conducted on the MSRS [39] dataset. The fused results of all methods on the training set are randomly partitioned into new training and validation sets with a 9:1 ratio. The testing set is provided by MSRS [39]. The mIoU metric is employed for qualitative evaluation of the segmentation results.

As shown in Fig. 10, our method achieves outstanding results in the multimodal object segmentation task. By employing sequential decomposition to obtain multiple intermediate states and thereby mitigating the loss of critical information, our fused results exhibit clearer edges, which leads to superior segmentation performance.

For instance, our method excels in segmenting the Bump (highlighted in green), a particularly subtle object, outperforming other methods such as SwinFusion, CrossFuse, and DDFM, which struggle with incomplete segmentation due to unclear edges.

Similarly, as demonstrated in Table 4, our method also demonstrates strong performance in quantitative comparisons. This evidence confirms that our approach effectively retains key information from the source images, benefiting high-level downstream tasks.

5. Ablation studies

In this section, the ablation studies are conducted to verify the rationality of module design, the effectiveness of decomposition loss, and the appropriateness of parameter selection.

5.1. The number of transition states

To determine the most suitable number of transition states K, we conducted ablation experiments. Starting from $K = 3$, we incremented the number of transition states by two in each subsequent experiment.

As shown in Table 5, it can be observed that as K increases, the model's performance gradually improves. However, when K reaches around 7, the performance gains slow down and show a tendency to decline. Therefore, considering the overall performance of the model, $K = 7$ is reasonable.

The first row of Fig. 11 presents a qualitative comparison for different values of K. It can be seen that when K is small, the model lacks sufficient representational capacity, leading to significant loss of detail information. Conversely, when K becomes excessively large and exceeds the suitable representation range, unnecessary redundant information may be overrepresented (for instance, harmful background information from the visible image), while the originally salient information may become diluted by the excess of transition states, leading to the potential loss of some important information.

Table 5
The ablation experiment of the number of transition states K. The bolded values represent the best results.

	MI	SF	AG	VIF	Qabf	LIQE	TOPIQ
K=3	5.252	11.474	3.710	1.029	0.706	1.064	0.236
K=5	5.374	11.464	3.682	1.039	0.708	1.062	0.238
K=7	**5.457**	**11.478**	**3.718**	**1.040**	**0.710**	**1.067**	**0.245**
K=9	5.437	11.459	3.703	1.021	0.703	1.060	0.241
K=11	5.423	11.447	3.697	1.038	0.705	1.061	0.239

5.2. Visualization of transition states

Fig. 12 presents the visualization results related to transition states. We visualize the transition state features $\{Z_1^{(1)}, Z_2^{(1)}, \ldots, Z_7^{(1)}\}$ of image "00004N" from the MSRS dataset, as well as the features of the visible and infrared images $\{Z_0^{(1)}, Z_8^{(1)}\}$ according to Eq. (3).

From left to right and top to bottom, the feature maps exhibit a continuous changing trend, consistent with our previous discussion. Clearly, these decomposed states are not mutually exclusive. There are overlapping parts and unique parts among them, with each transition state retaining some critical information from the source images.

For instance, $Z_4^{(1)}$ preserves low-frequency information from both source images, while $Z_5^{(1)}$ retains more high-frequency information from the infrared image. These transition states, which encapsulate rich information from the source images, contribute to retaining more important information from the source images.

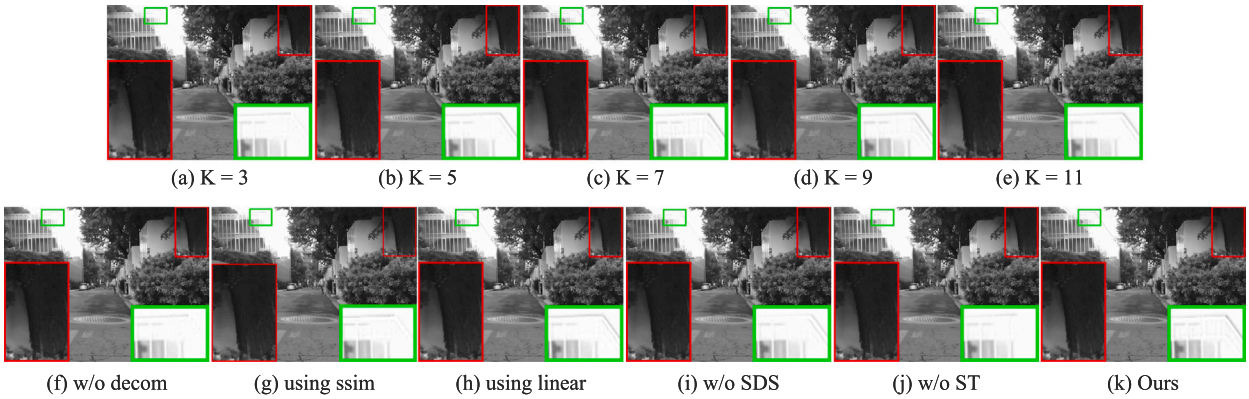

Fig. 11. Visualization of ablation experiments in MSRS "00327D".

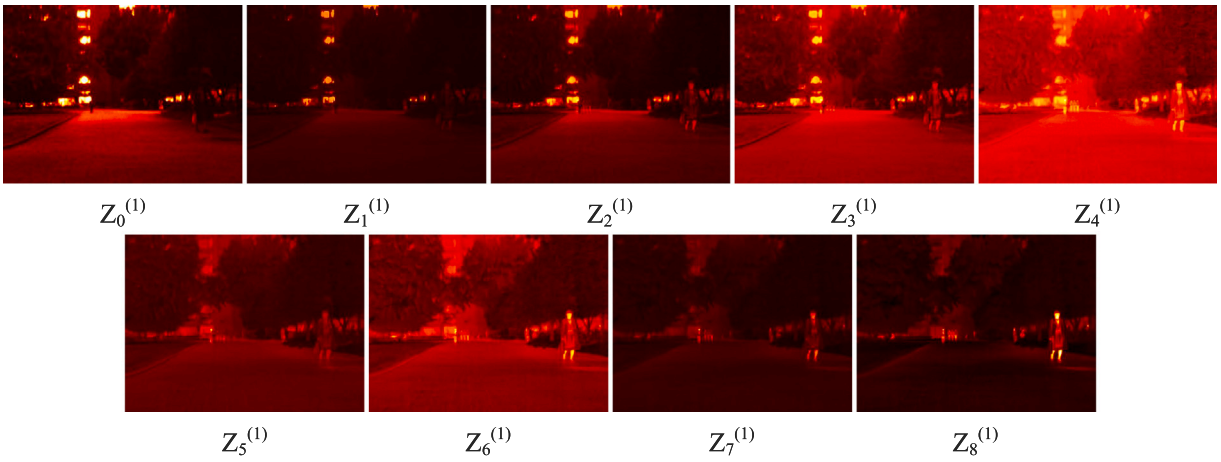

Fig. 12. Visualization of transition states and two source images' feature maps in 1-th layer.

Table 6
The ablation experiments of other factors. The bolded values represent the best results.

	MI	SF	AG	VIF	Qabf	LIQE	TOPIQ	FLOPs↓	Memory↓
w/o decom	5.337	11.432	3.707	1.038	0.706	1.060	0.238	–	–
using linear	5.442	11.248	3.581	1.030	0.704	1.061	0.240	–	–
using ssim	5.412	11.473	3.702	1.044	0.709	1.064	0.241	–	–
w/o SDS	5.455	11.472	3.714	1.039	0.701	1.065	0.244	7178M	19452M
w/o ST	5.316	11.476	3.701	1.036	0.708	1.063	0.239	–	–
Ours	**5.457**	**11.478**	**3.718**	**1.040**	**0.710**	**1.067**	**0.245**	**2871M**	**14356M**

5.3. Other factors

The ablation experiments on other influencing factors are shown in Table 6 and Fig. 11.

5.3.1. Decomposition loss L_{decom}

Firstly, we conduct an ablation study on the proposed loss function by removing the decomposition loss L_{decom} while keeping other conditions constant. The results, presented in the first row of the table and the "w/o decom" plot, indicate that the absence of the decomposition constraint leads to a significant loss of detail in the fused images, resulting in decreased model performance.

5.3.2. Distance metric

We experimented with using another common metric for measuring image distance, SSIM, as a replacement for the Pearson correlation coefficient. The results are shown in the second row of Table 6 and the "using ssim" plot in Fig. 11. As can be observed, image quality deteriorates when using SSIM. SSIM emphasizes structural differences within images, whereas CC considers differences based on pixel values. Measuring structural differences between deep features using conventional structural metrics for images is somewhat inappropriate; instead, directly measuring value differences between features in deep representations aligns better with intuition.

5.3.3. Decay function of L_{decom}

Secondly, we replaced the decay function with a linear one. The linear decay function applies an arithmetic decay with a fixed step size along the sub-diagonal of the distance matrix M_c:

$$\Omega_l(p, \mu, s) = 1 - p\frac{1-\mu}{s-1} \tag{13}$$

where Ω_l represents the linear decay function.

The outcomes, shown in the third row of the Table 6 and the "using linear" plot, reveal that the window region within the green box becomes blurred, leading to a decline in image quality. We hypothesize that this is due to the linear attenuation being too gradual compared to Gaussian attenuation, failing to effectively separate low-frequency and high-frequency information.

5.3.4. SDS strategy

To verify the effectiveness of the SDS strategy, we removed the SDS component. The results, illustrated in the fourth row of the table and the "w/o SDS" plot, clearly show a significant increase in the computational load of the loss function, with a quadratic growth trend.

It can be observed that without SDS, both the computational cost and memory usage during training significantly increase. It should be noted that the computational cost mentioned here only includes the calculation of L_{decom}. Additionally, we surprisingly found that the SDS can enhance the model's ability to preserve edge textures. This may be because the randomness introduced by SDS increases the model's robustness.

5.3.5. State transformer

Finally, to validate the effectiveness of the State Transformer, we replaced it with several convolution operations while maintaining a similar number of parameters.

The results, as shown in the fifth row of the table and the "w/o ST" plot, demonstrate that without the State Transformer, the model fails to effectively capture complementary information between transitional states, resulting in a substantial loss of critical information.

6. Conclusion

In this paper, a novel fusion framework (Conti-Fuse) based on the designed continuous decomposition strategy is proposed. Conti-Fuse densely samples the trajectory in the unified high-dimensional feature space to decompose the source deep features into multiple transition states, thereby mitigating the loss of critical information from the source input. Additionally, the CDM (including the State Transformer) is introduced to leverage its powerful feature interaction capability, capturing complementary information between transition states and enabling continuous decomposition. Finally, to drive the process of continuous decomposition, a novel and efficient loss function L_{decom} is proposed.

Both qualitative and quantitative comparisons of our proposed method with seven state-of-the-art (SOTA) approaches from the past three years are conducted. In quantitative evaluations, Conti-Fuse achieved best or second-best results across most metrics, demonstrating the superiority of our approach. Visually, our method effectively preserves the salient information of the source images while capturing fine details, thereby mitigating the loss of important information. In the multimodal semantic segmentation task, Conti-Fuse's fusion results retain more edge information from the source images, which enhances segmentation performance compared to the above methods.

CRediT authorship contribution statement

Hui Li: Writing – review & editing, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. **Haolong Ma:** Writing – original draft, Methodology, Formal analysis, Conceptualization. **Chunyang Cheng:** Writing – review & editing, Investigation, Formal analysis. **Zhongwei Shen:** Writing – review & editing, Visualization, Validation, Resources. **Xiaoning Song:** Visualization, Validation, Formal analysis, Data curation. **Xiao-Jun Wu:** Writing – review & editing, Visualization, Validation.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Hui Li reports financial support was provided by National Natural Science Foundation of China. Xiao-Jun Wu reports was provided by National Key Research and Development Program of China. Hui Li reports was provided by Fundamental Research Funds for the Central Universities. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (62202205, 62306203, 62332008, 62020106012), National Key Research and Development Program of China (2023YFF1105100, 2023YFE0116300), the National Social Science Foundation of China (21&ZD166), the Natural Science Foundation of Jiangsu Province, China (BK20221535), and the Fundamental Research Funds for the Central Universities (JUSRP123030).

Data availability

Data will be made available on request.

References

[1] R. Xu, Z. Xiao, M. Yao, Y. Zhang, Z. Xiong, Stereo video super-resolution via exploiting view-temporal correlations, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 460–468.
[2] Z. Yang, M. Yao, J. Huang, M. Zhou, F. Zhao, Sir-former: Stereo image restoration using transformer, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 6377–6385.
[3] W. Zhao, S. Xie, F. Zhao, Y. He, H. Lu, Metafusion: Infrared and visible image fusion via meta-feature embedding from object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 13955–13965.
[4] H. Zhang, H. Xu, X. Tian, J. Jiang, J. Ma, Image fusion meets deep learning: A survey and perspective, Inf. Fusion 76 (2021) 323–336.
[5] L. Tang, X. Xiang, H. Zhang, M. Gong, J. Ma, DIVFusion: Darkness-free infrared and visible image fusion, Inf. Fusion 91 (2023) 477–493.
[6] J. Liu, R. Dian, S. Li, H. Liu, SGFusion: A saliency guided deep-learning framework for pixel-level image fusion, Inf. Fusion 91 (2023) 205–214.
[7] H. Li, X.-J. Wu, J. Kittler, RFN-Nest: An end-to-end residual fusion network for infrared and visible images, Inf. Fusion 73 (2021) 72–86.
[8] H. Hermessi, O. Mourali, E. Zagrouba, Multimodal medical image fusion review: Theoretical background and recent advances, Signal Process. 183 (2021) 108036.
[9] P. Liang, J. Jiang, X. Liu, J. Ma, Fusion from decomposition: A self-supervised decomposition approach for image fusion, in: European Conference on Computer Vision, Springer, 2022, pp. 719–735.
[10] Y. Lu, F. Wang, X. Luo, F. Liu, Novel infrared and visible image fusion method based on independent component analysis, Front. Comput. Sci. 8 (2014) 243–254.
[11] S. Li, X. Kang, L. Fang, J. Hu, H. Yin, Pixel-level image fusion: A survey of the state of the art, Inf. Fusion 33 (2017) 100–112.
[12] C. Li, B. Zhang, D. Hong, J. Yao, J. Chanussot, LRR-Net: An interpretable deep unfolding network for hyperspectral anomaly detection, IEEE Trans. Geosci. Remote Sens. (2023).
[13] Z. Zhao, S. Xu, C. Zhang, J. Liu, P. Li, J. Zhang, DIDFuse: Deep image decomposition for infrared and visible image fusion, 2020, arXiv preprint arXiv:2003.09210.
[14] G. Pajares, J.M. De La Cruz, A wavelet-based image fusion tutorial, Pattern Recognit. 37 (9) (2004) 1855–1872.
[15] J. Shen, Y. Zhao, S. Yan, X. Li, et al., Exposure fusion using boosting Laplacian pyramid, IEEE Trans. Cybern. 44 (9) (2014) 1579–1590.
[16] H. Zhang, J. Ma, SDNet: A versatile squeeze-and-decomposition network for real-time image fusion, Int. J. Comput. Vis. 129 (10) (2021) 2761–2785.
[17] H. Li, X.-J. Wu, J. Kittler, MDLatLRR: A novel decomposition method for infrared and visible image fusion, IEEE Trans. Image Process. 29 (2020) 4733–4746.
[18] H. Tang, G. Liu, L. Tang, D.P. Bavirisetti, J. Wang, MdedFusion: A multi-level detail enhancement decomposition method for infrared and visible image fusion, Infrared Phys. Technol. 127 (2022) 104435.
[19] J. Nunez, X. Otazu, O. Fors, A. Prades, V. Pala, R. Arbiol, Multiresolution-based image fusion with additive wavelet decomposition, IEEE Trans. Geosci. Remote Sens. 37 (3) (1999) 1204–1211.
[20] C. Pohl, J.L. Van Genderen, Review article multisensor image fusion in remote sensing: concepts, methods and applications, Int. J. Remote Sens. 19 (5) (1998) 823–854.
[21] M.K. Panda, P. Parida, D.K. Rout, A weight induced contrast map for infrared and visible image fusion, Comput. Electr. Eng. 117 (2024) 109256.
[22] M.K. Panda, B.N. Subudhi, T. Veerakumar, M.S. Gaur, Edge preserving image fusion using intensity variation approach, in: 2020 IEEE Region 10 Conference, TENCON, IEEE, 2020, pp. 251–256.
[23] M.K. Panda, B.N. Subudhi, T. Veerakumar, M.S. Gaur, Pixel-level visual and thermal images fusion using maximum and minimum value selection strategy, in: 2020 IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security, ISSSC, IEEE, 2020, pp. 1–6.

[24] M.K. Panda, V. Thangaraj, B.N. Subudhi, V. Jakhetiya, Bayesian's probabilistic strategy for feature fusion from visible and infrared images, Vis. Comput. 40 (6) (2024) 4221–4233.
[25] W. Tang, F. He, Y. Liu, Y. Duan, T. Si, DATFuse: Infrared and visible image fusion via dual attention transformer, IEEE Trans. Circuits Syst. Video Technol. (2023).
[26] Z. Zhou, E. Fei, L. Miao, R. Yang, A perceptual framework for infrared–visible image fusion based on multiscale structure decomposition and biological vision, Inf. Fusion 93 (2023) 174–191.
[27] Y. Li, G. Liu, D.P. Bavirisetti, X. Gu, X. Zhou, Infrared-visible image fusion method based on sparse and prior joint saliency detection and LatLRR-FPDE, Digit. Signal Process. 134 (2023) 103910.
[28] M.K. Panda, B.N. Subudhi, T. Veerakumar, V. Jakhetiya, Integration of bi-dimensional empirical mode decomposition with two streams deep learning network for infrared and visible image fusion, in: 2022 30th European Signal Processing Conference, EUSIPCO, IEEE, 2022, pp. 493–497.
[29] H. Li, X. Qi, W. Xie, Fast infrared and visible image fusion with structural decomposition, Knowl.-Based Syst. 204 (2020) 106182.
[30] X. Zhang, Y. Demiris, Visible and infrared image fusion using deep learning, IEEE Trans. Pattern Anal. Mach. Intell. (2023).
[31] P. Liang, J. Jiang, Q. Ma, X. Liu, J. Ma, Fusion from decomposition: A self-supervised approach for image fusion and beyond, 2024, arXiv preprint arXiv: 2410.12274.
[32] S.W. Zamir, A. Arora, S. Khan, M. Hayat, F.S. Khan, M.-H. Yang, Restormer: Efficient transformer for high-resolution image restoration, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5728–5739.
[33] J. Ma, L. Tang, F. Fan, J. Huang, X. Mei, Y. Ma, SwinFusion: Cross-domain long-range learning for general image fusion via swin transformer, IEEE/CAA J. Autom. Sin. 9 (7) (2022) 1200–1217.
[34] Z. Zhao, H. Bai, Y. Zhu, J. Zhang, S. Xu, Y. Zhang, K. Zhang, D. Meng, R. Timofte, L. Van Gool, DDFM: denoising diffusion model for multi-modality image fusion, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 8082–8093.
[35] W. Di, L. Jinyuan, R. Liu, F. Xin, An interactively reinforced paradigm for joint infrared-visible image fusion and saliency object detection, Inf. Fusion 98 (2023) 101828.
[36] W. Wang, L.-J. Deng, G. Vivone, A general image fusion framework using multi-task semi-supervised learning, Inf. Fusion 108 (2024) 102414.
[37] H. Li, X.-J. Wu, CrossFuse: A novel cross attention mechanism based infrared and visible image fusion approach, Inf. Fusion 103 (2024) 102147.
[38] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: International Conference on Learning Representations, 2019.
[39] L. Tang, J. Yuan, H. Zhang, X. Jiang, J. Ma, PIAFusion: A progressive infrared and visible image fusion network based on illumination aware, Inf. Fusion 83 (2022) 79–92.
[40] J. Liu, X. Fan, Z. Huang, G. Wu, R. Liu, W. Zhong, Z. Luo, 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, 2022, pp. 5802–5811.
[41] P.F. A.M. Eskicioglu, Image quality measures and their performance, IEEE Trans. Commun. 43 (1995) 2959–2965.
[42] Y. Han, Y. Cai, Y. Cao, X. Xu, A new image fusion performance metric based on visual information fidelity, Inf. Fusion 14 (2) (2013) 127–135.
[43] C.S. Xydeas, V. Petrovic, et al., Objective image fusion performance measure, Electron. Lett. 36 (4) (2000) 308–309.
[44] W. Zhang, G. Zhai, Y. Wei, X. Yang, K. Ma, Blind image quality assessment via vision-language correspondence: A multitask learning perspective, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 14071–14081.
[45] C. Chen, J. Mo, J. Hou, H. Wu, L. Liao, W. Sun, Q. Yan, W. Lin, Topiq: A top-down approach from semantics to distortions for image quality assessment, IEEE Trans. Image Process. (2024).
[46] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J.M. Alvarez, P. Luo, SegFormer: Simple and efficient design for semantic segmentation with transformers, in: Neural Information Processing Systems (NeurIPS), 2021.