

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository: <https://orca.cardiff.ac.uk/id/eprint/189032/>

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Zhao, Cheng, Song, Tianyun, Wu, Zhiliang, Wang, Tianfu, Gabbouj, Moncef, Yue, Guanghui, Lei, Baiying and Zhou, Wei 2026. Uncertainty-guided spatiotemporal consistency fusion network for infrared-visible video fusion under extremely low-light conditions. IEEE Transactions on Image Processing 35 , pp. 8894-8909. 10.1109/tip.2026.3719477

Publishers page: <https://doi.org/10.1109/tip.2026.3719477>

Please note:

Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See <http://orca.cf.ac.uk/policies.html> for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.



# Uncertainty-Guided Spatiotemporal Consistency Fusion Network for Infrared-Visible Video Fusion Under Extremely Low-Light Conditions

Cheng Zhao<sup>✉</sup>, Tianyun Song, Zhiliang Wu<sup>✉</sup>, *Member, IEEE*, Tianfu Wang<sup>✉</sup>, Moncef Gabbouj<sup>✉</sup>, *Fellow, IEEE*, Guanghui Yue<sup>✉</sup>, *Senior Member, IEEE*, Baiying Lei<sup>✉</sup>, *Senior Member, IEEE*, and Wei Zhou<sup>✉</sup>, *Senior Member, IEEE*

**Abstract**—Infrared-visible video fusion under extremely low-light conditions is critically important yet remains underexplored, largely due to the scarcity of high-quality datasets and challenges posed by spatiotemporal uncertainty and modality bias. To address the dataset shortage, we built a dataset of 4,739 infrared and visible registration video pairs captured under extremely low-light conditions, spanning 5 scene types and 17 subcategories. Further, we proposed an Uncertainty-guided Spatiotemporal Consistency Fusion Network, termed USCFNet, for the infrared-visible video fusion. At each layer of the encoder, an Entropy-Gated SpatioTemporal Attention (EGSTA) module is introduced to capture temporal instability and spatial reliability variations through entropy-aware attention modulation, thereby enhancing feature spatiotemporal consistency. The refined infrared and visible features are then fused via a Difference-Guided Fusion (DGF) module, which adaptively exploits their content and edge differences to improve structural integrity and detail clarity. By progressively connecting DGF modules from shallow to deep layers, the network achieves the synergistic fusion of shallow textures and deep semantics. Subsequently, the output of the last DGF module is fused with the modality features of the last layer through a hierarchical

mixture-of-experts fusion module. This module enables the balanced integration of modality information while preserving fine local details. Finally, the fusion feature is fed into the decoder to produce the final fused video. Extensive experiments on our dataset and two public datasets show that USCFNet outperforms competing methods, achieving lower distortion and stronger spatiotemporal consistency. The source code and dataset are available at <https://github.com/Zhaocheng1/ELVID>

**Index Terms**—Infrared-visible video fusion, extremely low-light conditions, spatiotemporal consistency, uncertainty, mixture-of-experts.

## I. INTRODUCTION

IN A dual-spectrum imaging system, infrared and visible sensors observe the same scene based on distinct physical imaging mechanisms. The former shows the scene's energy distribution by capturing thermal radiation, while the latter captures the scene's texture and color details through the reflection of visible-band light [1]. Data captured by two sensors are complementary, and their effective fusion can provide reliable visual information for downstream tasks such as detection [2], [3], tracking [4], [5], and segmentation [6], [7]. Therefore, it is essential to propose effective methods for infrared and visible fusion.

Current research on infrared-visible fusion primarily focuses on normal lighting conditions [8], [9], with limited attention paid to extremely low-lighting conditions, such as nighttime video surveillance and search-and-rescue missions. One major obstacle to progress in this field is the scarcity of large-scale public datasets, which limits researchers' understanding of the challenges of such fusion tasks. Compared to normal illumination, visible data captured under extremely low-light conditions often suffers from severe degradation issues, such as insufficient brightness and loss of detail. While infrared data exhibits greater robustness in extremely low-light conditions, its inherent structural blurring and lack of detail are more pronounced [10]. These limitations result in a limited amount of information available for fusion tasks. For example, the first frame  $V_1$  of the visible video exhibits extremely low brightness and less information in the dark areas. The

Received 18 December 2025; revised 25 May 2026 and 7 July 2026; accepted 21 July 2026. Date of publication 7 August 2026; date of current version 19 August 2026. This work was supported in part by Shenzhen Medical Research Special Fund Project under Grant A2503036, Grant A2403035, and Grant C2401023; in part by the National Natural Science Foundation of China under Grant 62301329, Grant 62171312, Grant 62502447, and Grant 62071309; in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2024A1515030025; and in part by the Earth System Big Data Platform of the School of Earth Sciences, Zhejiang University. The associate editor coordinating the review of this article and approving it for publication was Dr. Tianyang Xu. (*Corresponding authors: Guanghui Yue; Baiying Lei.*)

Cheng Zhao, Tianyun Song, Tianfu Wang, Guanghui Yue, and Baiying Lei are with Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging, National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, School of Biomedical Engineering, Shenzhen University Medical School, Shenzhen University, Shenzhen 518060, China (e-mail: chengzhao@szu.edu.cn; 2400241038@mails.szu.edu.cn; tfwang@szu.edu.cn; yueguanghui@szu.edu.cn; leiby@szu.edu.cn).

Zhiliang Wu is with Zhejiang University, Hangzhou 310007, China (e-mail: wu\_zhiliang@zju.edu.cn).

Moncef Gabbouj is with the Faculty of Information Technology and Communication Sciences, Tampere University, 33720 Tampere, Finland (e-mail: moncef.gabbouj@tuni.fi).

Wei Zhou is with the School of Computer Science and Informatics, Cardiff University, CF10 3AT Cardiff, U.K. (e-mail: zhouw26@cardiff.ac.uk).

Digital Object Identifier 10.1109/TIP.2026.3719477

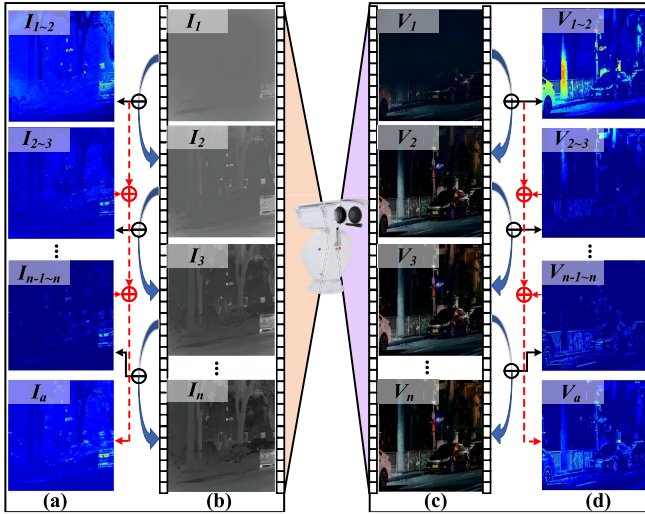


Fig. 1. Illustration of the infrared and visible inter-frame difference heatmaps under extremely low-light conditions. (a) The inter-frame difference heatmaps  $\{I_{t-1-t}\}_2^n$  and the average inter-frame difference heatmap  $I_a$  of the infrared video; (b) the infrared video frames  $\{I_t\}_1^n$ ; (c) the visible video frames  $\{V_t\}_1^n$ ; (d) the inter-frame difference heatmaps  $\{V_{t-1-t}\}_2^n$  and the average inter-frame difference heatmap  $V_a$  of the visible video.

infrared video  $\{I_t\}_1^n$  shows noticeable blurring, as shown in Figs. 1(b-c). On the other hand, there are also noticeable inter-frame brightness fluctuations in extremely low-light videos, which significantly reduce the temporal stability. The inter-frame difference heatmaps of the infrared and visible videos in Figs. 1(a) and 1(d) show that adjacent frames exhibit significant differences in regions such as vehicles and utility poles. Further, the averaged difference heatmaps, obtained by accumulating and averaging all inter-frame difference maps, show that these variations are mainly concentrated in salient structural regions and areas with evident brightness changes, whereas responses in low-light background regions are relatively weak. This indicates that both infrared and visible video sequences exhibit varying degrees of temporal instability under extremely low-light conditions. Overall, infrared and visible videos captured under extremely low-light conditions exhibit both spatial degradation (weak illumination, missing details, and blurred structures) and temporal instability (significant inter-frame brightness fluctuations). This makes effective infrared and visible video fusion under extremely low-light conditions considerably more challenging than those under normal-light conditions.

Existing fusion methods for low-light conditions can be broadly divided into two categories. The first category comprises methods that sequentially apply the visible enhancement network and the fusion network to improve the brightness or contrast of visible and alleviate poor fusion performance caused by visible degradation [11], [12]. In the other categories, those methods embed a brightness enhancement unit within the fusion framework to improve the learning ability of visible features [1], [10], [13]. However, because dark regions suffer from low information utilization and reduced structural reliability, whereas normally illuminated regions retain higher structural reliability, differences in spatial feature reliability provide useful cues for estimating the confidence

distribution across regions. After that, the confidence-guided feature extraction strategy can be used to adaptively adjust regional feature contributions for recovering key structures and enhancing overall fusion quality.

In addition to spatial degradation, video fusion also suffers from temporal instability. If the static-image fusion model is applied frame by frame to video sequences, the model cannot explicitly capture cross-frame dependencies. As a result, the fused video often exhibits temporal inconsistencies, such as brightness flickering and loss of frame-to-frame detail. Existing video fusion studies typically use only cross-attention mechanisms [14], [15], [16], [17] between adjacent frames (e.g., frame  $t$  and frame  $t-1$  or  $t+1$ ) to model very limited spatiotemporal relationships [4], [18]. Furthermore, these methods typically assume that all frames contribute equally and ignore temporal instability. This inevitably leads to poor fusion results and misleads subsequent spatiotemporal feature modeling due to the unexpected high weight of low-quality frames. Previous studies found that uncertainty estimation in visual tasks can reflect the model's confidence in different regions and adaptively emphasize high-confidence regions while suppressing low-confidence regions [19], [20]. Introducing this mechanism into video fusion tasks can be considered an effective way to achieve more robust feature selection in both temporal and spatial dimensions, thereby effectively improving the spatiotemporal consistency of the fusion results.

Furthermore, infrared and visible features differ significantly in semantic representation and detail expression, which can easily lead to modality learning imbalance during the fusion process. This imbalance makes it challenging to simultaneously ensure energy consistency and detail fidelity during the fusion process. The Mixture-of-Experts (MoE) model introduces multiple expert networks with complementary capabilities and dynamically selects the most suitable expert based on input features via a gating mechanism [21], [22], [23]. It enables differentiated modeling and dynamic fusion of modality-specific information, which has been proven to be an effective image fusion processing scheme [24], [25].

In this research, we first take up the challenge of data scarcity and construct a dataset for infrared-visible video fusion, named ELVID, comprising 4739 video pairs captured under extremely low-light conditions. We then propose an Uncertainty-guided Spatiotemporal Consistency Fusion Network (USCFNet) for infrared-visible video fusion under extremely low-light conditions. It follows an encoder-decoder framework and embeds an Entropy-Gated SpatioTemporal Attention (EGSTA) module in each encoder layer. It adaptively suppresses low-quality frames and low-confidence spatial regions to enhance the spatiotemporal consistency of modality-specific features. Then, the refined features from the two imaging modalities are fed into a Difference-Guided Fusion (DGF) module. The DGF module adjusts the contribution of each modality by leveraging content and edge differences to ensure effective feature fusion. Subsequently, the output of the DGF module at the last layer and the modality-specific features are fed into a hierarchical Mixture-of-Experts fusion (HMoE\_Fusion) block to generate the fusion feature. Finally,

the fusion feature is passed through the decoder to generate a high-quality fused video. Experimental results demonstrate that USCNet outperforms comparison methods on both our curated dataset and public datasets, and shows potential for downstream object detection tasks. The main contributions of this paper are summarized as follows:

- We constructed the ELVID dataset, the first infrared-visible video fusion dataset tailored to extremely low-light conditions, covering five scene types and 17 subcategories. It serves as a new benchmark for low-light video fusion.
- Considering the spatiotemporal uncertainties in extremely low-light infrared and visible video data, we propose an EGSTA unit to enhance the spatiotemporal consistency of modality-specific features. It suppresses the propagation of low-quality frames to enhance temporal consistency, while mitigating the influence of low-confidence regions to improve spatial consistency.
- To address the challenge of modality bias during learning and local detail degradation, we propose a HMoE\_Fusion module that synchronously models global modality complementary fusion and local fine-grained feature fusion.

## II. RELATED WORKS

### A. Infrared-Visible Fusion Method

With the continuous advancement of deep learning techniques, infrared-visible image fusion methods have achieved remarkable breakthroughs [26]. Early approaches extracted modality-specific features using CNN and combined them with traditional fusion rules to accomplish the fusion task [27], [28]. Subsequent studies introduced GANs to achieve end-to-end fusion [29], [30], [31]. At the same time, more recent works have incorporated Transformer [32], [33], [34], [35] and diffusion models [36], [37], [38], marking a transition from local feature fusion toward global semantic modeling. This trend has significantly improved the generative quality and perceptual naturalness of the fused image. However, most methods still rely on datasets collected under normal lighting conditions (e.g., TNO [39] and RoadScene [40]). With the emergence of low-light datasets such as LLVIP [41] and MSRS [42], research has begun to focus on infrared-visible image fusion under low-light conditions.

Considering the severe degradation of visible images in low-light environments, existing methods mostly employ two strategies: enhancement followed by fusion, or embedding a brightness enhancement module within the fusion network. For example, Chen et al. [43] jointly design a brightness enhancement network, a fusion network, and a brightness feedback network to improve fusion quality. Lv et al. [10] introduced a structure-aware mechanism to enhance the contributions of visible features. Hua et al. [13] improved low-light fusion performance with an illumination-aware module. Although these methods alleviate the degradation of visible images, they generally lack explicit modeling of spatial uncertainty, making it challenging to adaptively adjust feature contributions in critical structural regions.

Meanwhile, with the release of video fusion datasets such as M3SVD [18] and HDO [44], researchers have begun to

explore infrared-visible fusion at the video level. For example, Tang et al. [18] proposed a bidirectional cross-attention based temporal fusion framework that effectively captures inter-frame dependencies. Zhu et al. [4] extracted frame-wise features using CNNs and progressively fused adjacent frame features through a progressive fusion module. Although these methods alleviate the flickering problem in video fusion to some extent, their temporal modeling range remains limited. They generally assume equal contribution from all frames and lack explicit modeling of temporal uncertainty, making the fusion process vulnerable to low-quality frames. Recently, Han et al. [45] proposed the Seq-IF network, which enhances temporal consistency and structural fidelity in infrared and visible video fusion under complex illumination conditions by decoupling static and dynamic features and refining luminance consistency, further highlighting the importance of temporal modeling in video fusion. Based on the above analysis, we first construct an infrared and visible video fusion dataset for extremely low-light scenes to alleviate the problem of insufficient data. In terms of method design, this paper introduces the EGSTA module to encode multimodal video features under extremely low-light conditions. Specifically, the module enhances temporal consistency by suppressing the propagation of low-quality frame information, and improves spatial consistency by reducing interference from low-confidence regions in feature representations, thereby strengthening the spatiotemporal consistency of multimodal features.

### B. Uncertainty Estimation

Uncertainty in visual tasks includes aleatoric uncertainty and epistemic uncertainty [46]. Epistemic uncertainty reflects the model's judgment of the credibility of different regions. It can be used to adaptively focus attention on high-contribution regions during feature learning, thereby enabling more refined region-selective processing [19]. In recent years, uncertainty modeling has been increasingly incorporated into visual tasks to enhance model robustness and reliability. For instance, Zheng et al. [19] employed an uncertainty-aware convolution to fuse high-resolution panchromatic images with low-resolution multi-spectral images. Guo et al. [20] proposed a hybrid network that combines CNN and Transformer architectures and employs uncertainty estimation to identify unreliable features in medical images. The network dynamically allocates feature learning to minimize functional redundancy and improve representation efficiency. Hou et al. [47] developed a confidence-guided self-prediction optimization network to assess uncertainty in video action recognition, thereby improving the stability of action prediction in untrimmed videos. These studies suggest that uncertainty-based regulatory mechanisms can enhance information reliability. Motivated by this insight, we propose leveraging spatiotemporal uncertainty estimation to guide feature extraction and consistency modeling in low-light video fusion, thereby enabling the model to better handle spatial degradation and temporal instability under extremely low-light conditions. Here, uncertainty mainly refers to an entropy-based confidence measure for characterizing low-confidence or unstable feature states, rather than uncertainty in the strict probabilistic or bayesian sense.

### C. Mixture-of-Experts (MoE) Model

The MoE framework aims to achieve efficient feature processing and dynamic modeling through multi-expert collaboration and a gating-based selection mechanism [48]. It typically consists of multiple functionally complementary expert networks and a gating network. The gating module adaptively assigns weights to experts according to the input feature representations, enabling input-dependent conditional computation. Compared with conventional deep networks, the MoE framework enhances the model's representational capacity through a sparse expert selection mechanism while maintaining low computational cost. In recent years, MoE has demonstrated remarkable adaptability across various computer vision tasks. For example, Yang et al. [49] proposed a point-set metric-gated MoE model that automatically integrates complementary information from multiple source domains via MoE gating, thereby improving fault-diagnosis performance. Zhu et al. [50] constructed a cognitive perception-collaborative MoE network, where expert models simulate the collaboration among brain regions during cognitive tasks, thereby revealing the underlying pathological changes. Sun et al. [51] introduced an adaptive hybrid expert mechanism that dynamically integrates multimodal features for person re-identification. Liu et al. [52] proposed a weather-aware dynamic specificity and commonality collaborative mechanism based on an MoE framework, which adaptively allocates the optimal sub-network for specific weather types, enabling nighttime image restoration across multiple weather conditions. These studies indicate that MoE exhibits strong advantages in multi-source information fusion and heterogeneous feature modeling, enabling efficient information integration and cross-modal collaboration. However, existing MoE methods often employ the same expert network structure to uniformly model features from different modalities, which rarely incorporate specific designs to account for differences in information representation and structural details across modalities. Inspired by this, Xing et al. [53] proposed a semantic-guided multi-feature mixture-of-experts network that enhances learning of frequency- and spatial-domain features by assigning dedicated expert networks to different feature types, thereby achieving infrared and visible image fusion. This method focuses more on modeling distinct feature categories, whereas the representational differences and complementary characteristics across modalities have not been fully considered. Therefore, this paper proposes an HMoE module that introduces an expert routing mechanism at both the modality and token levels to dynamically select and combine different expert models. At the modality level, it enables complementary information fusion across multimodal features, while at the token level, it enhances local fine-grained feature modeling, thereby improving the accuracy of multimodal fused features.

## III. CURATED DATASET

Infrared-visible fusion algorithms have predominantly been evaluated over static image datasets (e.g., RoadScene, LLVIP). However, these datasets lack temporal continuity, thereby

failing to capture motion dynamics and temporal consistency in real-world scenarios. In recent years, several researchers have begun constructing infrared-visible video fusion datasets, such as HDO and M3SVD, which have promoted the research and evaluation of fusion methods in video scenes to some extent. However, existing video fusion datasets primarily cover regular illumination conditions (daytime or nighttime), and the nighttime scenes rarely contain large-scale extremely low-light regions, making them insufficient for research on infrared-visible video fusion under extremely low-light conditions. For example, in the publicly available HDO test dataset, which contains 24 scenes, only six scenes can be classified as low-light scenes. In the publicly available M3SVD test dataset, which comprises 30 scenes, only three are classified as low-light scenes. Given the severe scarcity of video data for extremely low-light scenes, we constructed an infrared-visible video fusion dataset specifically for such scenarios. By analyzing the spatiotemporal uncertainty and modality learning imbalance faced by infrared-visible video fusion tasks under extremely low-light conditions, we provide a more representative experimental benchmark for researching and evaluating video fusion tasks.

The ELVID dataset was captured using the IRS-PT4 dual-spectrum imaging system, with recording sessions conducted between 8:00 p.m. and 11:00 p.m. Due to differences in the field of view between the infrared and visible sensors in the IRS-PT4 system, the original video pairs exhibit slight spatial misalignment. To ensure precise cross-modal alignment, a geometric correction and temporal registration procedure based on LightGlue [54] was applied to all collected videos. To further improve the reliability of the registration results, we conducted manual verification after the automatic registration to minimize potential misregistration, thereby ensuring the accuracy of the dataset construction. The ELVID dataset contains 4739 paired infrared-visible video sequences, totaling 78,299 frames after frame-level conversion. As illustrated in Fig. 2, ELVID covers five representative scene categories, including campus sport scenes, traffic scenes, building scenes, pedestrian scenes, and campus environment scenes. Specifically, the campus sport scenes are divided into three subcategories: basketball courts, football fields, and tennis courts. The traffic scenes include four types: roads, parking lots, traffic signs, and temporarily parked vehicles. The building scenes are categorized into two groups based on the focal length: long focus (distant view) and short focus (close view). The pedestrian scenes comprise three patterns: single-person, multiple-person walking, and crowd gathering. The campus environment scenes include campus roads, vegetation, stairways, signboards, and other campus surroundings. The number of videos and their corresponding frame counts for each category are summarized in Fig. 2. Overall, the scene diversity and extremely low-light conditions of ELVID provide strong support for evaluating infrared-visible video fusion methods.

## IV. METHODOLOGY

### A. Architecture Overview

As shown in Fig. 3, given the infrared video sequence  $\mathbf{I} = \{\mathbf{I}_t\}_1^n$  and visible video sequence  $\mathbf{V} = \{\mathbf{V}_t\}_1^n$ , USCFNet

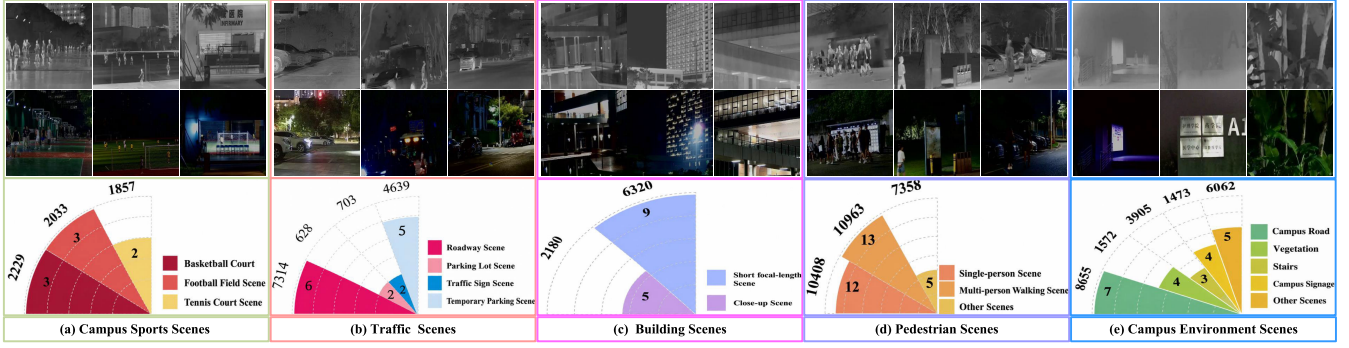


Fig. 2. Typical examples in our ELVID dataset. For the scene category distribution, the numbers on the periphery indicate the total number of frames across different categories, and the numbers in the fan-shaped area indicate the number of videos across different categories.

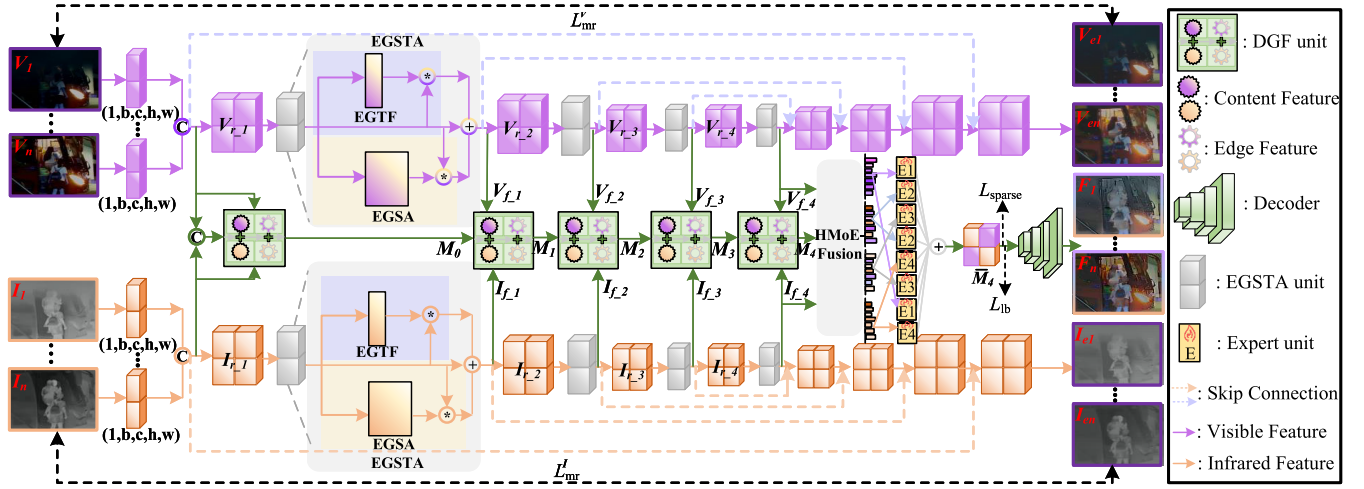


Fig. 3. The framework of USCFNet. USCFNet employs three parallel U-shaped encoder-decoder branches: the purple feature stream is the visible branch, the orange stream is the infrared branch, and the green stream is the feature fusion branch. EGSTA comprises two complementary units: EGTF and EGSA. EGTF represents the entropy-gated temporal fusion. EGSA represents the entropy-gated spatial attention. HMoE-Fusion represents the hierarchical Mixture-of-Experts fusion block.  $\mathcal{L}_{lb}$  is the load-balance loss,  $\mathcal{L}_{sparse}$  is the sparsity loss.  $I_{r-s}$  and  $V_{r-s}$  represent the initial infrared and visible multi-scale features, respectively.  $I_{f-s}$  and  $V_{f-s}$  represent the spatio-temporal features of infrared and visible, respectively.  $M_s$  represents the multi-scale fusion feature.  $\mathcal{L}_{mr}^I$  and  $\mathcal{L}_{mr}^V$  represent the reconstruction loss of infrared and visible, respectively.

employs three parallel U-shaped encoder-decoder branches (infrared branch, visible branch, and fusion branch) to generate the enhanced infrared video sequence  $I_e = \{I_{et}\}_1^n$ , the enhanced visible video sequence  $V_e = \{V_{et}\}_1^n$ , and the final fused video sequence  $F = \{F_t\}_1^n$ .

In the infrared branch, a ResNet encoder is applied to each frame of the infrared video sequence  $I$  to extract the corresponding multi-scale features  $I_{r-s} = \{I_{r-st}\}_{1,1}^{4,n}$ .  $s$  is the number of scales with  $s \in (1, 4)$ , and  $t$  denotes the frame index with  $t \in (1, n)$ . Subsequently, to mitigate unreliable cross-frame information introduced by low-quality frames, we introduce an EGSTA unit to enhance the spatio-temporal consistency of the multi-scale features  $I_{r-s}$ , yielding the refined spatio-temporal modeling representations  $I_{f-s} = \{I_{f-st}\}_{1,1}^{4,n}$ . Finally, the refined infrared features  $I_{f-s}$  are fed into the infrared-branch decoder to produce the enhanced infrared video sequence  $I_e$ . Notably, the visible and infrared branches share an identical architectural design.

For the fusion branch, we first develop a DGF module to fuse the refined infrared features  $I_{f-s}$ , visible features  $V_{f-s}$ , and the fusion features  $M_{s-1} = \{M_{s-1,t}\}_{1,1}^{4,n}$  from the previous

scale, to obtain the cross-modal fusion feature  $M_s = \{M_{s,t}\}_{1,1}^{4,n}$  for each scale. Subsequently,  $I_{f-4}$ ,  $V_{f-4}$  and  $M_4$  are jointly fed into the HMoE\_Fusion module to obtain the refined fusion feature  $\bar{M}_4 = \{\bar{M}_{4,t}\}_1^n$ . The HMoE\_Fusion module employs a dynamic expert selection mechanism to ensure that the fused features maintain balanced modality information while effectively preserving local details. Finally, the features  $\bar{M}_4$  are fed into a decoder to obtain the fused video sequence  $F$ . The EGSTA, DGF, and HMoE\_Fusion modules are the core components of USCFNet, and the details are provided below.

### B. Entropy-Gated SpatioTemporal Attention

Under extremely low-light conditions, visible video often suffers from insufficient brightness, reduced contrast, and the loss of information in small-object or weak-texture regions. In contrast, infrared video provides relatively stable contour and energy information but lacks fine-grained texture. Low-quality video frames often mislead feature learning between adjacent frames, thereby introducing erroneous or invalid

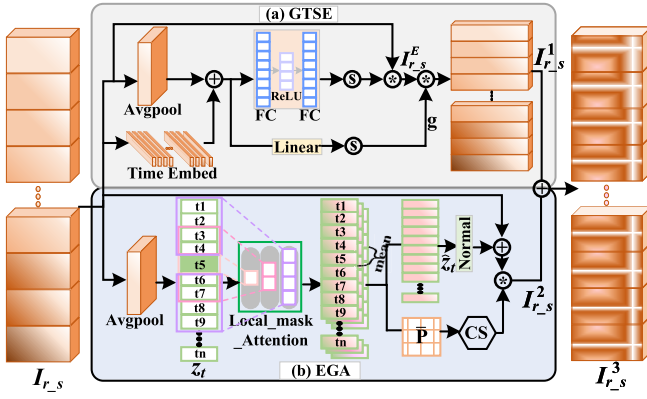


Fig. 4. The framework diagram of the EGTF module. CS is the temporal confidence score.  $\bar{P}$  is the attention distributions.  $S$  is the sigmoid function.  $g$  is the frame level gate.  $z_t$  is the temporal token.  $I_{r-s}^1$  is the enhanced intra-frame features.  $I_{r-s}^2$  is the enhanced inter-frame features.  $I_{r-s}^3$  is the temporal consistency feature.

information into the fusion process and generating artifacts. To mitigate the influence of low-quality frames or low-confidence regions, we propose the EGSTA module to ensure spatiotemporal consistency in modality-specific features. EGSTA comprises two complementary units: the Entropy-Gated Temporal Fusion (EGTF) unit and the Entropy-Gated Spatial Attention (EGSA) unit. EGTF divides video modeling from a temporal perspective into intra-frame enhancement and inter-frame reasoning, enhancing single-frame representation while maintaining temporal consistency across frames. EGSA models the spatial consistency of video sequences and introduces attention entropy as an uncertainty measure to suppress unreliable information. Taking the infrared branch as an example,  $I_{r-s}$  is used as the input of EGSTA, and the EGTF and EGSA units are executed at the same time to obtain the corresponding temporal consistency and spatial consistency features. The two are added together to obtain the final spatiotemporal modeling feature  $I_{f-s}$ .

1) *Entropy-Gated Temporal Fusion*: In low-light video sequences, temporal continuity can supplement the missing information in a single frame. However, it also faces the challenge of low-quality video frames misleading the learning of other frames. To address this challenge, we propose the EGTF unit, which decomposes temporal modeling into intra-frame enhancement and inter-frame inference to improve feature temporal consistency learning, as shown in Fig. 4. Specifically, the former feeds  $I_{r-s}$  into a Gated Time-aware SE (GTSE) block to produce the enhanced intra-frame feature  $I_{r-s}^1 = \{I_{r-st}^1\}_{1,1}^{4,n}$ . The latter inputs  $I_{r-s}$  into an Entropy-Gated Attention (EGA) unit, which characterizes inter-frame dependencies through multi-scale local masked attention. The attention entropy is then mapped to frame-level confidence scores to adjust video sequence information adaptively, thereby obtaining the enhanced inter-frame feature  $I_{r-s}^2 = \{I_{r-st}^2\}_{1,1}^{4,n}$ . Finally, the intra-frame enhanced features  $I_{r-s}^1$  and inter-frame enhanced features  $I_{r-s}^2$  are added to obtain the temporal consistency features  $I_{r-s}^3 = \{I_{r-st}^3\}_{1,1}^{4,n}$ .

a) *Gated time-aware SE block*: Pronounced brightness fluctuations and frame quality inconsistencies are common in

low-light video sequences. To evaluate the reliability of each frame in the video, the proposed GTSE block first applies global average pooling to the input features  $I_{r-s}$  to extract frame-wise channel features. The positional vector for each frame is then added to its corresponding channel features, enabling the model to capture semantic information across time positions. Subsequently, a squeeze-and-excitation (SE) [55] operation (FC-ReLU-FC) is performed on each frame to recalibrate the channel responses, thereby producing the corresponding weights  $I_{r-s}^w = \{I_{r-st}^w\}_{1,1}^{4,n}$ . Finally, the input features  $I_{r-s}$  are multiplied with the weights  $I_{r-s}^w$  to obtain the enhanced features  $I_{r-s}^E = \{I_{r-st}^E\}_{1,1}^{4,n}$ . To further ensure that the model can selectively emphasize high-quality frames while suppressing invalid ones, a linear operation is applied to the channel token features to obtain the corresponding frame-level gate  $g$ . Then multiply the enhanced features  $I_{r-s}^E$  with the frame-level gating  $g$  to get the enhanced intra-frame feature  $I_{r-s}^1$ , as shown in Fig. 4 (a).

b) *Entropy-gated attention unit*: EGA is used to characterize the dependencies between frames in a video sequence and suppress unreliable cross-frame information injection, as shown in Fig. 4 (b). Given the input feature sequence  $I_{r-s}$ , global average pooling is first applied to each frame of  $I_{r-s}$  to extract its channel features, yielding the temporal token  $z_t$ . Subsequently, a multi-scale sliding window centered at the current frame index  $t$  is constructed with window radii  $h \in \{1, 2, 4\}$ .

$$\omega_h(t) = \{k \mid |k - t| \leq h\}, \quad h \in \{1, 2, 4\}, \quad (1)$$

where  $k$  is the index of the neighboring frames within the window. This design enables frame  $t$  to focus on short-term dynamic features within its local neighborhood, specifically the  $\pm 1$ ,  $\pm 2$ , and  $\pm 4$  adjacent frames.

Meanwhile, we define a local mask with  $L_{t,k}^{(h)} = 0$  if  $|k - t| \leq h$  and  $L_{t,k}^{(h)} = -\infty$  otherwise. Subsequently, for each sliding window, the attention distribution  $\{A_{st}^{(h)}\}_{1,1}^{4,n}$  is computed from the current frame  $t$  to its neighboring frames  $k \in \omega_h(t)$ . And then  $\{A_{st}^{(h)}\}_{s,t=1,1}^{4,n}$  are multiplied with the value term of  $z_t$  to generate the local features corresponding to each window scale. The local features across all window scales are then averaged to obtain the locally enhanced feature  $\hat{z}_t$ , which incorporates information from neighboring frames.

$$\{A_{st}^{(h)}\}_{s,t=1,1}^{4,n} = \text{Softmax}_k \left( \frac{(z_t W_q)(z_k W_k)^T}{\sqrt{d}} + L_{t,k}^{(h)} \right), \quad (2)$$

$$\hat{z}_t = \text{mean} \left( \{A_{st}^{(h)}\}_{s,t=1,1}^{4,n} \cdot (z_t W_v) \right), \quad (3)$$

where  $z_t$  denotes the temporal token,  $W_q$ ,  $W_k$ , and  $W_v$  are the learnable projection matrices for query, key, and value, respectively.  $d$  is the feature dimension used for normalization.

After obtaining the cross-frame attention map, we further evaluate the uncertainty estimation of the video sequence. Specifically, the attention distributions across different window scales are averaged to obtain  $\bar{P} = \text{mean}(\{A_{st}^{(h)}\}_{1,1}^{4,n})$ , and the temporal attention entropy  $H_t$  is subsequently computed from  $\bar{P}$  and normalized via min-max scaling to obtain the  $\hat{H}_t$ .

When the attention entropy is low, it indicates that the current frame's dependence on a few neighboring frames is

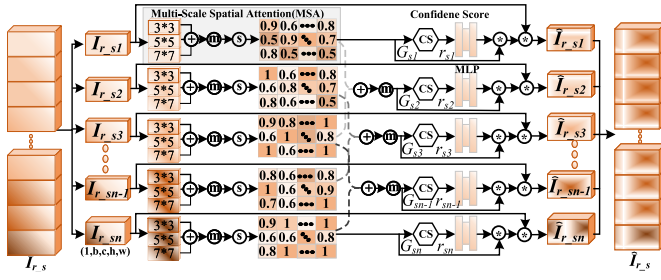


Fig. 5. Schematic diagram of the EGSA. CS is the spatial confidence score.  $I_{r,s}$  is the original input feature.  $s$  is the sigmoid activation.  $m$  represents the operator for calculating the average value.  $\hat{I}_{r,s}$  is the spatial consistency feature.  $\hat{I}_{r,s}$  denotes the spatial consistency feature.  $G_{st}$  represents the attention map with  $t \in \{1, 2, \dots, n\}$ .  $\gamma_{st}$  denotes the spatial confidence score. MLP is the multi-layer perceptron.

more stable; when the attention entropy increases, it suggests the presence of motion blur or interference from low-quality frames. Based on this, we map the attention entropy to a frame-level confidence score  $\beta_t = 1 - \hat{H}_t$ , which is used to gate the injection of cross-frame information. Finally, the locally enhanced feature  $\hat{z}_t$  is processed by layer normalization and added to the original feature  $I_{r,s}$  to obtain the temporally enhanced representation. Multiply the frame-level confidence  $\beta_t$  with the temporally enhanced representation to obtain the enhanced inter-frame feature  $I_{r,s}^2$ , as shown in Fig. 4 (b).

2) *Entropy-Gated Spatial Attention(EGSA)*: Due to the severe degradation present in extremely low-light video data, overly dark regions often suffer from insufficient brightness, missing textures, and blurred structures, resulting in pronounced spatial uncertainty. To address this, the EGSA unit estimates regional confidence based on attention entropy, which reflects the importance of different spatial areas and guides feature extraction accordingly, as shown in Fig. 5. Specifically, for each frame feature  $I_{r,st}$  in the input feature sequence  $I_{r,s}$ , we apply convolution operations with kernel sizes of  $3 \times 3$ ,  $5 \times 5$ , and  $7 \times 7$  to generate multi-scale spatial attention maps. These maps are then averaged and passed through a sigmoid activation to obtain the spatial attention matrix  $\{B_{s,t}\}_{1,1}^{4,n}$ .

Subsequently, the attention weights of boundary frames are kept unchanged. For non-boundary frames, a sliding window mechanism is introduced to average the attention maps of the current frame with those of its adjacent frames, thereby obtaining a cross-frame consistent attention map  $\hat{B}_{s,t}$ .

$$\hat{B}_{s,t} = \begin{cases} \frac{1}{3}(B_{s,t-1} + B_{s,t} + B_{s,t+1}), & 1 \leq t \leq n-2, \\ B_{s,t}, & t \in \{0, n-1\}. \end{cases} \quad (4)$$

Then, the spatial attention entropy is computed based on the attention map  $\hat{B}_{s,t}$  and normalized to obtain  $G_{st}$ . The spatial confidence score  $\gamma_{st} = 1 - G_{st}$  is derived as a gate of spatial reliability, which guides the network to suppress unreliable regions while emphasizing high-confidence spatial features during training.

Finally, a multi-layer perceptron (MLP) is applied to  $\gamma_{st}$  to obtain an adaptive fusion coefficient  $\alpha_{st} \in [0, 1]$ , which serves as a modulation factor to assist the model in determining

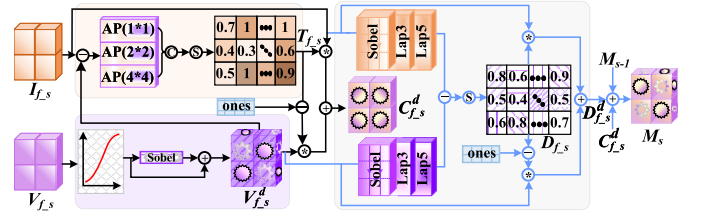


Fig. 6. Schematic diagram of the DGF.  $\odot$  is the concatenated operator.  $S$  is the sigmoid activation. The purple area represents the learnable contrast stretching and edge compensation operations on the visible features. The orange area represents the fusion of features at the content difference level. The gray area represents the fusion of features at the edge difference level.  $T_{f,s}$  is the attention weight map.  $V_{f,s}^d$  is the visible enhanced features.  $D_{f,s}^d$  is the edge consistency mask.  $C_{f,s}^d$  is the content-guided features.  $D_{f,s}^d$  is the edge-guided features.  $M_s$  is the fused feature.

whether to trust the confidence score  $\gamma_{st}$ . Subsequently,  $\alpha_{st}$  is used to weight the original feature  $I_{r,st}$ , yielding the refined feature  $\hat{I}_{r,st}$ . By concatenating all frame-level features, we obtain the video representation  $\hat{I}_{r,s} = \{\hat{I}_{r,st}\}_{1,1}^{4,n}$  with the spatial consistency, as shown in Fig. 5.

$$\hat{I}_{r,st} = I_{r,st} \odot (\alpha_{st} G_{st} + (1 - \alpha_{st}) \cdot \mathbf{E}), \quad (5)$$

where  $\mathbf{E}$  denotes an all-ones identity tensor having the same dimension as  $I_{r,st}$ , and  $\odot$  represents element-wise multiplication. Based on the original frame feature  $I_{r,st}$ , the adaptive coefficient  $\alpha_{st}$  controls the influence strength of the confidence factor  $\gamma_{st}$ . In this way, unreliable features can be suppressed while salient regions and key structures are highlighted under extremely low-light conditions.

### C. Difference Guided Fusion Module

The DGF module designs the feature fusion process around the content and edge differences between infrared and visible features, thereby preserving the fused feature structure and detailed information. As shown in the purple area in Fig. 6, DGF first performs the contrast enhancement operation  $\sigma_c(\cdot)$  on visible feature  $V_{f,s}$  to enhance the contrast of dark areas. And then, an edge compensation operation  $Sobel(\cdot)$  is performed on the contrast enhancement feature  $V_{f,s}^c$  to obtain the  $V_{f,s}^d$ :

$$\sigma_c = \frac{1}{1 + \exp(-\delta(V_{f,s} - \lambda))}, \quad (6)$$

$$V_{f,s}^d = V_{f,s}^c + \epsilon \cdot Sobel(V_{f,s}^c), \quad (7)$$

where  $V_{f,s}^c = V_{f,s} \odot \sigma_c(V_{f,s})$ , and  $\odot$  represents element-wise multiplication.  $\delta$  and  $\lambda$  are learnable parameters controlling contrast stretching. In this study, we set  $\delta = 4$ ,  $\lambda = 0.5$ .  $\epsilon$  is balance coefficient,  $\epsilon = 0.1$ .

Subsequently, the semantic information of specific modalities is fused based on content differences, as shown in the orange area in Fig. 6. Concretely, the content difference between modalities is calculated as  $D_{f,s} = |I_{f,s} - V_{f,s}^d|$ . Then, multi-scale spatial pyramid pooling is performed (using convolution kernels of  $1 \times 1$ ,  $2 \times 2$ , and  $4 \times 4$ , respectively) on the  $D_{f,s}$  to obtain the corresponding multi-scale features. The obtained features are subsequently upsampled, concatenated,

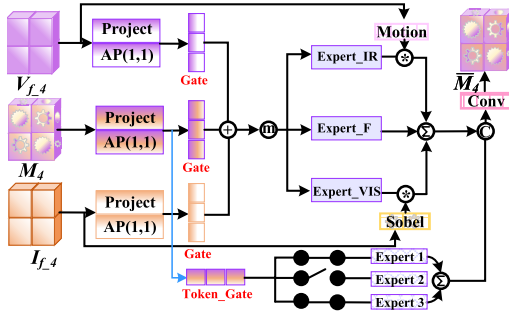


Fig. 7. Schematic diagram of the HMoE\_Fusion.

and activated by Sigmoid to generate the attention weight map  $T_{f-s} = \{T_{f-s}\}_{1,1}^{4,n}$ . After that, adaptive weighting is employed to dynamically determine which modality dominates the selection of fusion features, thereby achieving modality fusion at the content level. The specific adaptive weighting method is as follows:

$$C_{f-s}^d = T_{f-s} \odot I_{f-s} + (\mathbf{E} - T_{f-s}) \odot V_{f-s}^d, \quad (8)$$

where  $\mathbf{E}$  denotes an all-ones identity tensor having the same dimension as  $T_{f-s}$ .

To avoid loss of edge information in fused features, the DGF module further modulates feature weights based on edge differences. Given that the Sobel operator is robust to noise and the Laplacian operator facilitates the extraction of global structures, we propose to combine them to ensure fused features with reliable edge information. Specifically, Laplacian operators with kernel sizes of  $3 \times 3$  and  $5 \times 5$  together with Sobel operators are employed to extract multi-scale edge features  $D_s^{\text{ir}} = \{D_{st}^{\text{ir}}\}_{1,1}^{4,n}$  and  $D_s^{\text{vis}} = \{D_{st}^{\text{vis}}\}_{1,1}^{4,n}$ . Subsequently, edge consistency mask  $D_{f-s} = \text{Sigmoid}(|D_s^{\text{ir}} - D_s^{\text{vis}}|)$  is generated. Through adaptive weighting, the model is guided to select more reliable edge information at the structural level, as shown in the gray area in Fig. 6. The specific adaptive weighting method is as follows:

$$D_{f-s}^d = D_{f-s} \odot I_{f-s} + (\mathbf{E} - D_{f-s}) \odot V_{f-s}^d, \quad (9)$$

where  $\mathbf{E}$  denotes an all-ones identity tensor having the same dimension as  $D_{f-s}$ . To enhance inter-layer consistency, the fused features of the previous layer  $M_{s-1}$  are aligned and mapped to the current scale, and then added to the content-guided features  $C_{f-s}^d$  and the edge-guided features  $D_{f-s}^d$  to obtain the final fused features  $M_s = \{M_{s,t}\}_{1,1}^{4,n}$  on each layer.

#### D. Hierarchical Mixture-of-Experts Fusion Block

To further enhance the representational power of fused features and ensure balanced learning of modality-specific features, we design an HMoE\_Fusion, as shown in Fig. 7. This module employs a two-level routing mechanism that incorporates modality-level and token-level expert routing to dynamically select and combine experts, thereby achieving global modality complementarity fusion and local fine-grained modeling.

In the modal-level routing, the infrared feature  $I_{f-4} = \{I_{f-4t}\}_{t=1}^n$ , visible feature  $V_{f-4} = \{V_{f-4t}\}_{t=1}^n$ , and fused feature

$M_4 = \{M_{4t}\}_{t=1}^n$  are first processed by global average pooling to obtain the corresponding global representations. Subsequently, three independent parameter matrices are employed to generate expert selection distributions, each perturbed with noise, and a top- $k$  strategy is used to improve both robustness and sparsity. Finally, the gating results of the three modalities are averaged to obtain an integrated expert selection distribution, thereby avoiding single-modality bias and ensuring stable expert selection. Correspondingly, the infrared expert is combined with a motion map generated from temporal differences to highlight dynamic regions. In contrast, the visible expert is combined with edge map extracted using the Sobel operator to enhance texture details. The fusion expert directly processes the fused feature  $M_4$ . The outputs of the three types of experts are then weighted and fused under the guidance of the gating weights to produce the modal-level fusion result.

Based on the model-level routing mechanism, we introduce a token-level expert routing mechanism to perform expert allocation and aggregation of local fragments of the fused feature  $M_4$  on a token-by-token basis. Specifically, a projection matrix is first employed to generate the token-level routing distribution, based on which the fused feature  $M_4$  is assigned to multiple lightweight MLP experts. Each expert independently models the assigned tokens, and their outputs are subsequently aggregated to form the token-level fusion result. Finally, the outputs of the modal-level and token-level experts are concatenated along the channel dimension, followed by a  $1 \times 1$  convolution to compress and reorganize the feature, thereby generating the final fused representation  $\bar{M}_4 = \{\bar{M}_{4t}\}_1^n$  with global consistency and enhanced local details.

#### E. Loss Function

In general, infrared sequences capture thermal radiation intensity and salient targets but lack structural detail. In contrast, visible sequences exhibit rich textures and edges, but degrade severely under low illumination. Given this, we utilize a modality-aware hybrid loss  $\mathcal{L}_h$  to train the network, which incorporates content-structural constraint  $\mathcal{L}_{cs}$ , contrast constraint  $\mathcal{L}_c$ , temporal consistency  $\mathcal{L}_{tc}$ , modality-specific reconstruction  $\mathcal{L}_{mr}$ , and regularization constraints  $\mathcal{L}_r$ :

$$\mathcal{L}_h = \mathcal{L}_{cs} + \mathcal{L}_c + \mathcal{L}_{tc} + \mathcal{L}_{mr} + \mathcal{L}_r. \quad (10)$$

To simultaneously constrain the content consistency, gradient fidelity, and structural similarity of the fusion video with the modality-specific video, this paper combines the content loss  $\mathcal{L}_{con}$ , gradient loss  $\mathcal{L}_{grad}$ , and structural similarity loss  $\mathcal{L}_{ssim}$  into a unified content-structure loss  $\mathcal{L}_{cs}$ :

$$\mathcal{L}_{cs} = \mathcal{L}_{con} + \mathcal{L}_{grad} + \mathcal{L}_{ssim}. \quad (11)$$

Considering that visible images often suffer from detail loss and contrast degradation, while infrared images can stably provide energy distribution information under low-light conditions,  $\mathcal{L}_{con}$  consists of two components, each of which corresponds to an imaging modality:

$$\mathcal{L}_{con} = 1.5\mathcal{L}_{content}^{\text{ir}} + \mathcal{L}_{content}^{\text{vis}}, \quad (12)$$

where  $\mathcal{L}_{\text{content}}^{\text{ir}} = \frac{1}{n} \sum_{t=1}^n \|\omega_{\text{ir}} \odot (\mathbf{F}_t - \mathbf{I}_t)\|_1$  denotes the infrared content loss and  $\mathcal{L}_{\text{content}}^{\text{vis}} = \frac{1}{n} \sum_{t=1}^n \|\mathbf{F}_t - \mathbf{V}_t\|_1$  denotes the visible content loss. Notably, for infrared content loss, an infrared weight map  $\omega_{\text{ir}}$  is constructed from saliency and gradient features to emphasize “thermally salient regions” and structural edges, thereby strengthening the constraint on the fusion results in these critical areas.

$$\omega_{\text{ir}} = \mathbf{1} + 2\text{Hot}(\{\mathbf{I}_t\}_{t=1}^n) + \nabla(\{\mathbf{I}_t\}_{t=1}^n), \quad (13)$$

where  $\text{Hot}(\{\mathbf{I}_t\}_{t=1}^n)$  is obtained by applying an upper quantile threshold (e.g., top 10% high-response regions) on the infrared intensity map to extract thermally salient regions.  $\nabla(\{\mathbf{I}_t\}_{t=1}^n)$  denotes the normalized edge magnitude derived from gradient amplitudes. The grad loss  $\mathcal{L}_{\text{grad}}$  consists of two components, each of which corresponds to an imaging modality:

$$\mathcal{L}_{\text{grad}} = 8\mathcal{L}_{\text{grad}}^{\text{ir}} + 3\mathcal{L}_{\text{grad}}^{\text{vis}}, \quad (14)$$

where  $\mathcal{L}_{\text{grad}}^{\text{ir}} = \frac{1}{n} \sum_{t=1}^n \|\omega_{\text{ir}} \odot (\nabla \mathbf{F}_t - \nabla \mathbf{I}_t)\|_1$  denotes the infrared grad loss,  $\nabla$  denotes the gradient operator.  $\mathcal{L}_{\text{grad}}^{\text{vis}} = \frac{1}{n} \sum_{t=1}^n \|\omega_{\text{vis},t} \odot (\nabla \mathbf{F}_t - \nabla \mathbf{V}_t)\|_1$  denotes the visible grad loss. For the visible gradient loss, we define a visible weight map as  $\omega_{\text{vis},t} = \mathbf{1} + 1.5\nabla(\{\mathbf{V}_t\}_{t=1}^n)$  based on the edge intensity of the visible modality, which can emphasize the consistency of fine details and edges to prevent texture blurring under extremely low-light conditions. In addition to pixel- and gradient-level constraints, we further introduce a structural similarity (SSIM) loss  $\mathcal{L}_{\text{ssim}} = \mathcal{L}_{\text{ssim}}^{\text{ir}}(\mathbf{F}, \mathbf{I}) + \mathcal{L}_{\text{ssim}}^{\text{vis}}(\mathbf{F}, \mathbf{V})$  to preserve perceptual-level structural consistency between the fused result and modalities-specific.

The contrast constraint  $\mathcal{L}_c$  is introduced to align the global contrast of the fused video with that of the visible video, thereby mitigating the negative effects of low-light degradation. Specifically, the horizontal and vertical gradients are computed using Sobel operators to define the contrast map  $C(\cdot)$ . The loss is then formulated as the  $\mathbb{L}_1$  distance between the fused and visible sequences, i.e.,  $\mathcal{L}_c = \|C(\mathbf{F}) - C(\mathbf{V})\|_1$ , encouraging the fused images to preserve the local gradients and contrast of the visible modality.

The temporal consistency loss  $\mathcal{L}_{\text{tc}}$  is designed as  $\mathcal{L}_{\text{tc}} = \mathcal{L}_{\text{tc}}^{\text{I}} + \mathcal{L}_{\text{tc}}^{\text{V}}$ . Each component consists of a frame-difference term  $\mathcal{L}_{\text{fd}}^{\text{I}}$  and a spatial-gradient term  $\mathcal{L}_{\text{sg}}^{\text{I}}$ :

$$\mathcal{L}_{\text{fd}}^{\text{I}} = \frac{1}{n-1} \sum_{t=2}^n \|(\mathbf{F}_t - \mathbf{F}_{t-1}) - (\mathbf{R}_t - \mathbf{R}_{t-1})\|_1, \quad (15)$$

$$\mathcal{L}_{\text{sg}}^{\text{I}} = \frac{1}{n-1} \sum_{t=2}^n (\|\nabla_x \mathbf{F}_t - \nabla_x \mathbf{F}_{t-1}\|_1 + \|\nabla_y \mathbf{F}_t - \nabla_y \mathbf{F}_{t-1}\|_1). \quad (16)$$

The former enforces the inter-frame variations of the fused video  $\mathbf{F}$  to remain consistent with those of the reference modality  $\mathbf{R}$ ,  $\mathbf{R} \in (\mathbf{I} \vee \mathbf{V})$ , thereby avoiding motion artifacts. The latter enforces smooth gradient transitions between consecutive frames, reducing temporal jitter and flickering.

The modality-specific reconstruction loss  $\mathcal{L}_{\text{mr}}^{\text{R}} (\mathbf{R} \in (\mathbf{I} \vee \mathbf{V}))$  jointly constrains the reconstructed sequences by combining pixel-level errors and structural similarity, aiming to preserve modality-specific fidelity. Specifically, the infrared branch

output  $\mathbf{I}_e$  and the visible branch output  $\mathbf{V}_e$  are required to approximate their respective input sequences  $\mathbf{I}$  and  $\mathbf{V}$ .

$$\mathcal{L}_{\text{mr}} = 2 \cdot \mathcal{L}_{\text{mr}}^{\text{I}}(\mathbf{I}_e, \mathbf{I}) + 1 \cdot \mathcal{L}_{\text{mr}}^{\text{V}}(\mathbf{V}_e, \mathbf{V}). \quad (17)$$

In the hybrid expert architecture, the regularization term  $\mathcal{L}_{\text{r}}$  prevents the router from uniformly activating all experts by combining the load-balance loss  $\mathcal{L}_{\text{lb}}$  with the sparsity loss  $\mathcal{L}_{\text{sparse}}$ , thereby improving sparsity and diversity and stabilizing the training process [25]. The specific calculation method of  $\mathcal{L}_{\text{r}}$  is as follows:

$$\mathcal{L}_{\text{r}} = \mathcal{L}_{\text{lb}} + \mathcal{L}_{\text{sparse}}, \quad (18)$$

where  $\mathcal{L}_{\text{lb}}$  is the load-balancing loss, preventing routing imbalance across experts,  $\mathcal{L}_{\text{sparse}}$  encourages sparse expert activation.

## V. EXPERIMENTS

### A. Experiment Setup

1) *Datasets*: The proposed fusion scheme is evaluated on the ELVID dataset and two publicly available datasets, HDO and M3SVD. In the ELVID dataset, all original videos were first divided into 16-frame sub-sequences. Subsequently, the training and testing sets were randomly split on a per-video basis within the same scene. To mitigate potential data leakage, we carefully ensured that temporally adjacent clips or highly similar sequences were not assigned to both the training and testing sets. Finally, the ELVID dataset contains 4,180 training sequence pairs and 559 testing sequence pairs. Similarly, the HDO and M3SVD datasets are split into sub-sequences for model evaluation, containing 371 and 1,379 pairs, respectively.

2) *Evaluation Metrics*: This paper selects Average Gradient (AG) [56], Spatial Frequency (SF) [57], Enhancement Measure Evaluation (EME) [58], and Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [59] as evaluation metrics for the fusion performance of video frames. AG measures the overall sharpness of the image. SF measures the image’s detail and texture richness. EME measures the degree of contrast enhancement in local areas of the image. BRISQUE assesses the visual perception quality of the image.

3) *Compared Methods*: For comparative analysis, the proposed method is evaluated against six state-of-the-art fusion approaches, including CrossFuse [60], DANet [61], ITFuse [62], TextFusion [63], HAIAFusion [64], and SeAFusion [7]. To ensure fairness, the official pretrained weights released by the respective authors are employed for testing on three datasets. Since few existing studies address video-level fusion and no publicly available code is available [4], [18], the first frame of each test video is used as the test image for quantitative evaluation across all methods.

4) *Implementation Details*: Our method is implemented in PyTorch and conducted on an NVIDIA GeForce RTX 3090 GPU. During training, the Adam optimizer is employed to optimize the network parameters, with a learning rate of  $1 \times 10^{-4}$ . The batch size is set to 2, and the network is trained for 50 epochs. Each video sequence comprises 16 frames, and each frame has a spatial resolution of  $256 \times 256$  pixels.

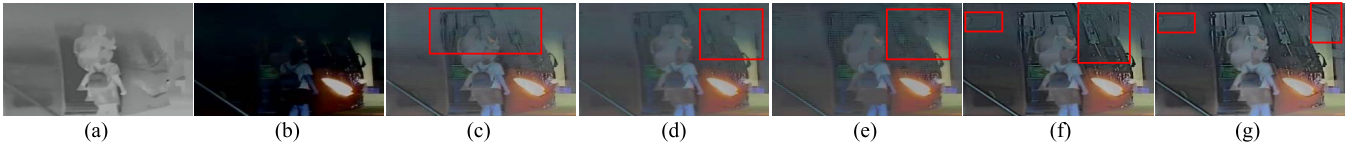


Fig. 8. Visualization results of the fused image under different loss configurations.

### B. Ablation Experiments

1) *Loss Function Analysis*: In our research, we employ a modality-aware hybrid loss to fully train the network, as formulated in Eq. (11). Here, we conduct a series of ablation studies to investigate the impact of different loss components. Fig. 8 presents a visualization of the fusion results under different loss configurations. (a) denotes the original infrared image, (b) the original visible image, and (c–j) correspond to the visualization of the fusion results under different loss functions. Fig. 8(c) shows the fusion result using only content-structure constraints loss  $\mathcal{L}_{cs}$ . Fig. 8(d) shows the fusion result using  $\mathcal{L}_{cs} + \mathcal{L}_{tc}$ . Fig. 8(e) shows the fusion result using  $\mathcal{L}_{cs} + \mathcal{L}_{tc} + \mathcal{L}_c$ . Fig. 8(f) shows the fusion result using  $\mathcal{L}_{cs} + \mathcal{L}_{tc} + \mathcal{L}_c + \mathcal{L}_{mr}$ . Fig. 8(j) shows the fusion result using  $\mathcal{L}_{cs} + \mathcal{L}_{tc} + \mathcal{L}_c + \mathcal{L}_{mr} + \mathcal{L}_r$ . Fig. 8(g) is used to verify the effectiveness of the regularization constraint, which is introduced by the HMoE\_Fusion module to avoid imbalance in expert assignments during model training. Therefore, Fig. 8(g) reflects the effectiveness of the HMoE\_Fusion module.

When only the content-structural loss is applied (Fig. 8(c)), the fusion result preserves the basic infrared energy distribution and visible detail information, but exhibits over-sharpening in the red box region, leading to slightly unnatural edge structures. With the additional incorporation of the temporal consistency constraint (Fig. 8(d)), the fusion result becomes smoother. However, the texture information in the low-light region (red box region) remains insufficiently revealed. After further introducing the visible contrast constraint (Fig. 8(e)), the local texture and details of the fused image are enhanced, but the fused image suffers from severe artifacts. Figs. 8 (c–e) show that temporal consistency can enable the model to learn cross-frame dynamic information more fully. When modality-specific reconstruction loss is added further (Fig. 8(f)), the alignment of the infrared and visible features becomes more stable, and the outline of the person and the local details (such as the front windshield and door windows of vehicles in the red box regions) are significantly improved. Finally, after introducing the regularization constraint term (Fig. 8(g)), the fusion result showed the best performance in terms of structural detail, brightness, and artifact suppression. In the red box area, it can be observed that the windows, headlights, and ceiling are clearer, and local textures are preserved more accurately. Overall, the fusion result generated by the modality-aware hybrid loss is the most outstanding in terms of visual consistency and structural fidelity.

2) *SpatioTemporal Attention Unit Analysis*: This paper proposes an EGSTA module to ensure the spatiotemporal

TABLE I  
ABLATION EXPERIMENTAL ANALYSIS OF EGSTA

| Methods   | BRISQUE↓     | SF↑          | AG↑          | EME↑         |
|-----------|--------------|--------------|--------------|--------------|
| EGSA      | 18.24        | 20.40        | 6.596        | 7.782        |
| EGTF+EGSA | <b>16.09</b> | <b>21.53</b> | <b>6.843</b> | <b>8.394</b> |

TABLE II  
ABLATION EXPERIMENTAL ANALYSIS OF DGF

| Methods             | BRISQUE↓     | SF↑          | AG↑          | EME↑         |
|---------------------|--------------|--------------|--------------|--------------|
| Content Difference  | 17.54        | 13.49        | 4.102        | 3.880        |
| Edge Difference     | 22.47        | 22.22        | 6.626        | 6.611        |
| Weighted Fusion     | 22.38        | <b>23.87</b> | <b>6.997</b> | 7.619        |
| Concatenated Fusion | 33.42        | 21.37        | 6.365        | 7.751        |
| DGF Module          | <b>16.09</b> | 21.53        | 6.843        | <b>8.394</b> |

consistency of feature modalities. EGSTA consists of two units: EGTF and EGSA. Table I shows the changes in the evaluation metrics of the proposed method after gradually introducing EGSA and EGTA. Compared with the baseline model, which employs only EGSA, the enhanced model with EGTF and EGSA achieves obvious improvements in terms of all evaluation metrics. Specifically, the BRISQUE value decreases by 2.15, indicating that incorporating temporal consistency modeling effectively reduces perceptual distortion in the fused images. Meanwhile, the SF, AG, and EME metrics all increase, indicating that the fused results exhibit higher local contrast and improved spatial sharpness. The overall results verify that introducing temporal consistency constraints based on spatial consistency modeling can more effectively leverage cross-frame correlations and enhance the detail preservation and visual perception quality of the fused image.

3) *Fusion Unit Analysis*: Table II presents the ablation results of the proposed DGF module. Since the DGF module consists of a content difference layer and an edge difference layer, we separately evaluate the fusion performance obtained by using only content differences and only edge differences. In addition, we compare the proposed DGF module with conventional weighted fusion and multi-feature concatenation strategies to further verify its effectiveness. The experimental results show that the DGF module achieves the lowest BRISQUE value, indicating that it can generate fused images with better perceptual quality and fewer visual distortions. Meanwhile, the DGF module obtains the best EME value, demonstrating its effectiveness in enhancing local contrast and improving the structural clarity and detailed information of the fused images. Although weighted fusion slightly outperforms

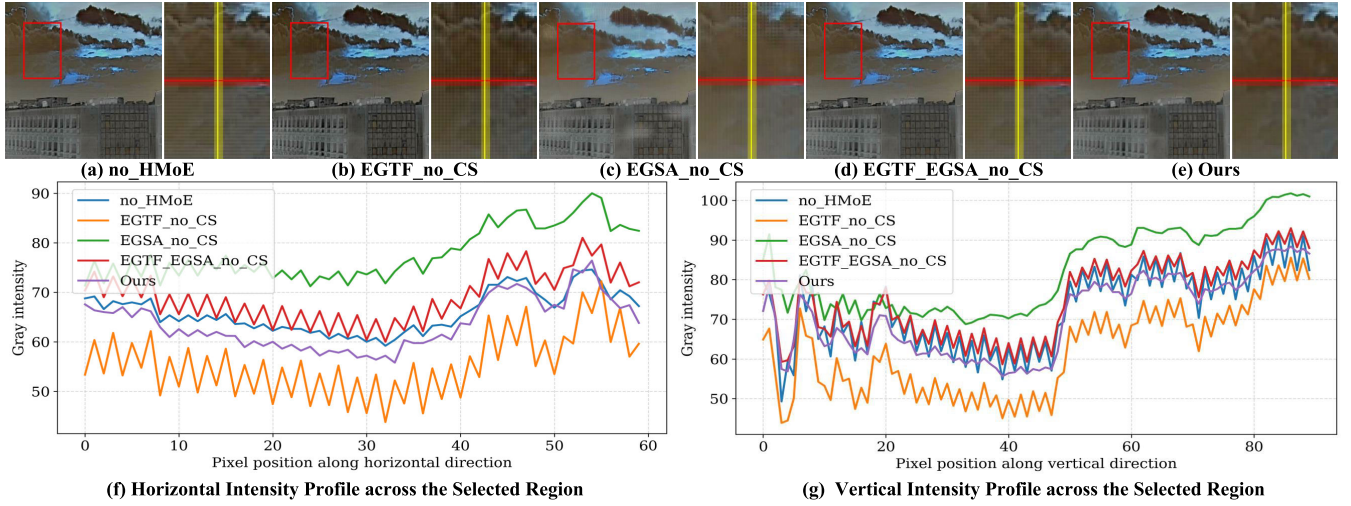


Fig. 9. Ablation analysis of the HMoE and the EGSTA unit without entropy gating.

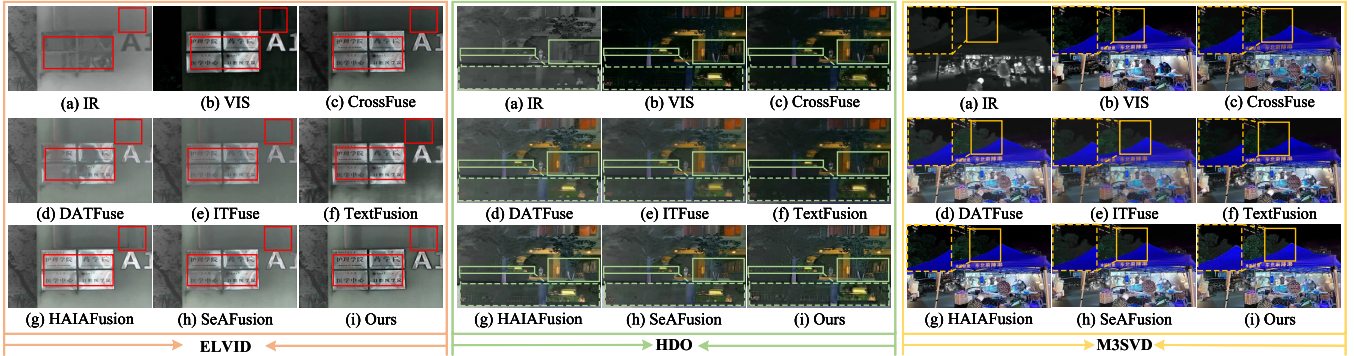


Fig. 10. Fusion performance analysis of different methods. The left side shows the fusion results on the private dataset ELVID. The middle section displays the fusion results on the public dataset HDO. The right side shows the fusion results on the public dataset M3SVD.

DGF on the SF and AG metrics, its higher BRISQUE score suggests that this fusion scheme may introduce more noise or visual distortion. In contrast, the DGF module can further enhance image naturalness and local contrast while preserving detailed information, thereby demonstrating superior fusion performance.

4) *HMoE and EGSTA Without Entropy Gating Analysis:* Fig. 9 presents the ablation analysis of the HMoE and EGSTA modules, where the EGSTA module consists of EGTF and EGSA. Figs. 9(a–e) show the fusion results under different ablation settings, together with the corresponding enlarged local regions. The red rectangles indicate the selected regions, while the red horizontal and yellow vertical lines in the enlarged regions denote the areas used to extract the horizontal and vertical gray-level intensity profiles, respectively. As shown in Fig. 9(a), removing the HMoE module results in noticeable horizontal artifacts in local regions. Figs. 9(b–d) further show that when the EGTF or (EGSA is used without entropy-gating constraints, the fused images are more likely to suffer from stripe-like artifacts in both horizontal and vertical directions. These observations suggest that standard attention mechanisms may introduce unstable responses when processing low-quality video frames or unreliable spatial regions. In contrast, Fig. 9(e) shows that when both HMoE and entropy

gating are incorporated, the fused image exhibits a more natural visual appearance and effectively suppresses artifacts. Figs. 9(f–g) further present the grayscale intensity distribution curves of selected regions in both the horizontal and vertical directions. Compared with the ablated variants, USCNet exhibits smoother, more stable intensity variations in both directions, indicating a stronger ability to suppress streak artifacts while preserving local details and structural information. These results further validate the effectiveness of the HMoE and EGSTA modules in enhancing detail representation and suppressing artifacts.

### C. Comparison of Frame-Based Fusion Performance

Since there is a lack of publicly available video fusion code, this paper first selects the first frame of the test video dataset as the analysis data to analyze the performance of different fusion methods. The choice of the first frame is motivated by two considerations. First, it serves as a unified starting frame across all video sequences, avoiding the subjective bias introduced by manually selecting the “best frame” and ensuring fair comparisons between methods. Second, the first frame is not affected by subsequent temporal propagation errors or accumulated motion blur, and thus can more stably

TABLE III  
QUANTITATIVE COMPARISON OF DIFFERENT FUSION METHODS ON ELVID, M3SVD, AND HDO DATASETS. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD** AND THE SECOND-BEST RESULTS ARE IN UNDERLINED. LOWER BRISQUE AND HIGHER SF, AG, AND EME INDICATE BETTER PERFORMANCE

| Methods               | ELVID        |              |              |              | M3SVD        |              |              |               | HDO          |              |              |              |
|-----------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|---------------|--------------|--------------|--------------|--------------|
|                       | BRISQUE↓     | SF↑          | AG↑          | EME↑         | BRISQUE↓     | SF↑          | AG↑          | EME↑          | BRISQUE↓     | SF↑          | AG↑          | EME↑         |
| CrossFuse             | 33.03        | 18.45        | 5.325        | 7.427        | 40.04        | 29.40        | 7.437        | 8.784         | 27.77        | 22.18        | 6.357        | 7.018        |
| DATFuse               | <u>26.37</u> | 12.79        | 4.199        | 3.493        | <b>21.07</b> | 21.19        | 6.372        | 4.908         | <u>15.97</u> | 18.65        | 6.565        | 4.732        |
| ITFuse                | 47.34        | 4.71         | 1.780        | 1.821        | 35.83        | 11.85        | 2.806        | 2.813         | 43.14        | 6.800        | 2.758        | 2.466        |
| TextFusion            | 31.35        | 17.47        | <u>5.402</u> | 6.144        | 32.43        | <u>31.68</u> | <u>8.015</u> | 8.559         | 24.60        | 22.14        | 6.690        | 6.412        |
| HAIAFusion            | 36.05        | 12.98        | 4.231        | 4.491        | 30.92        | 30.14        | 7.934        | <u>10.45</u>  | 35.69        | 19.56        | 6.882        | <b>7.422</b> |
| SeAFusion             | 27.54        | 11.09        | 3.887        | 3.610        | 26.01        | 25.04        | 7.187        | 6.993         | 28.89        | 17.75        | 6.499        | 5.285        |
| <b>USCFNet (Ours)</b> | <b>16.09</b> | <b>21.53</b> | <b>6.843</b> | <b>8.394</b> | <u>23.21</u> | <b>32.44</b> | <b>9.173</b> | <b>14.181</b> | <b>6.363</b> | <b>22.63</b> | <b>7.247</b> | 6.240        |

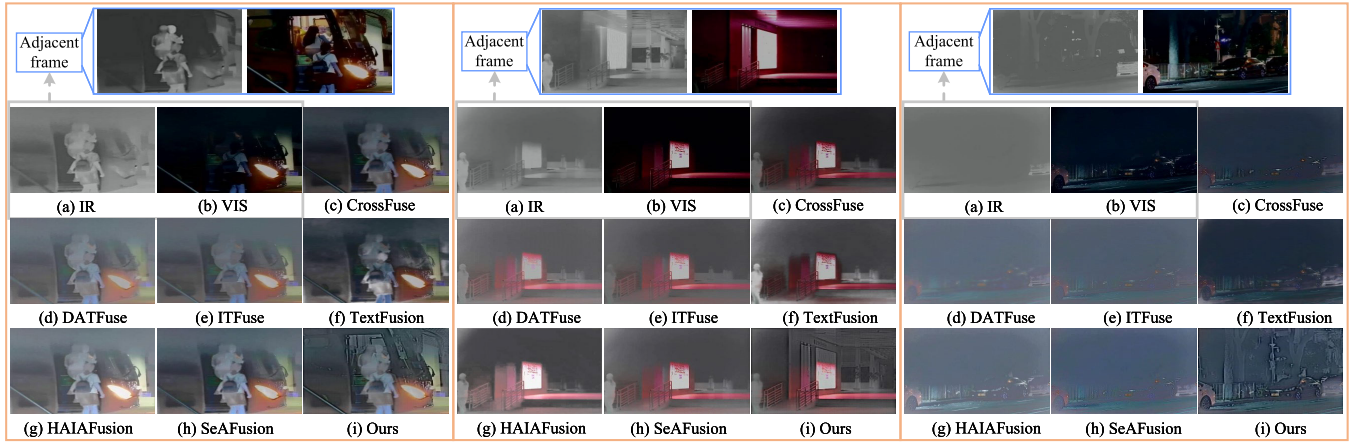


Fig. 11. Fusion performance analysis of different compared methods on low-quality frames. All data are from the curated ELVID dataset. The gray and blue boxes indicate adjacent video frames.

reflect model performance under static input conditions. As shown in Figs. 10(a) and (b), the original video frame data are used as representative examples. From the fusion results, it is evident that methods such as CrossFuse, DATFuse, ITFuse, and TextFusion exhibit varying degrees of detail loss across all three datasets. For instance, the wall contours in the ELVID dataset appear blurred, while the fence structures in HDO are missing. In addition, the fusion result of ITFuse is generally over-smoothed, while CrossFuse exhibits noticeable brightness inconsistencies and locally dark regions. SeAFusion improves the fusion performance to some extent. However, it still suffers from edge information loss in the key areas of interest (highlighted by the red, green, and yellow boxes). In contrast, both HAIAFusion and the proposed method show better fusion quality, but the former has lower contrast in the region of interest. The proposed USCFNet effectively preserves the energy characteristics of the infrared modality while retaining the texture details of the visible modality. Across different datasets, key structural elements such as the text on the sign in the red box region, the fences in the green box region, and the trees in the yellow region remain clearly discernible, and background transitions are natural, resulting in the best overall visual quality. These observations fully verify that the proposed model possesses stronger energy preservation and detail reconstruction capabilities in extremely low-light.

Table III presents the quantitative results of different fusion methods on the ELVID, M3SVD, and HDO datasets. On the ELVID dataset, the proposed USCFNet achieves the best performance across all metrics, indicating that the fused images generated by our method exhibit the lowest distortion and the highest perceptual naturalness, while maintaining clearer detail preservation and better contrast. On the M3SVD dataset, USCFNet achieves the highest values in SF, AG, and EME, demonstrating its ability to produce rich texture details and maintain balanced brightness across diverse scene sequences. Although DATFuse attains the lowest BRISQUE score, its AG and EME values are relatively low, indicating that its fused results exhibit slightly better perceptual naturalness. However, its fusion result lacks sufficient structural sharpness and contrast. On the HDO dataset, USCFNet achieves a significantly lower BRISQUE score than all other methods, while also attaining the highest SF and AG values. This further validates the proposed approach's effectiveness in detail preservation and contrast enhancement. Overall, the results demonstrate that the proposed model not only exhibits outstanding fusion performance under extremely low light but also maintains strong generalization and robustness across diverse scenes.

#### D. Comparison of Video-Based Fusion Performance

Fig. 11 illustrates the fusion performance of different methods in the low-quality frames. (a) and (b) show the original

TABLE IV

QUANTITATIVE RESULTS OF THE FUSION PERFORMANCE IN DIFFERENT METHODS UNDER LOW-LIGHT VIDEO FUSION TASKS.  $\uparrow$  INDICATES HIGHER IS BETTER,  $\downarrow$  INDICATES LOWER IS BETTER

| Methods        | FDE $\downarrow$                  | T-SSIM $\uparrow$                 | TE $\downarrow$                   | TGC $\uparrow$                    |
|----------------|-----------------------------------|-----------------------------------|-----------------------------------|-----------------------------------|
| CrossFuse      | 1.309 $\pm$ 1.457                 | 0.980 $\pm$ 0.033                 | 1.775 $\pm$ 0.652                 | 0.962 $\pm$ 0.061                 |
| DATFuse        | 1.176 $\pm$ 1.888                 | 0.981 $\pm$ 0.031                 | 1.643 $\pm$ 0.634                 | 0.956 $\pm$ 0.063                 |
| ITFuse         | <b>0.875<math>\pm</math>1.043</b> | 0.986 $\pm$ 0.021                 | <b>1.446<math>\pm</math>0.613</b> | 0.936 $\pm$ 0.062                 |
| TextFusion     | 1.406 $\pm$ 1.677                 | 0.980 $\pm$ 0.033                 | 1.828 $\pm$ 0.664                 | 0.961 $\pm$ 0.067                 |
| HAIAFusion     | 1.761 $\pm$ 2.120                 | 0.975 $\pm$ 0.040                 | 1.707 $\pm$ 0.765                 | 0.958 $\pm$ 0.065                 |
| SeAFusion      | 1.494 $\pm$ 1.585                 | 0.978 $\pm$ 0.036                 | 1.567 $\pm$ 0.705                 | 0.960 $\pm$ 0.060                 |
| USCFNet (Ours) | 0.937 $\pm$ 0.998                 | <b>0.989<math>\pm</math>0.020</b> | 1.536 $\pm$ 0.571                 | <b>0.981<math>\pm</math>0.036</b> |

low-quality frames, where the regions highlighted by blue boxes correspond to their adjacent frames. This paper aims to enhance the fusion performance of low-quality frames through the information transfer between neighboring frames using a spatiotemporal consistency network. Figs. 11(c-k) show the fusion results of different image fusion methods for various scenes. It can be observed that these methods can fuse infrared energy information and visible texture details at the single-frame level. However, due to the lack of spatiotemporal consistency modeling, these methods cannot effectively utilize the key features of adjacent frames to improve the fusion performance of low-quality frames. In contrast, the proposed USCFNet effectively leverages infrared energy and visible texture features from adjacent frames to reconstruct missing-frame information. This significantly enhances the fusion performance of low-quality frames and enables high-quality reutilization of discarded frames. These results further validate the superiority of the proposed model in spatio-temporal feature modeling.

Based on time-series data, we evaluate the fusion performance of video sequences using several temporal metrics, including frame difference energy (FDE), temporal SSIM (T-SSIM), temporal entropy (TE), and temporal gradient consistency (TGC), thereby providing an objective assessment of different methods. Specifically, FDE measures inter-frame energy fluctuations; T-SSIM evaluates structural similarity across frames; TE quantifies temporal uncertainty; and TGC assesses gradient continuity of the fused sequences. As shown in Table IV, USCFNet achieves the best performance on T-SSIM and TGC, indicating that it maintains the most stable structural consistency and the smoothest temporal edge changes during video fusion. For FDE and TE metrics, USCFNet results are slightly lower than ITFuse, but the gap is small. Notably, although ITFuse is slightly better in terms of inter-frame energy and entropy fluctuations, its structural consistency and gradient continuity are inferior to USCFNet. In summary, the spatiotemporal modeling mechanism of USCFNet effectively suppresses inter-frame fluctuations, achieving higher spatiotemporal consistency while maintaining detail stability.

### E. Discussion

1) *Spatiotemporal Consistency Heatmap Analysis:* Fig. 12 illustrates the spatiotemporal consistency heatmaps of the

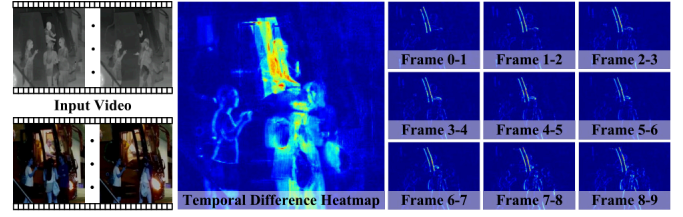


Fig. 12. The heatmaps of spatio-temporal consistency analysis. The left panel shows the original video sequence, the large image in the center presents the average inter-frame difference heatmap, and the right panel displays the frame-wise difference heatmaps.

USCFNet. In the average inter-frame difference heatmap, the background regions exhibit predominantly low difference responses (blue), while moderately high responses (yellow) appear around moving objects. This indicates that the proposed method maintains stability in static areas while preserving reasonable temporal variations in dynamic regions. The frame-wise difference heatmaps further demonstrate that the structural details remain highly consistent across consecutive frames, without noticeable flickering or spatial drift. This confirms that the proposed spatio-temporal consistency modeling effectively suppresses inter-frame instability and preserves the continuity of fine details.

2) *Object Detection Task:* To further validate USCFNet's potential for downstream object detection tasks, we conducted experiments on the ELVID dataset, as shown in Fig. 13. We used a pre-trained YOLOv11 as the object detection network to detect objects in fused images from different comparison methods. Figs. 13(c-h) shows that the comparison methods failed to detect the athletes within the red-marked areas, exhibiting significant target misses. In contrast, USCFNet successfully detected all football players in the moving scene, demonstrating higher detectability. This result indicates that USCFNet can yield more discriminative features for downstream visual tasks, such as object detection, by producing high-quality fused videos.

3) *Extremely Degraded Video Frame Analysis:* Fig. 14 illustrates the fusion performance of different methods under extreme conditions: the original infrared frame (a) appears heavily blurred, while the visible frame (b) provides almost no usable structure except for a localized bright region. Although the fusion results in Figs. 14 (c-h) integrate certain modal information, they still fail to recover fine details in the key target area highlighted by the red box. In contrast, the proposed method not only reconstructs the electronic display in the red box but also restores the staircase railing in the green box by leveraging temporal information from adjacent frames. Meanwhile, the network adaptively suppresses unreliable regions in the infrared modality. These results indicate that USCFNet, built on spatiotemporal uncertainty modeling, remains robust and effective even when handling extremely degraded low-quality video inputs.

Fig. 15 shows the fusion visualization results of different methods in scenes with severe occlusion and extremely weak textures. The first row shows the occluded scene, and the second row shows the scene with extremely weak textures.

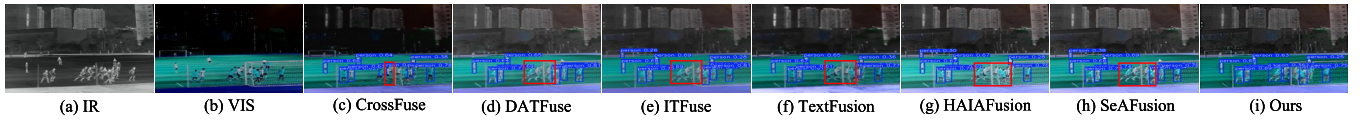


Fig. 13. Performance analysis of object detection tasks on the ELVID dataset.

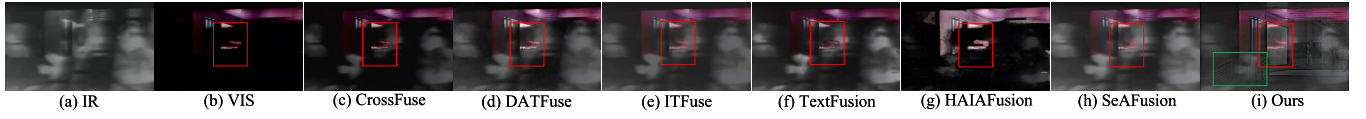


Fig. 14. Fusion performance analysis of extremely degraded video frames.

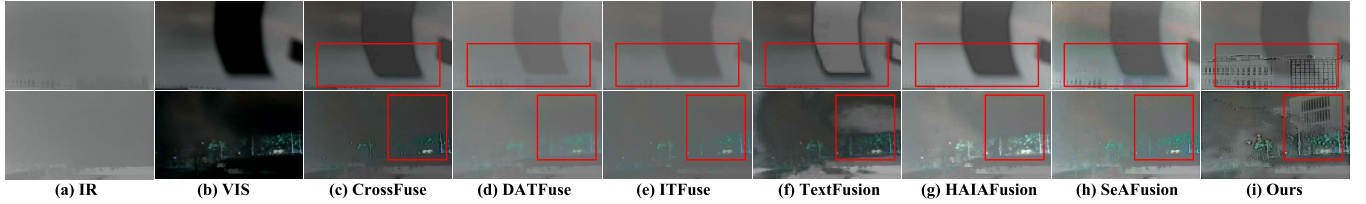


Fig. 15. Comparison of fusion performance under severe occlusion and extremely weak-texture scenes.

TABLE V  
COMPUTATIONAL EFFICIENCY COMPARISON OF DIFFERENT  
METHODS. ↓ INDICATES LOWER IS BETTER

| Methods        | Params (M)↓  | FLOPs (G)↓   | Time (ms)↓   |
|----------------|--------------|--------------|--------------|
| CrossFuse      | 1.161        | 5.202        | 15.04        |
| DATFuse        | 10.76        | <b>1.185</b> | 129.8        |
| ITFuse         | 0.082        | 5.676        | 282.735      |
| TextFusion     | <b>0.074</b> | 4.439        | 14.384       |
| HAIAFusion     | 0.591        | 38.68        | 39.41        |
| SeAFusion      | 0.167        | 10.88        | <b>4.014</b> |
| USCFNet (ours) | 36.70        | 17.90        | 20.27        |

As the infrared images are relatively blurry, the methods shown in Figs. 15 (c–e) produce fused results that are overall blurry as well. Although the methods in Figs. 15 (f–h) can better fuse infrared and visible image information, but they struggle to effectively restore the architectural details in the red-boxed area because they do not fully utilize temporal information from adjacent frames. In contrast, the proposed method can combine information from adjacent frames to mitigate interference from occlusions and low-texture scenes to some extent, thereby enabling the recovery of key region features. These results suggest that USCFNet maintains favorable spatiotemporal modeling capabilities under highly challenging conditions and shows certain advantages in recovering structures in key regions.

4) *Computational Complexity and Efficiency*: Table V compares the parameter counts, FLOPs, and average inference times of different methods. Since USCFNet incorporates a spatiotemporal relationship modeling mechanism tailored to video fusion, it has more parameters than the competing methods. Nevertheless, USCFNet requires fewer FLOPs than HATFusion and achieves shorter average inference time than DATFuse, ITFuse, and HAIAFusion. These results suggest that the proposed method enhances spatiotemporal feature modeling without introducing excessive computational

overhead. Overall, USCFNet achieves a favorable balance between fusion performance and computational complexity.

5) *Limitations and Future Directions*: While the ELVID dataset provides a new benchmark for infrared-visible video fusion under extremely low illumination conditions, its scene diversity remains limited compared to large-scale normal-illumination datasets. From a methodological perspective, the proposed USCFNet has high computational complexity, which may pose challenges to real-time performance and deployment on resource-constrained platforms. Furthermore, the framework is primarily optimized for the fusion task itself. Although the fused results show potential for downstream tasks such as object detection, quantitative evaluation on these tasks has not yet been conducted due to the lack of task-specific annotations. These limitations point to important directions for future research, including improving model efficiency and developing task-driven fusion frameworks that jointly optimize with downstream targets.

## VI. CONCLUSION

This paper proposes an uncertainty-guided spatio-temporal consistency fusion network for infrared-visible video fusion in extremely low-light conditions. Unlike conventional image-level fusion methods, USCFNet integrates spatio-temporal uncertainty analysis, a modality-specific difference-guided mechanism, and a hierarchical expert collaboration strategy within a unified framework. This design enables the model to not only extract spatiotemporal consistency features of modality-specific, but also to ensure balanced learning of information from different modalities, thereby enhancing the visual integrity of the fusion result. Furthermore, we curated a dataset called ELVID. It provides a reliable experimental benchmark for future research on extremely low-light video fusion. Experimental results demonstrate that USCFNet significantly outperforms existing methods on our curated

dataset and two public datasets. In particular, under extremely low-light conditions, it produces fused results with higher spatio-temporal consistency and better detail preservation, thereby establishing a solid basis for robust infrared-visible video fusion under extreme illumination conditions. In future work, we plan to extend the proposed framework to more diverse real-world scenarios and downstream tasks, such as object detection and tracking, while further improving model efficiency for practical deployment.

## REFERENCES

- [1] L. Tang, X. Xiang, H. Zhang, M. Gong, and J. Ma, "DIVFusion: Darkness-free infrared and visible image fusion," *Inf. Fusion*, vol. 91, pp. 477–493, Mar. 2023.
- [2] M.-J. Tan, S.-B. Gao, and S.-C. Han, "Boosting visible-infrared image fusion and target detection performance with sleep-wake joint learning," *IEEE Trans. Instrum. Meas.*, vol. 74, pp. 1–19, 2025.
- [3] H. Fu et al., "LRAF-Net: Long-range attention fusion network for visible-infrared object detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 10, pp. 13232–13245, Oct. 2024.
- [4] Y. Zhu, Q. Wang, C. Li, J. Tang, C. Gu, and Z. Huang, "Visible-thermal multiple object tracking: Large-scale video dataset and progressive fusion approach," *Pattern Recognit.*, vol. 161, May 2025, Art. no. 111330.
- [5] X. Zhang, P. Ye, H. Leung, K. Gong, and G. Xiao, "Object fusion tracking based on visible and infrared images: A comprehensive review," *Inf. Fusion*, vol. 63, pp. 166–187, Nov. 2020.
- [6] J. Nie, L. Lai, W. Wei, L. Zhang, and Y. Zhang, "Interactive visible and infrared image fusion and segmentation," *Inf. Fusion*, vol. 127, Mar. 2026, Art. no. 103832.
- [7] L. Tang, J. Yuan, and J. Ma, "Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network," *Inf. Fusion*, vol. 82, pp. 28–42, Jun. 2022.
- [8] X. Zhang and Y. Demiris, "Visible and infrared image fusion using deep learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 10535–10554, Aug. 2023.
- [9] J. Liu et al., "Infrared and visible image fusion: From data compatibility to task adaption," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 4, pp. 2349–2369, Apr. 2025.
- [10] G. Lv, X. Gao, A. Dong, Z. Wei, and J. Cheng, "SLFusion: A structure-aware infrared and visible image fusion network for low-light scenes," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 12, pp. 11877–11892, Dec. 2025, doi: [10.1109/TCSVT.2025.3586679](https://doi.org/10.1109/TCSVT.2025.3586679).
- [11] D. Liu, Y. Shi, G. Zhang, S. Li, and L. Zhang, "ZVEFusion: Zero-shot visual enhancement fusion for infrared and visible images in low light," in *Proc. ICASSP - IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2025, pp. 1–5.
- [12] J. Nie, B. Wu, W. Wei, L. Zhang, and Y. Zhang, "MCINet: Fusing low-light visible-infrared image via max-merge complementary information," *IEEE Trans. Multimedia*, vol. 27, pp. 7769–7782, 2025.
- [13] D. Hua, Q. Chen, Z. Wu, Y. Zuo, W. Wen, and Y. Fang, "Perceptual transform fusion of infrared and visible images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 12, pp. 11935–11949, Dec. 2025, doi: [10.1109/TCSVT.2025.3588203](https://doi.org/10.1109/TCSVT.2025.3588203).
- [14] Z. Wu, K. Li, H. Fan, and Y. Yang, "Drafting and revision: Advancing high-fidelity video inpainting," in *Proc. 33rd Int. Joint Conf. Artif. Intell.*, Aug. 2024, pp. 2063–2071.
- [15] Z. Wu, C. Sun, H. Xuan, K. Zhang, and Y. Yan, "Divide-and-conquer completion network for video inpainting," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 6, pp. 2753–2766, Jun. 2023.
- [16] Z. Wu, H. Xuan, C. Sun, W. Guan, K. Zhang, and Y. Yan, "Semi-supervised video inpainting with cycle consistency constraints," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 22586–22595.
- [17] Z. Wu, C. Sun, H. Xuan, and Y. Yan, "Deep stereo video inpainting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 5693–5702.
- [18] L. Tang et al., "VideoFusion: A spatio-temporal collaborative network for multi-modal video fusion," 2025, *arXiv:2503.23359*.
- [19] K. Zheng, J. Huang, M. Zhou, D. Hong, and F. Zhao, "Deep adaptive pansharpening via uncertainty-aware image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Apr. 2023, Art. no. 5403715, doi: [10.1109/T](https://doi.org/10.1109/T)
- [20] X. Guo, X. Lin, X. Yang, L. Yu, K.-T. Cheng, and Z. Yan, "UCTNet: Uncertainty-guided CNN-transformer hybrid networks for medical image segmentation," *Pattern Recognit.*, vol. 152, Aug. 2024, Art. no. 110491.
- [21] C. Riquelme et al., "Scaling vision with sparse mixture of experts," in *Proc. NIPS*, 2021, pp. 8583–8595.
- [22] Y. Chen, L. Ren, H. Xia, Z. Wang, C. Gao, and F. Wang, "A compound fault diagnosis method based on multi-task learning with multi-gate mixture-of-experts," in *Proc. 14th Int. Conf. Measuring Technol. Mechatronics Autom. (ICMTMA)*, Jan. 2022, pp. 281–285.
- [23] W. Sun et al., "HotMoE: Exploring sparse mixture-of-experts for hyperspectral object tracking," *IEEE Trans. Multimedia*, vol. 27, pp. 4072–4083, 2025.
- [24] Y. Sun et al., "Task-gated multi-expert collaboration network for degraded multi-modal image fusion," in *Proc. Int. Conf. Mach. Learn.*, 2025, pp. 57571–57586.
- [25] B. Cao, Y. Sun, P. Zhu, and Q. Hu, "Multi-modal gated mixture of local-to-global experts for dynamic image fusion," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 23555–23564.
- [26] S. Karim, G. Tong, J. Li, A. Qadir, U. Farooq, and Y. Yu, "Current advances and future perspectives of image fusion: A comprehensive review," *Inf. Fusion*, vol. 90, pp. 185–217, Feb. 2023.
- [27] M. Amin-Naji, A. Aghagolzadeh, and M. Ezoji, "Ensemble of CNN for multi-focus image fusion," *Inf. Fusion*, vol. 51, pp. 201–214, Nov. 2019.
- [28] X. Liu, C. Deng, J. Chanussot, D. Hong, and B. Zhao, "StfNet: A two-stream convolutional neural network for spatiotemporal image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6552–6564, Sep. 2019.
- [29] C. Zhao et al., "MHW-GAN: Multidiscriminator hierarchical wavelet generative adversarial network for multimodal image fusion," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 10, pp. 13713–13727, Oct. 2024.
- [30] J. Ma, H. Xu, J. Jiang, X. Mei, and X.-P. Zhang, "DDcGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 4980–4995, 2020.
- [31] H. Xu, J. Ma, and X.-P. Zhang, "MEF-GAN: Multi-exposure image fusion via generative adversarial networks," *IEEE Trans. Image Process.*, vol. 29, pp. 7203–7216, 2020.
- [32] Z. Wu et al., "WaveFormer: Wavelet transformer for noise-robust video inpainting," in *Proc. AAAI Conf. Artif. Intell.*, 2024, vol. 38, no. 6, pp. 6180–6188.
- [33] Y. Liu, C. Yu, J. Cheng, Z. J. Wang, and X. Chen, "MM-net: A MixFormer-based multi-scale network for anatomical and functional image fusion," *IEEE Trans. Image Process.*, vol. 33, pp. 2197–2212, 2024.
- [34] Z. Wu, K. Chen, K. Li, H. Fan, and Y. Yang, "BVINet: Unlocking blind video inpainting with zero annotations," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2025, pp. 14017–14027.
- [35] Y. Xian, Y. Xu, Y. He, and Y. Yang, "From 2D grids to 1D tokens: Reforming shared representations for multimodal image fusion," 2026, *arXiv:2606.12303*.
- [36] J. Yue, L. Fang, S. Xia, Y. Deng, and J. Ma, "Dif-fusion: Toward high color fidelity in infrared and visible image fusion with diffusion models," *IEEE Trans. Image Process.*, vol. 32, pp. 5705–5720, 2023.
- [37] Y. Shi, Y. Liu, J. Cheng, Z. J. Wang, and X. Chen, "VDMUFusion: A versatile diffusion model-based unsupervised framework for image fusion," *IEEE Trans. Image Process.*, vol. 34, pp. 441–454, 2025.
- [38] P. Liang, J. Jiang, Q. Ma, C. Wang, X. Liu, and J. Ma, "FusionINV: A diffusion-based approach for multimodal image fusion," *IEEE Trans. Image Process.*, vol. 34, pp. 5355–5368, 2025.
- [39] A. Toet, "The TNO multiband image data collection," *Data Brief*, vol. 15, pp. 249–251, Dec. 2017.
- [40] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2Fusion: A unified unsupervised image fusion network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 502–518, Jan. 2022.
- [41] X. Jia, C. Zhu, M. Li, W. Tang, and W. Zhou, "LLVIP: A visible-infrared paired dataset for low-light vision," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 3496–3504.
- [42] L. Tang, J. Yuan, H. Zhang, X. Jiang, and J. Ma, "PIAFusion: A progressive infrared and visible image fusion network based on illumination

- [43] J. Chen, L. Yang, W. Liu, X. Tian, and J. Ma, "LENFusion: A joint low-light enhancement and fusion network for nighttime infrared and visible image fusion," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–15, 2024.
- [44] H. Xie, M. Sang, Y. Zhang, Y. Yang, S. Zhao, and J. Zhong, "RCVS: A unified registration and fusion framework for video streams," *IEEE Trans. Multimedia*, vol. 26, pp. 11031–11043, 2024.
- [45] Y. Han et al., "Seq-IF: Sequentially consistent infrared-visible video fusion under time-varying illumination for perception enhancement," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 36, no. 7, pp. 9765–9779, Jul. 2026, doi: [10.1109/TCSVT.2026.3669445](https://doi.org/10.1109/TCSVT.2026.3669445).
- [46] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
- [47] J. Hou, X. Wu, R. Wang, J. Luo, and Y. Jia, "Confidence-guided self refinement for action prediction in untrimmed videos," *IEEE Trans. Image Process.*, vol. 29, pp. 6017–6031, 2020.
- [48] N. Shazeer et al., "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," 2017, *arXiv:1701.06538*.
- [49] B. Yang, J. Zhang, R. Liu, D. Lin, P. Li, and C. L. P. Chen, "Point-to-set metric-gated mixture of experts for multisource domain adaptation fault diagnosis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 8, pp. 14893–14907, Aug. 2025.
- [50] X. Zhu, Y. Wang, Y. Ma, Z. Zhang, and Y. Gao, "CognitMoE: A cognition-aware collaborative multi-expert network for bipolar disorder diagnosis," *Neural Netw.*, vol. 192, Dec. 2025, Art. no. 107854.
- [51] Z. Sun et al., "FlexiReID: Adaptive mixture of expert for multi-modal person re-identification," 2025, *arXiv:2510.15595*.
- [52] Y. Liu, Y. Xu, Y. Wei, X. Bi, and B. Xiao, "Clear nights ahead: Towards multi-weather nighttime image restoration," in *Proc. AAAI Conf. Artif. Intell.*, 2026, vol. 40, no. 9, pp. 7395–7403.
- [53] Y. Xing, Z. Niu, S. Yang, S. Zhang, and Y. Zhang, "Less is more: Infrared and visible images fusion via semantic-guided mixture of multi-feature experts," *IEEE Trans. Image Process.*, vol. 35, pp. 3381–3394, 2026.
- [54] P. Lindenberger, P. E. Sarlin, and M. Pollefeys, "LightGlue: Local feature matching at light speed," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 17627–17638.
- [55] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
- [56] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, Mar. 2002.
- [57] A. M. Eskicioglu and P. S. Fisher, "Image quality measures and their performance," *IEEE Trans. Commun.*, vol. 43, no. 12, pp. 2959–2965, Dec. 1995.
- [58] S. S. Agaian, K. Panetta, and A. M. Grigoryan, "Transform-based image enhancement algorithms with performance measure," *IEEE Trans. Image Process.*, vol. 10, no. 3, pp. 367–382, Mar. 2001.
- [59] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.
- [60] Z. Wang, W. Shao, Y. Chen, J. Xu, and L. Zhang, "A cross-scale iterative attentional adversarial fusion network for infrared and visible images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 8, pp. 3677–3688, Aug. 2023.
- [61] W. Tang, F. He, Y. Liu, Y. Duan, and T. Si, "DATFuse: Infrared and visible image fusion via dual attention transformer," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 7, pp. 3159–3172, Jul. 2023.
- [62] W. Tang, F. He, and Y. Liu, "ITFuse: An interactive transformer for infrared and visible image fusion," *Pattern Recognit.*, vol. 156, Dec. 2024, Art. no. 110822.
- [63] C. Cheng et al., "TextFusion: Unveiling the power of textual semantics for controllable image fusion," *Inf. Fusion*, vol. 117, May 2025, Art. no. 102790.
- [64] Y. Sun, M. Dong, M. Yu, and L. Zhu, "HAIAFusion: A hybrid attention illumination-aware framework for infrared and visible image fusion," *IEEE Trans. Instrum. Meas.*, vol. 74, pp. 1–19, 2025.