S2F-Net: Shared-Specific Fusion Network for Infrared and Visible Image Fusion

Yijing Zhao[1,2] Yuchao Xia[3] Yi Ding[3] Yumeng Liu[1] Shuai Liu[1,2] Hongan Wang[1,2]

[1]Institute of Software, Chinese Academy of Sciences [2]University of Chinese Academy of Sciences

[3]University of Electronic Science and Technology of China

Abstract

A modality gap exists between infrared and visible images, presenting challenges for image fusion. Despite the modality heterogeneity, both types of images inherently capture the same scene, suggesting the presence of common information. Effectively extracting shared features while distinguishing modality-specific ones is pivotal for bridging this gap and achieving superior fusion outcomes. To address this, we propose the **S**hared-**S**pecific **F**usion **Net**work (S2F-Net). The S2F-Net introduces a three-branch feature extractor, which retains two branches for extracting features from each modality, innovatively creating an additional branch dedicated to facilitating the separation of shared features from modality-specific ones. This facilitates guiding the fusion of cross-modal information to generate efficient fusion features, ensuring the effective integration of complementary information from different modalities. To achieve feature fusion and image reconstruction, we propose two fusion modules: the Cross-modality Attention-Guided Fusion Module (CAGFM) and the Multi-Level Fusion Module (MLFM). The former utilizes shared and specific features by employing cross-modality channel attention, enabling effective integration of information across modalities. The latter facilitates feature interaction across different levels. Additionally, to effectively disentangle shared and specific features, we introduce the shared-specific learning module. Extensive experiments conducted on open-source datasets validate the superior performance of our proposed method.

Contributions

Our contributions can be summarized as follows:

- This paper proposes a novel three-branch multi-level Shared-Specific Fusion Network (S2F-Net) designed to extract both shared and specific features from two modalities. Leveraging a shared-specific learning module, the network effectively disentangles these features, facilitating superior fusion outcomes. To the best of our knowledge, this is the first three-branch network tailored for image fusion.
- A novel fusion strategy composed of two main modules is proposed. The Cross-modality Attention-Guided Fusion Module (CAGFM) fuses shared and specific features at the same level using cross-modality channel attention. The Multi-Level Fusion Module (MLFM) integrates features across different levels to enhance information exchange and interaction.
- Experiments on open-source datasets demonstrate the effectiveness of S2F-Net compared to other state-of-the-art (SOTA) methods. Comprehensive ablation experiments further confirm the effectiveness of the proposed modules.

Methodology

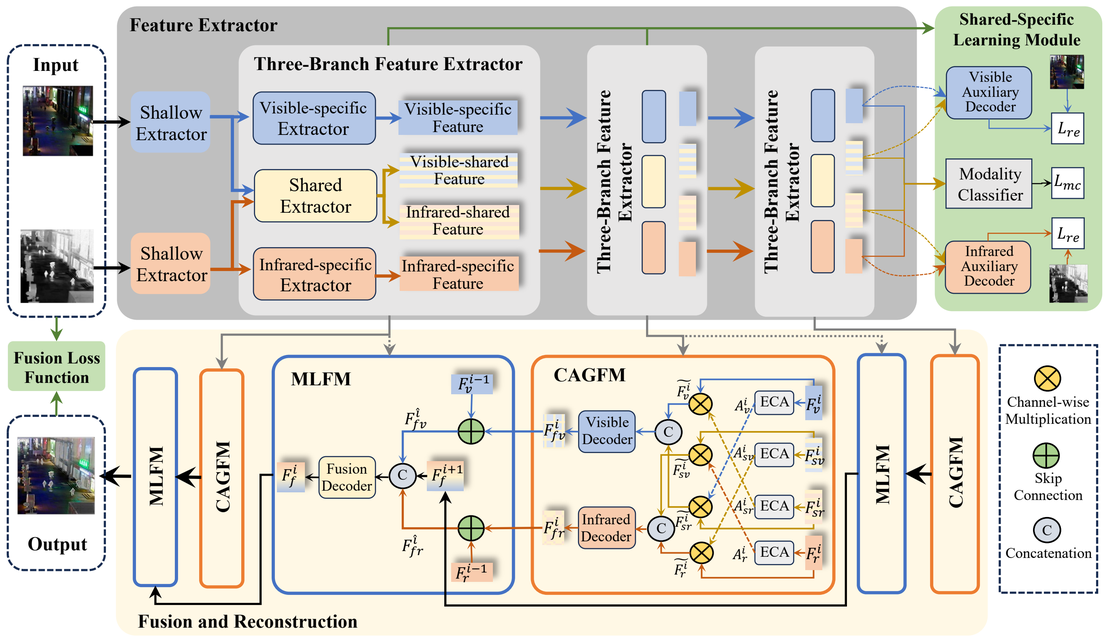

Figure 1. The framework of the proposed S2F-Net for infrared and visible image fusion

Framework and Three-Branch Feature Extractor

For the multi-modal image fusion task involving infrared and visible images, the objective is to generate a fused image I_f given a paired infrared image I_r and visible image I_v as input. This task involves three steps: feature extraction, feature fusion, and image reconstruction.

At i^{th} level, three branches exist. The novel branch $E_s^i(\cdot)$ extracts shared features from two modalities, with subscript s indicating "shared". These features from the two modalities are represented as F_{sv}^i for the shared-visible feature and F_{sr}^i for the shared-infrared feature. The other two branches are denoted as $E_r^i(\cdot)$ and $E_v^i(\cdot)$, responsible for extracting the specific features of each modality. The feature extraction at the i^{th} level is expressed as:

$$\begin{aligned} F_v^i &= E_v^i(F_v^{i-1}), \\ F_r^i &= E_r^i(F_r^{i-1}), \\ \{F_{sv}^i, F_{sr}^i\} &= E_s^i(F_{sv}^{i-1}, F_{sr}^{i-1}) \end{aligned} \tag{1}$$

Superscript $i-1$ denotes features from the prior level, while F_v^i and F_r^i represent the i^{th} specific features for visible and infrared. At the first level, without prior shared features, the shared branch uses shallow-level infrared and visible features as inputs.

$$\{F_{sv}^1, F_{sr}^1\} = E_s^1(F_v^{SF}, F_r^{SF}) \tag{2}$$

Cross-modality Attention-Guided Fusion Module

CAGFM utilizes Efficient Channel Attention to extract channel attention from the four features at each level. Channel attention is denoted as A_m^i, where the subscript m denotes the source feature type and belongs to the set $\{v, r, sv, sr\}$.

$$\begin{aligned} \tilde{F_v^i} &= F_v^i \odot A_{sr}^i, \tilde{F_{sr}^i} = F_{sr}^i \odot A_v^i \\ \tilde{F_r^i} &= F_r^i \odot A_{sv}^i, \tilde{F_{sv}^i} = F_{sv}^i \odot A_r^i \end{aligned} \tag{3}$$

Following this, shared features from one modality are channel-wise concatenated with the specific features from another modality. Using two decoders $\mathcal{D}_v$ and $\mathcal{D}_r$, the concatenated feature maps are downscaled to match the same dimension as the features at the $i-1$ level. The decoder facilitates cross-modal shared and specific information interaction.

$$\begin{aligned} F_{fv}^i &= \mathcal{D}_v^i(concat(\tilde{F_v^i}, \tilde{F_{sr}^i})), \\ F_{fr}^i &= \mathcal{D}_r^i(concat(\tilde{F_r^i}, \tilde{F_{sv}^i})) \end{aligned} \tag{4}$$

Multi-Level Fusion Module

MLFM is designed to foster inter-level information exchange. At i^{th} level, MLFM combines the fused features generated by CAGFM with the fusion results from the $i+1^{th}$ level and the F_v^{i-1} and F_r^{i-1} features obtained in $i-1^{th}$ level.

Initially, skip connections are used to element-wise add F_v^{i-1} to F_{fv}^i, and F_r^{i-1} to F_{fr}^i. The element-wise added fusion features are denoted as $\hat{F_{fr}^i}$ and $\hat{F_{fv}^i}$, as summarized in the following equation:

$$\hat{F_{fv}^i} = F_{fv}^i \oplus F_v^{i-1}, \hat{F_{fr}^i} = F_{fr}^i \oplus F_r^{i-1} \tag{5}$$

Skip connected features are concatenated with the $i+1^{th}$ level's fusion feature along the channel dimension, and then input into decoder $\mathcal{D}_f$ to get F_f^i, the final reconstructed fusion feature of the i^{th} level.

$$F_f^i = \mathcal{D}_f^i\{concat(\hat{F_{fv}^i}, F_f^{i+1}, \hat{F_{fr}^i})\} \tag{6}$$

Shared-Specific Learning Module

Modality Classifier. This ensures that the entire feature extraction network maximizes the distance between shared and specific features while minimizing the distance between shared features from the two modalities.

$$\mathcal{L}_{mc} = \mathbb{E}_m[-log(p(m|F_m^i, \Theta_{mc}))] \tag{7}$$

Auxiliary Decoder. The auxiliary decoder network $\mathcal{D}e_m$ is used, where m denotes the modality reconstructed. The shared and specific features of the same modality are concatenated along the channel dimension and input into the decoder for reconstruction.

$$\hat{I_v^i} = \mathcal{D}e_v(concat(F_v^i, F_{sv}^i)), \hat{I_r^i} = \mathcal{D}e_r(concat(F_r^i, F_{sr}^i)) \tag{8}$$

$$\mathcal{L}_{re} = \sum_i \mathcal{L}_2(\hat{I_v^i}, I_v) + \sum_i \mathcal{L}_2(\hat{I_r^i}, I_r) \tag{9}$$

Results

We conducted experiments on three different publicly available datasets (VTUAV, MSRS and LLVIP) to demonstrate the superior performance of S2F-Net, showing its adaptability across various scenarios.

Quality Comparison

The first row demonstrates S2F-Net's superior visual performance in nighttime scenes by effectively integrating modalities and clearly distinguishing pedestrians. The second row depicts a challenging scene with a slightly obscured small target, a pedestrian. S2F-Net preserves significant information about the pedestrian in the infrared modality as much as possible, while other methods tend to blur the pedestrian or struggle to distinguish the pedestrian from the background. In the third row, the nighttime visible image reveals a detailed telephone booth that is absent in the infrared image. S2F-Net disentangles shared and specific features and leverages visible-specific features to mitigate distortions caused by the missing telephone booth in the infrared image.

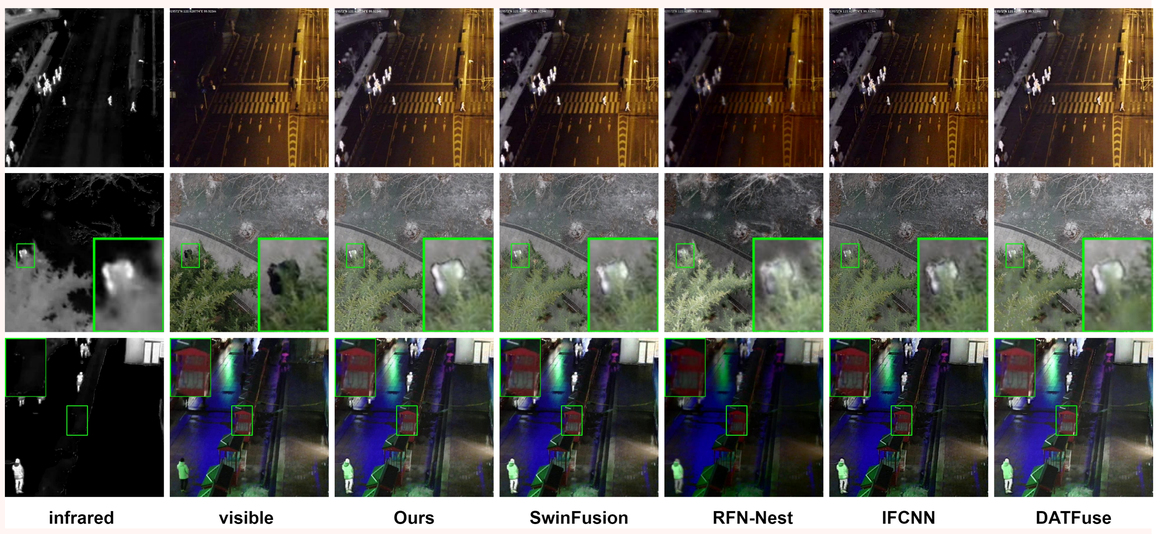

Figure 2. Quantitative comparison with state-of-the-arts in the UAV-view image fusion scenarios

Quantitative Comparison

In experiments on the LLVIP dataset, to facilitate comparison with previous methods, we selected four metrics used in prior studies: average gradient (AG), correlation coefficient (CC), sum of the correlations of differences (SCD), and Q_{abf}. Table 1 presents the experimental results on the LLVIP dataset. Even with the change in evaluation metrics, S2F-Net still achieves outstanding results. The prominence of AG and Q_{abf} indicates S2F-Net's excellence in detail processing, while the excellent performance in CC suggests that S2F-Net has learned relevant information for fusion from different modalities as much as possible.

Table 1. Quantitative comparisons on LLVIP

Method	AG↑	CC↑	SCD↑	Q_{abf} ↑
DenseFuse	3.2640	<u>0.7187</u>	1.2110	0.3094
DIDFuse	3.4474	0.6509	1.2575	0.2436
IFCNN	5.4137	0.6910	1.4270	0.5845
NestFuse	4.5400	0.6663	1.4816	0.4847
AUIF	3.8589	0.6945	1.2840	0.2765
RFN-Nest	2.7853	0.7138	1.4612	0.2288
SeAFusion	5.6836	0.6779	<u>1.5803</u>	0.5284
DetFusion	<u>6.1824</u>	0.7086	**1.6497**	<u>0.5899</u>
Ours	**6.7056**	**0.8371**	1.3979	**0.6288**

Conclusions

In this paper, we introduce S2F-Net, a novel three-branch method tailored for infrared and visible image fusion. S2F-Net is adept at extracting both shared and modality-specific features, rendering it adaptable to complex scenarios. Leveraging the innovative shared branch, we propose two fusion modules, namely CAGFM (Cross-modality Attention-Guided Fusion Module) and MLFM (Multi-Level Fusion Module), which are designed to effectively fuse information from the two modalities. Extensive experiments conducted on open-source datasets validate the superior performance of our proposed method.

contact: yijing@iscas.ac.cn