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

Yijing Zhao1,2, Yuchao Xia3, Yi Ding3, Yumeng Liu2, Shuai Liu1,2, Hongan Wang1,2
1Institute of System Science, Chinese Academy of Sciences
2University of Electronic Science and Technology of China
3University of Chinese Academy of Sciences

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 Shared-Specific Fusion Network (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 a cross-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 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 methods. Comprehensive ablation experiments further confirm the effectiveness of the proposed modules.

Methodology

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 ̂F from a given infrared image Ir and visible image Iv as input. This task involves three steps: feature extraction, feature fusion, and image reconstruction.

At the 3rd level, three branches exist. The novel branch Fc3,i(i) extracts shared features from two modalities, with subscript i indicating “shared”. These features from the two modalities are represented as Fc3,i for the shared-visible feature and Fc3,v for the shared-infrared feature. The other two branches are denoted as Fv,i and Fr,i, responsible for extracting the specific features of each modality. The feature extraction at the 3rd level is expressed as:

\[
F_c = F_c(F_v, F_r)
\]

Superscript i = 1 denotes features from the prior level, while Fc,i and Fr,i represent the specific features for visible and infrared. At the 1st level, without prior shared features, the shared branch uses shallow-level infrared and visible features as inputs. The Cross-modality Attention-Guided Fusion Module

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

Following this, shared features from each modality are channel-wise concatenated with the specific features from another modality. Using two decoders Dc and Ds, the concatenated feature maps are downsampled to match the same dimension as the features at the i-1 level. The decoder facilitates cross-modal shared and specific information interaction.

Multi-Level Fusion Module

MLFM is designed to foster inter-level information exchange. At the 4th level, MLFM combines the fused features generated by CAGFM with the fusion results from the i + 1th level and the Fc,i and Fr,i features obtained in i - 1th level.

\[
F_c = F_c(F_v, F_r)
\]

Auxiliary Decoder

The auxiliary decoder network Daux 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.

\[
D_m = D_m(D_c(F_v, F_r), D_m(F_v, F_r))
\]

\[
L_m = \mathbb{E}_{F_m}[-\log(D_m(F_v, F_r))]
\]

Results

We conducted experiments on three different publicly available datasets (UTAV, MSRS, and LUVIP) 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 obscure street 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.

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.