CDDFuse: Correlation-Driven Dual-Branch Feature Decomposition for Multi-Modality Image Fusion

Zixiang Zhao[1,2] Haowen Bai[1] Jiangshe Zhang[1*] Yulun Zhang[2]
Shuang Xu[3,4] Zudi Lin[5] Radu Timofte[2,6] Luc Van Gool[2]

[1] Xi'an Jiaotong University [2] Computer Vision Lab, ETH Zürich
[3] Research and Development Institute of Northwestern Polytechnical University in Shenzhen
[4] Northwestern Polytechnical University [5] Harvard University [6] University of Würzburg

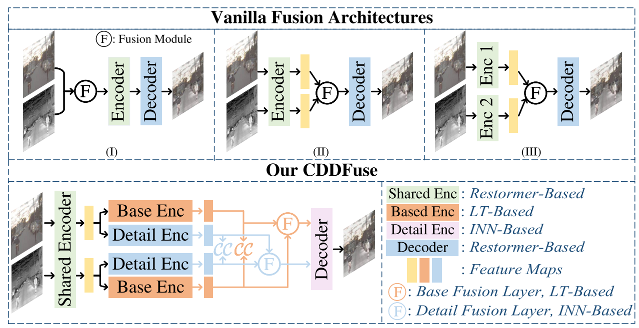
zixiangzhao@stu.xjtu.edu.cn, jszhang@mail.xjtu.edu.cn

Abstract

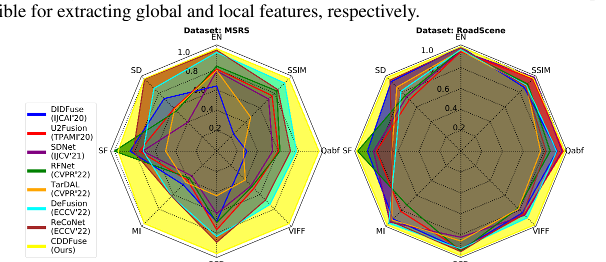
Multi-modality (MM) image fusion aims to render fused images that maintain the merits of different modalities, e.g., *functional highlight and detailed textures. To tackle the challenge in modeling cross-modality features and decomposing desirable modality-specific and modality-shared features, we propose a novel Correlation-Driven feature Decomposition Fusion (**CDDFuse**) network. Firstly, CDDFuse uses Restormer blocks to extract cross-modality shallow features. We then introduce a dual-branch Transformer-CNN feature extractor with Lite Transformer (LT) blocks leveraging long-range attention to handle low-frequency global features and Invertible Neural Networks (INN) blocks focusing on extracting high-frequency local information. A correlation-driven loss is further proposed to make the low-frequency features correlated while the high-frequency features uncorrelated based on the embedded information. Then, the LT-based global fusion and INN-based local fusion layers output the fused image. Extensive experiments demonstrate that our CDDFuse achieves promising results in multiple fusion tasks, including infrared-visible image fusion and medical image fusion. We also show that CDDFuse can boost the performance in downstream infrared-visible semantic segmentation and object detection in a unified benchmark. The code is available at https://github.com/Zhaozixiang1228/MMIF-CDDFuse.*

1. Introduction

Image fusion is a basic image processing topic that aims to generate informative fused images by combining the important information from source ones [47,75,78,79,87]. The fusion targets include digital [45,84], multi-modal [36,70,88]



(a) Existing MMIF methods *vs.* CDDFuse. The *base* and *detail* encoders are responsible for extracting global and local features, respectively.



(b) CDDFuse (highlighted in yellow) achieves the state-of-the-art performance on MSRS [57] and RoadScene [71] in eight metrics.

Figure 1. Workflow and performance comparison of our proposed CDDFuse with existing MM image fusion approaches.

and remote sensing [4, 76, 91] images, *etc*. The *Infrared-Visible image Fusion* (IVF) and *Medical Image Fusion* (MIF) are two challenging sub-categories of *Multi-Modality Image Fusion* (MMIF), focusing on modeling the cross-modality features from all the sensors and aggregating them into the output images. Specifically, IVF targets fused images that preserve thermal radiation information in the input infrared images and detailed texture information in the input visible images. The fused images can avoid the shortcomings of visible images being sensitive to illumination conditions as well as the infrared images being noisy and low-resolution. Downstream recognition tasks, *e.g.*, multi-modal saliency detection [33, 52, 63], object detection [6, 34, 58] and se-

*Corresponding author.

mantic segmentation [37, 49–51] can then benefit from the obtained clearer representations of scenes and objects in IVF images. Similarly, MIF aims to clearly exhibit the abnormalities by fusing multiple medical imaging modalities to reveal comprehensive information to assist diagnosis and treatment [21].

Many methods have been developed to tackle the MMIF challenges in recent years [35, 41, 44, 55, 73, 74, 82]. A common pipeline that demonstrated promising results utilizes CNN-based feature extraction and reconstruction in an Auto-Encoder (AE) manner [28, 29, 32, 88]. The workflow is illustrated in Fig. 1a. However, existing methods have three main shortcomings. First, the internal working mechanism of CNNs is difficult to control and interpret, causing insufficient extraction of cross-modality features. For example, in Fig. 1a, shared encoders in (I) and (II) cannot distinguish modality-specific features, while the private encoders in (III) ignore features shared by modalities. Second, the context-independent CNN only extracts local information in a relatively small receptive field, which can hardly extract global information for generating high-quality fused images [31]. Thus, it is still unclear whether the inductive biases of CNN are capable enough to extract features for all modalities. Third, the forward propagation of fusion networks often causes the loss of high-frequency information [42, 89]. Our work explores a more reasonable paradigm to tackle the challenges in feature extraction and fusion.

First, we aim to add correlation restrictions to the extracted features and limit the solution space, which improves the controllability and interpretability of feature extraction. Our assumption is that, in the MMIF task, the input features of the two modalities are correlated at low frequencies, representing the modality-shared information, while the high-frequency feature is irrelevant and represents the unique characteristics of the respective modalities. Taking IVF as an example, since infrared and visible images come from the same scene, the low-frequency information of the two modalities contains statistical co-occurrences, such as background and large-scale environmental features. On the contrary, the high-frequency information of the two modalities is independent, *e.g*., the texture and detail information in the visible image and the thermal radiation information in the infrared image. Therefore, we aim to facilitate the extraction of modality-specific and modality-shared features by increasing and decreasing the correlation between low-frequency and high-frequency features, respectively.

Second, from the architectural perspective, Vision Transformers [14, 38, 80] recently shows impressive results in computer vision, with self-attention mechanism and global feature extraction. However, Transformer-based methods are computationally expensive, which leaves room for further improvement with considering the efficiency-performance tradeoff of image fusion architectures. Therefore, we propose integrating the advantages of local context extraction and computational efficiency in CNN and the advantages of global attention and long-range dependency modeling in Transformer to complete the MMIF task.

Third, to solve the challenge of losing wanted high-frequency input information, we adopt the building block of *Invertible Neural networks* (INN) [13]. INN was proposed with invertibility by design, which prevents information loss through the mutual generation of input and output features and aligns with our goal of preserving high-frequency features in the fused images.

To this end, we proposed the *Correlation-Driven feature Decomposition Fusion* (**CDDFuse**) model, where modality-specific and modality-shared feature extractions are realized by a dual-branch encoder, with the fused image reconstructed by the decoder. The workflow is shown in Figs. 1a and 2. Our contributions can be summarized in four aspects:

- We propose a dual-branch Transformer-CNN framework for extracting and fusing global and local features, which better reflects the distinct modality-specific and modality-shared features.
- We refine the CNN and Transformer blocks for a better adaptation to the MMIF task. Specifically, we are the first to utilize the INN blocks for lossless information transmission and the LT blocks for trading-off fusion quality and computational cost.
- We propose a correlation-driven decomposition loss function to enforce the modality shared/specific feature decomposition, which makes the cross-modality base features correlated while decorrelates the detailed high-frequency features in different modalities.
- Our method achieves leading image fusion performance for both IVF and MIF. We also present a unified measurement benchmark to justify how the IVF fusion images facilitate downstream MM object detection and semantic segmentation tasks.

2. Related Work

This section briefly reviews the representative works of deep learning (DL)-based multi-modal image fusion (MMIF) approaches, and the Vision Transformer (LT, Restormer) as well as INN modules employed in CDDFuse.

2.1. DL-based multi-modal image fusion

In the era of DL, CNN-based models for MMIF can be categorized into four main classes: generative adversarial network (GAN)-based models [39, 42, 43], AE-based models [28, 29, 37, 82], unified models [26, 70, 71, 83, 85] and algorithm unrolling models [11, 15, 77, 89]. In GAN-based models, GAN [18, 46, 48] are utilized to simultaneously make the fusion images distributionally similar to inputs and perceptually satisfactory. AE-based methods can be regraded as the DL variant of transformation models with replacing the transformers and inverse transformers with encoders and decoders [88]. Through cross-task learning, the unified models can alleviate the problems of limited training data and missing ground-truth [83]. Algorithm unrolling models build a bridge between traditional optimization and DL methods, and establish model-driven interpretable CNN frameworks [90]. Recently, considering the combination of fusion and downstream pattern recognition tasks, Liu *et al*. [35] pioneered the exploration of the combination of image fusion and detection. Then the gradient of the loss function for segmentation [56] and detection [54] is used to guide the generation of fused images. Liang *et al*. [32] propose a self-supervised learning framework to complete the fusion task without paired images. Additionally, adding a pre-processing registration module before the fusion module is proved to solve the misregistration of source images effectively [19, 62, 72]. Jiang *et al*. [22] firstly developed a multi-view and multi-modality fusion based stitching method for comprehensive scene perception.

2.2. Vision transformer and variants

Transformer, firstly proposed by Vaswani *et al*. [61] for natural language processing (NLP) and ViT [14] for computer vision. Then numerous transformer-based models have gained satisfied results in classification [38, 60], object detection [8, 96], segmentation [64, 92] and multi-modal learning [25, 86]. For low-level vision tasks, Transformer combining with multi-task learning [9] and Swin Transformer block [31, 38] has achieved advanced results compared to CNN-based methods. Other advanced networks also obtain competitive results in various inverse problems [7, 16, 30, 66].

Considering the large computational overhead of spatial self-attention, Wu *et al*. [67] proposed a lightweight LT structure for mobile NLP tasks. Through Long-Short Range Attention and the Flattened feed-forward network, the amount of parameters is largely reduced while maintaining the model performance. Restormer [80] improves transformer block by the gated-Dconv network and multi-Dconv head attention transposed modules, which facilitate multi-scale local-global representation learning on high-resolution images. We adopt LT and Restormer blocks into our CDDFuse model.

2.3. Invertible neural networks

The invertible neural network is an important module of the Normalized Flow model, a popular kind of generative model [2]. It was first proposed by NICE [12], and later the additive coupling layer in NICE was replaced by the coupling layers in RealNVP [13]. Subsequently, 1×1 invertible convolution was used in Glow [27], which can generate realistic high-resolution images. INNs are also applied to classification tasks to save memory and improve the features extraction ability of the backbone [5, 17, 20]. Because of its lossless information-preserving property, INNs have been effectively integrated into image processing fields such as image coloring [3], image hiding [23], image rescaling [68] and image/video super-resolution [94, 95].

2.4. Comparison with existing approaches

The existing methods most relevant to our model are the AE-based methods. Compared with the conventional AE methods, our CDDFuse model that extracts local and long-range features with different structures is more reasonable and intuitive than a pure CNN framework. In addition, our proposed correlation-based decomposition loss can effectively suppress redundant information and improve the quality of extracted features than traditional loss functions.

3. Method

In this section, we first introduce the workflow of CDDFuse and the detailed structure of each module. For simplicity, we denote low-frequency long-range features as the *base* features and high-frequency local features as the *detail* features in the following discussion.

3.1. Overview

Our CDDFuse contains four modules, *i.e*., a dual-branch encoder for feature extraction and decomposition, a decoder for reconstructing original images (in training stage I) or generating fusion images (in training stage II), and the base/detail fusion layer to fuse the different frequency features, respectively. The detailed workflow is illustrated in Fig 2. Note that CDDFuse is a generic multi-modal image fusion network, and we only take the IVF task as an example to explain the working of CDDFuse.

3.2. Encoder

The encoder has three components: the Restormer block [80]-based *share feature encoder* (SFE), the Lite Transformer (LT) block [67]-based *base transformer encoder* (BTE) and the Invertible Neural networks (INN) block [13]-based *detail CNN encoder* (DCE). The BTE and DCE together form the *Long-short Range Encoder*.

First, we define some symbols for clarity in formulation. The input paired infrared and visible images are denoted as $I\in\mathbb{R}^{H\times W}$ and $V\in\mathbb{R}^{H\times W\times 3}$. The SFE, BTE and DCE are represented by $\mathcal{S}(\cdot)$, $\mathcal{B}(\cdot)$ and $\mathcal{D}(\cdot)$, respectively.

Share feature encoder. SFE aims to extracts shallow features $\{\Phi_I^S, \Phi_V^S\}$ from infrared and visible inputs $\{I, V\}$, *i.e*.,

$$\Phi_I^S = \mathcal{S}(I),\ \Phi_V^S = \mathcal{S}(V). \tag{1}$$

The reason we choose Restormer block in SFE is that Restormer can extract global features from high-resolution input images by applying self-attention across feature dimension [81]. Therefore, it can extract cross-modality shallow

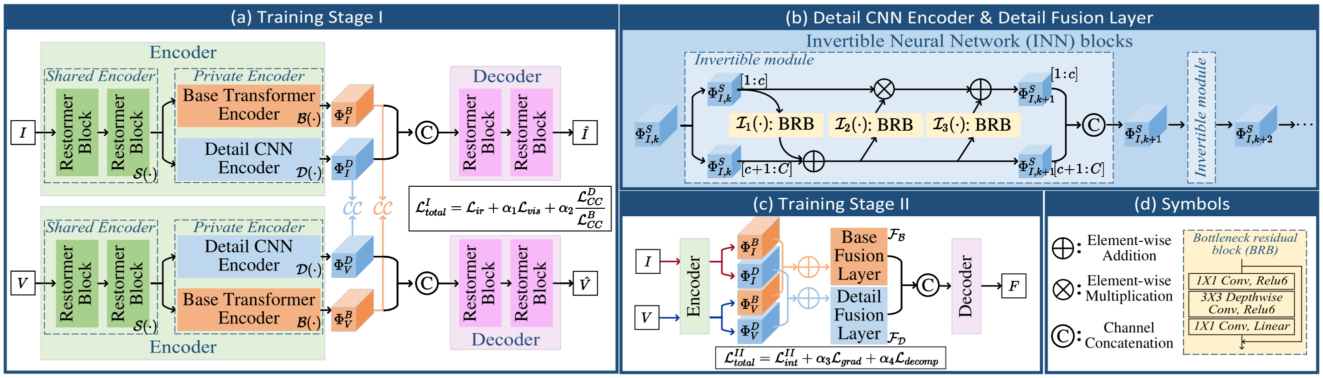


Figure 2. The architecture of our CDDFuse method (IVF as an example). **(a)** The pipeline in training stage I, which aims to train an AE structure for base/detail feature decomposition and reconstructing source images. **(b)** The INN block-based DCE and detail fusion layer, and the BRB block in the affine coupling layers of INN. **(c)** The pipeline in training stage II, which aims to obtain fusion images.

features without increasing too much computation. The architecture of Restormer block we use can be referred in supplementary material or the original paper [81].

Base transformer encoder. The BTE is to extract low-frequency base features from the shared features:

$$\Phi_I^B = \mathcal{B}\left(\Phi_I^S\right),\ \Phi_V^B = \mathcal{B}\left(\Phi_V^S\right). \tag{2}$$

where Φ_I^B and Φ_V^B are the base feature of I and V, respectively. In order to extract long-distance dependency features, we use a Transformer with spatial self-attention. Considering to balance the performance and computational efficiency, we use the LT block [67] as the basic unit of BTE. Through the structure of Flattened feed-forward network which flattens the bottleneck of Transformer blocks, the LT block shrinks the embedding to reduce the number of parameters while preserving the same performance, meeting our expectation.

Detail CNN encoder. Contrary to BTE, the DCE extracts high-frequency detail information from the shared features, which is formulated as:

$$\Phi_I^D = \mathcal{D}\left(\Phi_I^S\right),\ \Phi_V^D = \mathcal{D}\left(\Phi_V^S\right). \tag{3}$$

Considering that edge and texture information in detail features are very important for image fusion tasks, we hope that the CNN architecture in DCE can preserve as much detail information as possible. The INN [13] module enables the input information to be better preserved by making its input and output features mutually generated. Thus, it can be regarded as a lossless feature extraction module and is very suitable for use here. Therefore, we adopt the INN block with affine coupling layers [13, 93]. In each invertible layer, the transformation is:

$$\begin{aligned}
\Phi_{I,k+1}^S\left[c+1\!:\!C\right] &= \Phi_{I,k}^S\left[c+1\!:\!C\right] + \mathcal{I}_1\left(\Phi_{I,k}^S\left[1\!:\!c\right]\right),\\
\Phi_{I,k+1}^S\left[1\!:\!c\right] &= \Phi_{I,k}^S\left[1\!:\!c\right] \odot \exp\left(\mathcal{I}_2\left(\Phi_{I,k+1}^S\left[c+1\!:\!C\right]\right)\right)\\
&\quad + \mathcal{I}_3\left(\Phi_{I,k+1}^S\left[c+1\!:\!C\right]\right),\\
\Phi_{I,k+1}^S &= \mathcal{CAT}\left\{\Phi_{I,k+1}^S\left[1\!:\!c\right], \Phi_{I,k+1}^S\left[c+1\!:\!C\right]\right\}
\end{aligned} \tag{4}$$

where the $\odot$ is the Hadamard product, $\Phi_{I,k}^S\left[1\!:\!c\right] \in \mathbb{R}^{h\times w\times c}$ is the 1st to the cth channels of input feature for the kth invertible layer ($k = 1, \cdots, K$), $\mathcal{CAT}(\cdot)$ is the channel concatenation operation and $\mathcal{I}_i$ ($i = 1, \cdots, 3$) are the arbitrary mapping functions. The calculation details can be seen in Fig 2(d) and the supplementary meterial. In each invertible layer, $\mathcal{I}_i$ can be set to any mapping without affecting the lossless information transmission in this invertible layer. Considering the trade-off between computational consumption and feature extraction ability, we employ bottleneck residual block (BRB) block in MobileNetV2 [53] as $\mathcal{I}_i$. Finally, $\Phi_I^D = \Phi_{I,K}^S$ and Φ_V^D can be obtained in the same way, by replacing the subscript in Eq. (4) from I to V.

3.3. Fusion layer

The function of the base/detail fusion layer is to fuse base/detail features, respectively. Considering the inductive bias for base/detail feature fusion should be similar to base/detail feature extraction in the encoder, we employ LT and INN blocks for the base and detail fusion layer, where:

$$\Phi^B = \mathcal{F}_\mathcal{B}\left(\Phi_I^B, \Phi_I^B\right),\ \Phi^D = \mathcal{F}_\mathcal{D}\left(\Phi_I^D, \Phi_V^D\right), \tag{5}$$

$\mathcal{F}_\mathcal{B}$ and $\mathcal{F}_\mathcal{D}$ are the base and detail fusion layer, respectively.

3.4. Decoder

In the decoder $\mathcal{DC}(\cdot)$, the decomposed features are concatenated in the channel dimension as the input, and the original image (training stage I) or the fused image (training stage II) is the output of the decoder, which is formulated as:

$$\begin{aligned}
&\text{Stage I:} && \hat{I} = \mathcal{DC}\left(\Phi_I^B, \Phi_I^D\right), \hat{V} = \mathcal{DC}\left(\Phi_V^B, \Phi_V^D\right);\\
&\text{Stage II:} && F = \mathcal{DC}\left(\Phi^B, \Phi^D\right).
\end{aligned} \tag{6}$$

Since the inputs here involving cross-modality and multi-frequency features, we keep the decoder structure consistent with the design of SFE, *i.e.*, using the Restormer block as the basic unit of the decoder.

3.5. Two-stage training

A big challenge of the MMIF task is that due to its lack of ground truth, advanced supervised learning methods are ineffective. Here, inspired by [28], we use a two-stage learning scheme to train our CDDFuse end-to-end.

Training stage I. In the training stage I, the paired infrared and visible images $\{I, V\}$ are input into the SFE to extracts shallow features $\{\Phi_I^S, \Phi_V^S\}$. Then the LT block-based BTE and the INN-based DCE are employed to extract low-frequency base feature $\{\Phi_I^B, \Phi_V^B\}$ and high-frequency detail feature $\{\Phi_I^D, \Phi_V^D\}$ for the two different modalities, respectively. After that, the base and detail features of infrared $\{\Phi_I^B, \Phi_I^D\}$ (or visible $\{\Phi_V^B, \Phi_V^D\}$) images are concatenated and input into the decoder to reconstruct the original infrared image $\hat{I}$ (or visible image $\hat{V}$).

Training stage II. In the training stage II, the paired infrared and visible images $\{I, V\}$ are input into a nearly well-trained Encoder to obtain the decomposition features. Then the decomposed base features $\{\Phi_I^B, \Phi_V^B\}$ and detail features $\{\Phi_I^D, \Phi_V^D\}$ are input into the fusion layer $\mathcal{F}_\mathcal{B}$ and $\mathcal{F}_\mathcal{D}$, respectively. At last, the fused features $\{\Phi^B, \Phi^D\}$ are input into the decoder to obtain the fused image F.

Training losses. In training stage I, the total loss $\mathcal{L}_{total}^I$ is:

$$\mathcal{L}_{total}^I = \mathcal{L}_{ir} + \alpha_1 \mathcal{L}_{vis} + \alpha_2 \mathcal{L}_{decomp}, \tag{7}$$

where $\mathcal{L}_{ir}$ and $\mathcal{L}_{vis}$ are the reconstruction losses for infrared and visible images, $\mathcal{L}_{decomp}$ is the feature decomposition loss, and α_1 as well as α_2 are the tuning papameters. The reconstruction losses mainly ensure that the information contained in the images is not lost during the encoding and decoding process, *i.e.*

$$\mathcal{L}_{ir} = \mathcal{L}_{int}^I(I, \hat{I}) + \mu \mathcal{L}_{SSIM}(I, \hat{I}), \tag{8}$$

where $\mathcal{L}_{int}^I = \|I - \hat{I}\|_2^2$ and $\mathcal{L}_{SSIM}(I, \hat{I}) = 1 - SSIM(I, \hat{I})$. $SSIM(\cdot, \cdot)$ is the structural similarity index [65]. $\mathcal{L}_{vis}$ can be obtained in the same way. Additionally, our proposed feature decomposition loss $\mathcal{L}_{decomp}$ is:

$$\mathcal{L}_{decomp} = \frac{\left(\mathcal{L}_{CC}^D\right)^2}{\mathcal{L}_{CC}^B} = \frac{\left(\mathcal{CC}\left(\Phi_I^D, \Phi_V^D\right)\right)^2}{\mathcal{CC}\left(\Phi_I^B, \Phi_V^B\right) + \epsilon} \tag{9}$$

where $\mathcal{CC}(\cdot, \cdot)$ is the correlation coefficient operator, and ϵ here is set to 1.01 to ensure this term always being positive.

The motivation of this loss term is that, according to our MMIF assumption, the decomposed features $\{\Phi_I^B, \Phi_V^B\}$ will contain more modality-shared information, such as background and large-scale environment, so they are often highly correlated. In contrast, $\{\Phi_I^D, \Phi_V^D\}$ represents the texture and detail information in V and the thermal radiation as well as clear edge information in I, which is modality-specific. Thus, the feature maps are less correlated. Empirically, under the guidance of $\mathcal{L}_{decomp}$ in gradient descent, $\mathcal{L}_{CC}^D$ gradually approaches 0 and $\mathcal{L}_{CC}^B$ becomes larger, which satisfies our intuition for feature decomposition. The visualization of the decomposition effect will be shown in Fig. 5.

Subsequently in training stage II, inspaired by [56], the total loss becomes:

$$\mathcal{L}_{total}^{II} = \mathcal{L}_{int}^{II} + \alpha_3 \mathcal{L}_{grad} + \alpha_4 \mathcal{L}_{decomp}, \tag{10}$$

where $\mathcal{L}_{int}^{II} = \frac{1}{HW}\|I_f - \max(I_{ir}, I_{vis})\|_1$ and $\mathcal{L}_{grad} = \frac{1}{HW}\| |\nabla I_f| - \max(|\nabla I_{ir}|, |\nabla I_{vis}|)\|_1$. ∇ indicates the Sobel gradient operator. α_3 and α_4 are the tuning parameters.

4. Infrared and visible image fusion

Here we elaborate the implementation and configuration details of our networks for the IVF task. Experiments are conducted to show the performance of our models and the rationality of network structures.

4.1. Setup

Datasets and metrics. IVF experiments use three popular benchmarks to verify our fusion model, i.e., MSRS [57], RoadScene [71], and TNO [59]. We train our network on MSRS training set (1083 pairs) and 50 pairs in RoadScene are used for validation. MSRS test set (361 pairs), RoadScene (50 pairs) and TNO (25 pairs) are employed as test datasets, which the fusion performance can be verified comprehensively. Note that fine-tuning is not applied to the RoadScene and TNO datasets to verify the generalization performance of the fusion models.

We use eight metrics to quantitatively measure the fusion results: entropy (EN), standard deviation (SD), spatial frequency (SF), mutual information (MI), sum of the correlations of differences (SCD), visual information fidelity (VIF), $Q^{AB/F}$ and structural similarity index measure (SSIM). Higher metrics indicate that a fusion image is better. The details of these metrics can be found in [40].

Implement details. Our experiments are carried out on a machine with two NVIDIA GeForce RTX 3090 GPUs. The training samples are randomly cropped into 128×128 patches in the preprocessing stage. The number of epochs for training is set to 120 with 40 and 80 epochs in the first and second stages, respectively. The batch size is set to 16. We adopt the Adam optimizer with the initial learning rate set to 10^{-4} and decreasing by 0.5 every 20 epochs. For the network hyperparameters setting, the number of Restormer blocks in SFE is 4, with 8 attention heads and 64 dimensions. The dimension of the LT block in BTE is also 64 with 8 attention heads. The configuration of decoder is the same as encoder. As for loss functions Eq. (7) and (10), α_1 to α_4 are set to 1, 2, 10, and 2, in order to keep the same order of magnitude for each term.

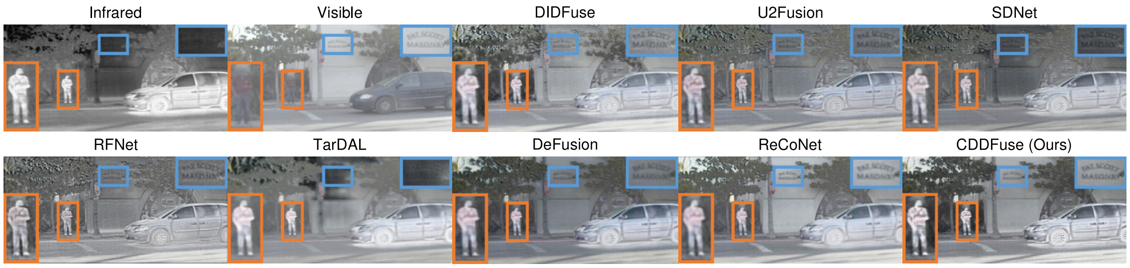


Figure 3. Visual comparison for "FLIR_04602" in RoadScene IVF dataset.

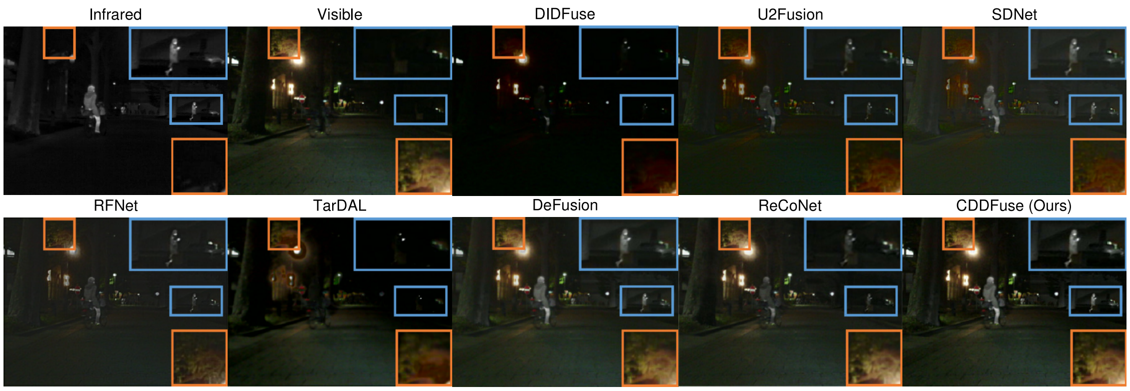


Figure 4. Visual comparison for "00706N" in MSRS IVF dataset.

4.2. Comparison with SOTA methods

In this section, we test CDDFuse on the three test sets and compare the fusion results with the state-of-the-art methods including DIDFuse [88], U2Fusion [70], SDNet [82], RFNet [72], TarDAL [35], DeFusion [32] and ReCoNet [19].

Qualitative comparison. We show the qualitative comparison in Figs. 3 and 4. Obviously, our method better integrates thermal radiation information in infrared images and detailed textures in visible images. Objects in dark regions are clearly highlighted, so that foreground targets can be easily distinguished from the background. Additionally, background details that are difficult to identify due to the low illumination have clear edges and abundant contour information, which help us understand the scene better.

Quantitative comparison. Afterward, eight metrics are employed to quantitatively compare the above results, which are displayed in Tab. 1. Our method has excellent performance on almost all metrics, proving that our method is suitable for various kinds of illumination and target categories.

Visualization of feature decomposition. Fig. 5 visualizes the decomposed features. Obviously, more background information in the base feature group is activated, and the activated areas are also relevant. In the detail feature group, infrared features instead focus more on object highlights, while visible features pay more attention to details and textures, showing that the modality-specific features are well extracted. The visualization is consistent with our analysis.

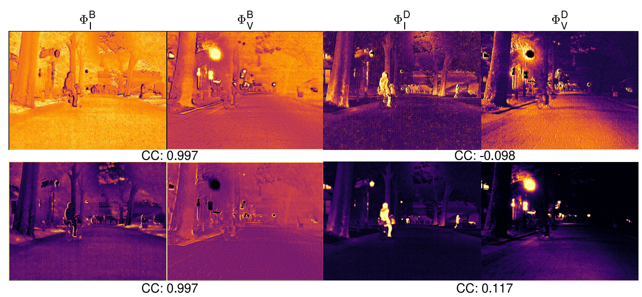


Figure 5. Visualization of the decomposed features

4.3. Ablation studies

Ablation experiments are set to verify the rationality of the different modules. EN, SD, VIF and SSIM are used to quantitatively validate the fusion effectiveness. The results of experimental groups are shown in Tab. 2.

Decomposition loss $\mathcal{L}_{decomp}$. In Exp. I, we change the definition in Eq. (9) from division to subtraction as $\mathcal{L}_{decomp} = (\mathcal{L}_{CC}^D)^2 - \mathcal{L}_{CC}^B$, which can also increase $\mathcal{L}_{CC}^B$ and decrease $(\mathcal{L}_{CC}^D)^2$.The results of Exp. I demonstrate that although the new loss can generate marginally satisfactory results, it produces poor results compared to the definition in Eq. (9). In Exp. II, we do not use the correlation-driven loss $\mathcal{L}_{decomp}$, and the results show that $\mathcal{L}_{decomp}$ is necessary for feature decomposition. There is no guarantee that BTE and DCE can learn the different frequency features without $\mathcal{L}_{decomp}$.

LT and the INN blocks. We then verify the necessity of LT blocks and INN blocks in the *Long-short Range Encoder*. In

Table 1. Quantitative results of the IVF task. **Boldface** and underline show the best and second-best values, respectively.

	EN	SD	SF	MI	SCD	VIF	Qbaf	SSIM
Dataset: MSRS Infrared-Visible Fusion Dataset [57]								
DID [88]	4.27	31.49	10.15	1.61	1.11	0.31	0.20	0.24
U2F [70]	5.37	25.52	9.07	1.40	1.24	0.54	0.42	0.77
SDN [82]	5.25	17.35	8.67	1.19	0.99	0.50	0.38	0.72
RFN [72]	5.56	24.09	**11.98**	1.30	1.13	0.51	0.43	0.83
TarD [35]	5.28	25.22	5.98	1.49	0.71	0.42	0.18	0.47
DeF [32]	6.46	37.63	8.60	<u>2.16</u>	1.35	<u>0.77</u>	<u>0.54</u>	<u>0.94</u>
ReC [19]	<u>6.61</u>	<u>43.24</u>	9.77	2.16	<u>1.44</u>	0.71	0.50	0.85
CDDFuse	**6.70**	**43.38**	<u>11.56</u>	**3.47**	**1.62**	**1.05**	**0.69**	**1.00**
Dataset: TNO Infrared-Visible Fusion Dataset [59]								
DID [88]	6.97	45.12	12.59	1.70	<u>1.71</u>	<u>0.60</u>	0.40	0.81
U2F [70]	6.83	34.55	11.52	1.37	1.71	0.58	<u>0.44</u>	0.99
SDN [82]	6.64	32.66	12.05	1.52	1.49	0.56	0.44	<u>1.00</u>
RFN [72]	6.83	34.50	**15.71**	1.20	1.67	0.51	0.39	0.92
TarD [35]	6.84	<u>45.63</u>	8.68	<u>1.86</u>	1.52	0.53	0.32	0.88
DeF [32]	6.95	38.41	8.21	1.78	1.64	0.60	0.41	0.96
ReC [19]	<u>7.10</u>	44.85	8.73	1.78	1.70	0.57	0.39	0.88
CDDFuse	**7.12**	**46.00**	<u>13.15</u>	**2.19**	**1.76**	**0.77**	**0.54**	**1.03**
Dataset: RoadScene Infrared-Visible Fusion Dataset [71]								
DID [88]	<u>7.43</u>	51.58	14.66	2.11	1.70	0.58	0.48	0.86
U2F [70]	7.09	38.12	13.25	1.87	1.70	0.60	<u>0.51</u>	0.97
SDN [82]	7.14	40.20	13.70	2.21	1.49	0.60	0.51	**0.99**
RFN [72]	7.21	41.25	<u>16.19</u>	1.68	1.73	0.54	0.45	0.90
TarD [35]	7.17	47.44	10.83	2.14	1.55	0.54	0.40	0.88
DeF [32]	7.23	44.44	10.22	<u>2.25</u>	1.69	<u>0.63</u>	0.48	0.89
ReC [19]	7.36	<u>52.54</u>	10.78	2.18	<u>1.74</u>	0.59	0.43	0.88
CDDFuse	**7.44**	**54.67**	**16.36**	**2.30**	**1.81**	**0.69**	**0.52**	<u>0.98</u>

Exp. III, we changed the LT blocks as INN, *i.e.*, the base and detail features are both extracted by INN blocks. Similarly, in Exp. IV, the features of different modalities are extracted by LT blocks. The results show that although the ability of feature extraction for LT blocks is slightly stronger than that of INN blocks, it is worse than that of CDDFuse which cooperates with LT and INN blocks. Subsequently, in Exp. V, we changed the INN module as a CNN module composed of BRBs with similar parameters in INN blocks, and its effect is slightly worse than that of the LT module alone, which proves that the information loss is serious when the CNN is employed to accomplish the fusion task.

Two-stage training. Finally, if we abandon the two-stage training and directly train the encoder, decoder and fusion layer simultaneously, the results are very unsatisfactory. It is proved that two-stage training can effectively reduce the difficulty of training and improve training robustness.

In summary, ablation results in Tab. 2 demonstrate the effectiveness and rationality of our network design.

4.4. Downstream IVF applications

To further study fusion performance in high-level MM computer vision tasks, this section applies the infrared, visible and fusion images of SOTA methods in Sec. 4.2 to MM object detection and semantic segmentation, and investigate information fusion's benefit for the downstream tasks. Due to space constraints, qualitative results are presented in the supplementary material.

Table 2. Ablation experiment results in the testset of MSRS. **Bold** indicates the best value.

	Configurations	EN	SD	VIF	SSIM
I	Division $\to$ Subtraction in $\mathcal{L}_{decomp}$	6.55	42.20	0.98	1.00
II	w/o $\mathcal{L}_{decomp}$	6.19	36.49	0.96	0.97
III	LT block $\to$ INN in BTE	6.47	41.39	1.00	0.98
IV	INN $\to$ LT block in DCE	6.56	42.18	1.00	0.99
V	INN $\to$ CNN block in DCE	6.54	42.10	0.98	0.98
VI	w/o two-stage training	6.28	38.42	0.97	0.99
	Ours	**6.70**	**43.38**	**1.05**	**1.00**

4.4.1 Infrared-visible object detection

Setup. The MM object detection is performed on M^{3}FD dataset [35] with 4200 pairs of infrared/visible images, and six categories of labels (*i.e.*, people, car, bus, motorcycle, truck and lamp). It is divided into training/validation/test sets with a proportion 8:1:1. YOLOv5 [24], a SOTA detector, is employed to evaluate the detection performance with the metric mAP@0.5. The training epoch, batch size, optimizer and initial learning rate are set as 400, 8, SGD optimizer and 1e-2, respectively.

Comparison with SOTA methods. Tab. 3 displays that CDDFuse has the best detection performance, especially in the People and Truck classes, demonstrating that CDDFuse can improve detection accuracy by fusing thermal radiation information and highlighting the difficult-to-observe targets.

4.4.2 Infrared-visible semantic segmentation

Setup. We conduct the MM semantic segmentation on the MSRS dataset [57] with semantic information of nine object categories (*i.e.*, background, bump, color cone, guardrail, curve, bike, person, car stop and car). The division of dataset follows [57]. The backbone we choose is DeeplabV3+ [10] and we compare the model effectiveness by intersection-over-union (IoU). All the models are supervised by cross-entropy loss and trained by SGD with the batchsize of 8 over 340 epochs, of which the first 100 epochs are trained by freezing the backbone network. The initial learning rate is 7e-3 and is decreased by the cosine annealing delay.

Comparison with SOTA methods. The segmentation results are exhibited in Tab. 4. CDDFuse better integrates the edge and contour information in the source images, which enhances the ability of model to perceive the object boundary, and makes the segmentation more accurate.

5. Medical image fusion

Setup. We selected 286 pairs of medical images from the Harvard Medical website [1] for MIF experiments, of which

Table 3. AP@0.5(%) values for MM detection on M^3FD dataset.

	Bus	Car	Lam	Mot	Peo	Tru	mAP@0.5
IR	78.75	88.69	70.17	63.42	80.91	65.77	74.62
VI	78.29	90.73	86.35	69.33	70.53	70.91	77.69
DID	79.65	92.51	84.70	68.72	79.61	68.78	78.99
U2F	79.15	92.29	87.61	66.75	80.67	71.37	79.64
SDN	81.44	92.33	84.14	67.37	79.35	69.29	78.99
RFN	78.15	91.94	84.95	**72.80**	79.41	69.04	79.38
TarD	81.33	**94.76**	87.13	69.34	81.52	68.65	80.45
DeF	**82.94**	92.49	**87.78**	69.45	80.82	71.44	80.82
ReC	78.92	91.79	87.41	69.34	79.41	69.98	79.48
Ours	82.60	92.54	86.88	71.62	**81.60**	**71.53**	**81.13**

Table 4. IoU(%) values for MM segmentation on MSRS dataset.

Models	Unl	Car	Per	Bik	Cur	CS	GD	CC	Bu	mIOU
VI	90.5	75.6	45.4	59.4	37.2	51.0	46.4	43.5	50.2	55.4
IR	84.7	67.8	56.4	51.8	34.6	39.3	42.2	40.2	48.4	51.7
DID [88]	97.2	78.3	58.7	60.9	36.2	52.9	**62.4**	44.0	55.7	60.7
U2F [70]	97.5	82.3	63.4	62.6	40.3	52.6	51.9	44.8	**59.5**	61.7
SDN [82]	97.3	78.4	62.5	61.7	35.7	49.3	52.4	42.2	52.9	59.2
RFN [72]	97.3	78.7	60.6	61.3	36.3	49.4	45.6	45.7	48.0	58.1
TarD [35]	97.1	79.1	55.4	59.0	33.6	49.4	54.9	42.6	53.5	58.3
DeF [32]	97.5	82.6	61.1	62.6	40.4	51.5	48.1	47.9	54.8	60.7
ReC [19]	97.4	81.0	59.9	61.4	41.0	51.3	54.4	47.4	55.9	61.1
Ours	**97.7**	**84.6**	**64.2**	**65.1**	**43.9**	**53.8**	61.7	**50.6**	57.3	**64.3**

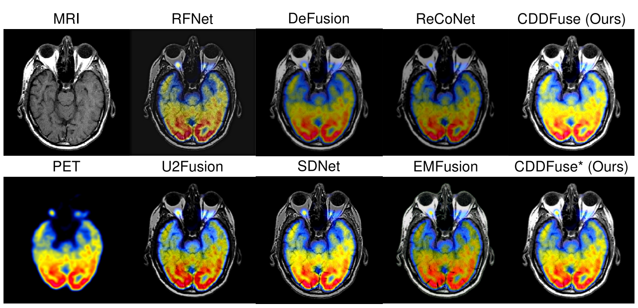


Figure 6. Visual comparison for "MRI-PET-16" in MRI-PET MIF.

130 pairs and 20 pairs are used for training and validation, respectively. 21 pairs of MRI-CT images, 42 pairs of MRI-PET images and 73 pairs of MRI-SPECT images are utilized as the test datasets. The training strategy and the metrics for evaluation are the same as that for IVF.

Comparison methods. We performed two groups of experiments. First, we compare fusion methods trained on the IVF task, *i.e*. TarDAL [35], RFNet [72], DeFusion [32], ReCoNet [19] and our CDDFuse, to demonstrate the generalization ability of the fusion methods. Note that the above methods are not fine-tuned on the MIF dataset. Then, we train CDDFuse on the MIF dataset (denoted as CDDFuse*), and compare it with U2Fusion [70], SDNet [82] and EMFusion [69], which are all trained on the MIF dataset.

Comparison with SOTA methods. Qualitative and Quantitative results are presented in Fig. 6 and Tab. 5. CDDFuse can preserve the detailed texture and highlight the structure information, and achieves leading performance on almost all metrics whether trained on the MIF dataset or not.

Table 5. Quantitative results of the MIF task. **Boldface** and underline show the best and second-best values, respectively. CDDFuse* represents the results after training on MIF datasets.

	EN	SD	SF	MI	SCD	VIF	$Q^{AB/F}$	SSIM
Dataset: MRI-CT Medical Image Fusion								
TarD [35]	4.75	61.14	28.38	1.94	0.81	0.32	0.35	0.61
RFN [72]	**5.30**	52.95	33.42	1.98	0.58	0.33	0.52	0.49
DeF [32]	4.63	66.38	21.56	2.20	1.12	0.47	0.44	1.29
ReC [19]	4.41	66.96	20.16	2.03	1.24	0.40	0.42	1.29
CDDFuse	4.83	**88.59**	**33.83**	**2.24**	**1.74**	**0.50**	**0.59**	**1.31**
U2F [70]	4.88	52.98	22.54	2.08	0.75	0.37	0.46	0.49
SDN [82]	**5.02**	60.07	29.41	2.14	0.97	0.38	0.47	0.51
EMF [69]	4.76	72.76	22.56	2.34	1.32	0.56	0.49	1.31
CDDFuse*	4.88	**79.17**	**38.14**	**2.61**	**1.41**	**0.61**	**0.68**	**1.34**
Dataset: MRI-PET Medical Image Fusion								
	EN	SD	SF	MI	SCD	VIF	$Q^{AB/F}$	SSIM
TarD [35]	3.81	57.65	23.65	1.36	1.46	0.57	0.58	0.68
RFN [72]	**4.77**	50.57	**29.11**	1.53	0.96	0.39	0.52	0.42
DeF [32]	4.17	64.65	22.35	1.74	1.48	0.58	0.56	1.45
ReC [19]	3.66	65.25	21.72	1.51	1.49	0.44	0.51	1.40
CDDFuse	4.24	**81.72**	28.04	**1.87**	**1.82**	**0.66**	**0.65**	**1.46**
U2F [70]	3.73	57.07	23.27	1.69	1.27	0.40	0.49	1.39
SDN [82]	3.83	61.40	**31.97**	1.71	1.40	0.47	0.57	1.46
EMF [69]	4.21	56.80	26.01	1.82	1.31	0.62	0.67	1.47
CDDFuse*	**4.23**	**70.73**	29.57	**2.03**	**1.69**	**0.71**	**0.71**	**1.49**
Dataset: MRI-SPECT Medical Image Fusion								
	EN	SD	SF	MI	SCD	VIF	$Q^{AB/F}$	SSIM
TarD [35]	3.66	53.46	18.50	1.44	0.90	0.64	0.52	0.36
RFN [72]	**4.39**	44.01	**23.77**	1.60	0.72	0.45	0.58	0.37
DeF [32]	3.81	56.65	15.45	1.80	1.27	0.61	0.56	**1.46**
ReC [19]	3.22	60.07	17.40	1.50	1.47	0.46	0.54	1.40
CDDFuse	3.91	**71.82**	20.68	**1.89**	**1.92**	**0.66**	**0.69**	1.44
U2F [70]	3.47	52.97	19.58	1.68	1.28	0.48	0.57	1.41
SDN [82]	3.43	49.62	**22.20**	1.69	1.09	0.55	0.66	1.48
EMF [69]	3.74	51.93	17.14	1.88	1.12	0.71	0.74	**1.49**
CDDFuse*	**3.90**	**58.31**	20.87	**2.49**	**1.35**	**0.97**	**0.78**	1.48

6. Conclusion

In this paper, we propose a dual-branch Transformer-CNN architecture for multi-modal image fusion. With the help of Restormer, Lite transformer and invertible neural network blocks, modality-specific and -shared features are better extracted, and the decomposition for them is more intuitive and effective by the proposed correlation-driven decomposition loss. Experiments demonstrate the fusion effect of our CDDFuse, and the accuracy of downstream multi-modal pattern recognition tasks can be also improved.

Acknowledgement

This work has been supported by the National Key Research and Development Program of China under grant 2018AAA0102201, the National Natural Science Foundation of China under Grant 61976174 and 12201497, the Guangdong Basic and Applied Basic Research Foundation under Grant 2023A1515011358, the Fundamental Research Funds for the Central Universities under Grant D5000220060, and partly supported by the Alexander von Humboldt Foundation.

References

[1] Harvard medical website. http://www.med.harvard.edu/AANLIB/home.html. 7

[2] Lynton Ardizzone, Jakob Kruse, Carsten Rother, and Ullrich Köthe. Analyzing inverse problems with invertible neural networks. In *ICLR*, 2019. 3

[3] Lynton Ardizzone, Carsten Lüth, Jakob Kruse, Carsten Rother, and Ullrich Köthe. Guided image generation with conditional invertible neural networks. *CoRR*, abs/1907.02392, 2019. 3

[4] Wele Gedara Chaminda Bandara and Vishal M. Patel. Hypertransformer: A textural and spectral feature fusion transformer for pansharpening. In *CVPR*, pages 1757–1767, 2022. 1

[5] Jens Behrmann, Will Grathwohl, Ricky T. Q. Chen, David Duvenaud, and Jörn-Henrik Jacobsen. Invertible residual networks. In *ICML*, volume 97, pages 573–582, 2019. 3

[6] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *CoRR*, abs/2004.10934, 2020. 1

[7] Jiezhang Cao, Yawei Li, Kai Zhang, and Luc Van Gool. Video super-resolution transformer. *CoRR*, abs/2106.06847, 2021. 3

[8] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229, 2020. 3

[9] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *CVPR*, pages 12299–12310, 2021. 3

[10] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, pages 833–851, 2018. 7

[11] Xin Deng and Pier Luigi Dragotti. Deep convolutional neural network for multi-modal image restoration and fusion. *IEEE TPAMI*, 43(10):3333–3348, 2021. 2

[12] Laurent Dinh, David Krueger, and Yoshua Bengio. NICE: non-linear independent components estimation. In *ICLR (Workshop)*, 2015. 3

[13] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *ICLR*, 2017. 2, 3, 4

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2, 3

[15] Fangyuan Gao, Xin Deng, Mai Xu, Jingyi Xu, and Pier Luigi Dragotti. Multi-modal convolutional dictionary learning. *IEEE TIP*, 31:1325–1339, 2022. 2

[16] Zhicheng Geng, Luming Liang, Tianyu Ding, and Ilya Zharkov. RSTT: real-time spatial temporal transformer for space-time video super-resolution. *CoRR*, abs/2203.14186, 2022. 3

[17] Aidan N. Gomez, Mengye Ren, Raquel Urtasun, and Roger B. Grosse. The reversible residual network: Backpropagation without storing activations. In *NIPS*, pages 2214–2224, 2017. 3

[18] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014. 2

[19] Zhanbo Huang, Jinyuan Liu, Xin Fan, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Reconet: Recurrent correction network for fast and efficient multi-modality image fusion. In *ECCV*, 2022. 3, 6, 7, 8

[20] Jörn-Henrik Jacobsen, Arnold W. M. Smeulders, and Edouard Oyallon. i-revnet: Deep invertible networks. In *ICLR*, 2018. 3

[21] Alex Pappachen James and Belur V. Dasarathy. Medical image fusion: A survey of the state of the art. *Inf. Fusion*, 19:4–19, 2014. 2

[22] Zhiying Jiang, Zengxi Zhang, Xin Fan, and Risheng Liu. Towards all weather and unobstructed multi-spectral image stitching: Algorithm and benchmark. In *ACM Multimedia*, pages 3783–3791, 2022. 3

[23] Junpeng Jing, Xin Deng, Mai Xu, Jianyi Wang, and Zhenyu Guan. Hinet: Deep image hiding by invertible network. In *ICCV*, pages 4713–4722, 2021. 3

[24] Glenn Jocher. ultralytics/yolov5. https://github.com/ultralytics/yolov5, oct 2020. 7

[25] Xincheng Ju, Dong Zhang, Junhui Li, and Guodong Zhou. Transformer-based label set generation for multi-modal multi-label emotion detection. In *ACM Multimedia*, pages 512–520, 2020. 3

[26] Hyungjoo Jung, Youngjung Kim, Hyunsung Jang, Namkoo Ha, and Kwanghoon Sohn. Unsupervised deep image fusion with structure tensor representations. *IEEE TIP*, 29:3845–3858, 2020. 2

[27] Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *NeurIPS*, pages 10236–10245, 2018. 3

[28] Hui Li, Xiao-Jun Wu, and Josef Kittler. Rfn-nest: An end-to-end residual fusion network for infrared and visible images. *Inf. Fusion*, 73:72–86, 2021. 2, 5

[29] Hui Li and Xiao-Jun Wu. Densefuse: A fusion approach to infrared and visible images. *IEEE TIP*, 28(5):2614–2623, 2018. 2

[30] Wenbo Li, Xin Lu, Jiangbo Lu, Xiangyu Zhang, and Jiaya Jia. On efficient transformer and image pre-training for low-level vision. *CoRR*, abs/2112.10175, 2021. 3

[31] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *ICCVW*, pages 1833–1844, 2021. 2, 3

[32] Pengwei Liang, Junjun Jiang, Xianming Liu, and Jiayi Ma. Fusion from decomposition: A self-supervised decomposition approach for image fusion. In *ECCV*, 2022. 2, 3, 6, 7, 8

[33] Aishan Liu, Xianglong Liu, Jiaxin Fan, Yuqing Ma, Anlan Zhang, Huiyuan Xie, and Dacheng Tao. Perceptual-sensitive GAN for generating adversarial patches. In *AAAI*, pages 1028–1035, 2019. 1

[34] Aishan Liu, Xianglong Liu, Hang Yu, Chongzhi Zhang, Qiang Liu, and Dacheng Tao. Training robust deep neural networks via adversarial noise propagation. *IEEE TIP*, 2021. 1

[35] Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *CVPR*, pages 5792–5801, 2022. 2, 3, 6, 7, 8

[36] Jinyuan Liu, Xin Fan, Ji Jiang, Risheng Liu, and Zhongxuan Luo. Learning a deep multi-scale feature ensemble and an edge-attention guidance for image fusion. *IEEE TCSVT*, 32(1):105–119, 2022. 1

[37] Risheng Liu, Zhu Liu, Jinyuan Liu, and Xin Fan. Searching a hierarchically aggregated fusion architecture for fast multi-modality image fusion. In *ACM Multimedia*, pages 1600–1608, 2021. 2

[38] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 9992–10002, 2021. 2, 3

[39] Jiayi Ma, Pengwei Liang, Wei Yu, Chen Chen, Xiaojie Guo, Jia Wu, and Junjun Jiang. Infrared and visible image fusion via detail preserving adversarial learning. *Inf. Fusion*, 54:85–98, 2020. 2

[40] Jiayi Ma, Yong Ma, and Chang Li. Infrared and visible image fusion methods and applications: A survey. *Inf. Fusion*, 45:153–178, 2019. 5

[41] Jiayi Ma, Linfeng Tang, Fan Fan, Jun Huang, Xiaoguang Mei, and Yong Ma. Swinfusion: Cross-domain long-range learning for general image fusion via swin transformer. *IEEE CAA J. Autom. Sinica*, 9(7):1200–1217, 2022. 2

[42] Jiayi Ma, Wei Yu, Pengwei Liang, Chang Li, and Junjun Jiang. Fusiongan: A generative adversarial network for infrared and visible image fusion. *Inf. Fusion*, 48:11–26, 2019. 2

[43] Jiayi Ma, Hao Zhang, Zhenfeng Shao, Pengwei Liang, and Han Xu. Ganmcc: A generative adversarial network with multiclassification constraints for infrared and visible image fusion. *IEEE TIM*, 70:1–14, 2021. 2

[44] Jiayi Ma, Ji Zhao, Junjun Jiang, Huabing Zhou, and Xiaojie Guo. Locality preserving matching. *Int. J. Comput. Vis.*, 127(5):512–531, 2019. 2

[45] Kede Ma, Zhengfang Duanmu, Hojatollah Yeganeh, and Zhou Wang. Multi-exposure image fusion by optimizing A structural similarity index. *IEEE TCI*, 4(1):60–72, 2018. 1

[46] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *CVPR*, pages 2794–2802, 2017. 2

[47] Bikash Meher, Sanjay Agrawal, Rutuparna Panda, and Ajith Abraham. A survey on region based image fusion methods. *Inf. Fusion*, 48:119–132, 2019. 1

[48] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 2

[49] Haotong Qin, Yifu Ding, Mingyuan Zhang, YAN Qinghua, Aishan Liu, Qingqing Dang, Ziwei Liu, and Xianglong Liu. Bibert: Accurate fully binarized bert. In *ICLR*, 2022. 2

[50] Haotong Qin, Ruihao Gong, Xianglong Liu, Mingzhu Shen, Ziran Wei, Fengwei Yu, and Jingkuan Song. Forward and backward information retention for accurate binary neural networks. In *CVPR*, pages 2250–2259, 2020. 2

[51] Haotong Qin, Xiangguo Zhang, Ruihao Gong, Yifu Ding, Yi Xu, and Xianglong Liu. Distribution-sensitive information retention for accurate binary neural network. *International Journal of Computer Vision*, pages 1–22, 2022. 2

[52] Xuebin Qin, Zichen Vincent Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jägersand. Basnet: Boundary-aware salient object detection. In *CVPR*, pages 7479–7489, 2019. 1

[53] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, pages 4510–4520, 2018. 4

[54] Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. Detfusion: A detection-driven infrared and visible image fusion network. In *ACM Multimedia*, pages 4003–4011, 2022. 3

[55] Linfeng Tang, Yuxin Deng, Yong Ma, Jun Huang, and Jiayi Ma. Superfusion: A versatile image registration and fusion network with semantic awareness. *IEEE CAA J. Autom. Sinica*, 9(12):2121–2137, 2022. 2

[56] Linfeng Tang, Jiteng Yuan, and Jiayi Ma. Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network. *Inf. Fusion*, 82:28–42, 2022. 3, 5

[57] Linfeng Tang, Jiteng Yuan, Hao Zhang, Xingyu Jiang, and Jiayi Ma. Piafusion: A progressive infrared and visible image fusion network based on illumination aware. *Inf. Fusion*, 83-84:79–92, 2022. 1, 5, 7

[58] Shiyu Tang, Ruihao Gong, Yan Wang, Aishan Liu, Jiakai Wang, Xinyun Chen, Fengwei Yu, Xianglong Liu, Dawn Song, Alan L. Yuille, Philip H. S. Torr, and Dacheng Tao. Robustart: Benchmarking robustness on architecture design and training techniques. *CoRR*, abs/2109.05211, 2021. 1

[59] Alexander Toet and Maarten A. Hogervorst. Progress in color night vision. *Optical Engineering*, 51(1):1 – 20, 2012. 5, 7

[60] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, volume 139, pages 10347–10357, 2021. 3

[61] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017. 3

[62] Di Wang, Jinyuan Liu, Xin Fan, and Risheng Liu. Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration. In *IJCAI*, pages 3508–3515, 2022. 3

[63] Jiakai Wang, Aishan Liu, Zixin Yin, Shunchang Liu, Shiyu Tang, and Xianglong Liu. Dual attention suppression attack: Generate adversarial camouflage in physical world. In *CVPR*, 2021. 1

[64] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *ICCV*, pages 548–558, 2021. 3

[65] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004. 5

[66] Zhendong Wang, Xiaodong Cun, Jianmin Bao, and Jianzhuang Liu. Uformer: A general u-shaped transformer for image restoration. *CoRR*, abs/2106.03106, 2021. 3

[67] Zhanghao Wu, Zhijian Liu, Ji Lin, Yujun Lin, and Song Han. Lite transformer with long-short range attention. In *ICLR*, 2020. 3, 4

[68] Mingqing Xiao, Shuxin Zheng, Chang Liu, Yaolong Wang, Di He, Guolin Ke, Jiang Bian, Zhouchen Lin, and Tie-Yan Liu. Invertible image rescaling. In *ECCV*, pages 126–144, 2020. 3

[69] Han Xu and Jiayi Ma. Emfusion: An unsupervised enhanced medical image fusion network. *Inf. Fusion*, 76:177–186, 2021. 8

[70] Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. U2fusion: A unified unsupervised image fusion network. *IEEE TPAMI*, 44(1):502–518, 2022. 1, 2, 6, 7, 8

[71] Han Xu, Jiayi Ma, Zhuliang Le, Junjun Jiang, and Xiaojie Guo. Fusiondn: A unified densely connected network for image fusion. In *AAAI*, pages 12484–12491, 2020. 1, 2, 5, 7

[72] Han Xu, Jiayi Ma, Jiteng Yuan, Zhuliang Le, and Wei Liu. Rfnet: Unsupervised network for mutually reinforcing multi-modal image registration and fusion. In *CVPR*, pages 19647–19656, 2022. 3, 6, 7, 8

[73] Han Xu, Xinya Wang, and Jiayi Ma. Drf: Disentangled representation for visible and infrared image fusion. *IEEE TIM*, 70:1–13, 2021. 2

[74] Han Xu, Hao Zhang, and Jiayi Ma. Classification saliency-based rule for visible and infrared image fusion. *IEEE TCI*, 7:824–836, 2021. 2

[75] Ruikang Xu, Zeyu Xiao, Mingde Yao, Yueyi Zhang, and Zhiwei Xiong. Stereo video super-resolution via exploiting view-temporal correlations. In *ACM Multimedia*, pages 460–468, 2021. 1

[76] Shuang Xu, Jiangshe Zhang, Zixiang Zhao, Kai Sun, Junmin Liu, and Chunxia Zhang. Deep gradient projection networks for pan-sharpening. In *CVPR*, pages 1366–1375, 2021. 1

[77] Shuang Xu, Zixiang Zhao, Yicheng Wang, Chunxia Zhang, Junmin Liu, and Jiangshe Zhang. Deep convolutional sparse coding networks for image fusion. *CoRR*, abs/2005.08448, 2020. 2

[78] Zizheng Yang, Mingde Yao, Jie Huang, Man Zhou, and Feng Zhao. Sir-former: Stereo image restoration using transformer. In *ACM Multimedia*, pages 6377–6385, 2022. 1

[79] Mingde Yao, Zhiwei Xiong, Lizhi Wang, Dong Liu, and Xuejin Chen. Spectral-depth imaging with deep learning based reconstruction. *Optics express*, 27(26):38312–38325, 2019. 1

[80] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, pages 5718–5729, 2022. 2, 3

[81] Syed Waqas Zamir, Aditya Arora, Salman H. Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. *CoRR*, abs/2111.09881, 2021. 3, 4

[82] Hao Zhang and Jiayi Ma. Sdnet: A versatile squeeze-and-decomposition network for real-time image fusion. *Int. J. Comput. Vis.*, 129(10):2761–2785, 2021. 2, 6, 7, 8

[83] Hao Zhang, Han Xu, Yang Xiao, Xiaojie Guo, and Jiayi Ma. Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity. In *AAAI*, pages 12797–12804, 2020. 2, 3

[84] Xingchen Zhang. Deep learning-based multi-focus image fusion: A survey and a comparative study. *IEEE TPAMI*, 2021. 1

[85] Yu Zhang, Yu Liu, Peng Sun, Han Yan, Xiaolin Zhao, and Li Zhang. IFCNN: A general image fusion framework based on convolutional neural network. *Inf. Fusion*, 54:99–118, 2020. 2

[86] Jiawei Zhao, Yifan Zhao, and Jia Li. M3TR: multi-modal multi-label recognition with transformer. In *ACM Multimedia*, pages 469–477, 2021. 3

[87] Zixiang Zhao, Shuang Xu, Chunxia Zhang, Junmin Liu, and Jiangshe Zhang. Bayesian fusion for infrared and visible images. *Signal Process.*, 177:107734, 2020. 1

[88] Zixiang Zhao, Shuang Xu, Chunxia Zhang, Junmin Liu, Jiangshe Zhang, and Pengfei Li. DIDFuse: Deep image decomposition for infrared and visible image fusion. In *IJCAI*, pages 970–976, 2020. 1, 2, 3, 6, 7, 8

[89] Zixiang Zhao, Shuang Xu, Jiangshe Zhang, Chengyang Liang, Chunxia Zhang, and Junmin Liu. Efficient and model-based infrared and visible image fusion via algorithm unrolling. *IEEE TCSVT*, 32(3):1186–1196, 2022. 2

[90] Zixiang Zhao, Jiangshe Zhang, Shuang Xu, Zudi Lin, and Hanspeter Pfister. Discrete cosine transform network for guided depth map super-resolution. In *CVPR*, pages 5697–5707, June 2022. 3

[91] Zixiang Zhao, Jiangshe Zhang, Shuang Xu, Kai Sun, Lu Huang, Junmin Liu, and Chunxia Zhang. FGF-GAN: A lightweight generative adversarial network for pansharpening via fast guided filter. In *ICME*, pages 1–6, 2021. 1

[92] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H. S. Torr, and Li Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *CVPR*, pages 6881–6890, 2021. 3

[93] Man Zhou, Xueyang Fu, Jie Huang, Feng Zhao, Aiping Liu, and Rujing Wang. Effective pan-sharpening with transformer and invertible neural network. *IEEE TGRS*, 60:1–15, 2022. 4

[94] Man Zhou, Keyu Yan, Jie Huang, Zihe Yang, Xueyang Fu, and Feng Zhao. Mutual information-driven pan-sharpening. In *CVPR*, pages 1788–1798, 2022. 3

[95] Xiaobin Zhu, Zhuangzi Li, Xiaoyu Zhang, Changsheng Li, Yaqi Liu, and Ziyu Xue. Residual invertible spatio-temporal network for video super-resolution. In *AAAI*, pages 5981–5988, 2019. 3

[96] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: deformable transformers for end-to-end object detection. In *ICLR*, 2021. 3