SigFusion: Unified Signal-Level Self-Supervised Learning Paradigm for Image Fusion

Zeyu Wang[1], Jiawei Feng[1], Jiayu Wang[1], Pengjie Wang[1], Haiyu Song[1,*]

[1]College of Computer Science and Engineering, Dalian Minzu University, Dalian 116600, China
{wangzeyu, 90251308, shy}@dlnu.edu.cn, pengjiewang@gmail.com, 202311054010@stu.dlnu.edu.cn

Abstract

Image Fusion (IF) aims to integrate complementary features from multiple source images into a single image. However, a key challenge in this field is the lack of large-scale real-world training datasets. Existing models typically rely on either small datasets or synthetic, less realistic datasets. To address this, we propose SigFusion, a unified signal-level self-supervised learning paradigm for various IF tasks. The core idea is to use signal-level Pseudo-Label Generation Networks (PLGN) to automatically synthesize training sets and pseudo-labels with real multi-source signal characteristics from vast unlabeled natural images. PLGN includes two critical components: learnable 1D Signal Modulators (SM) and SigFormer. SM learns implicit 1D signal patterns across various source images and embeds them into natural images, reducing the domain gap between synthetic and real datasets. SigFormer integrates Transformer with signal processing methods, establishing an appropriate signal representation space for SM. Its cascaded, multi-level design allows hierarchical feature learning from coarse to fine detail. Moreover, SigFormer can serve as a flexible backbone for IF, as its design adheres to the classic decomposition-reconstruction paradigm. Experimental results demonstrate that SigFusion achieves state-of-the-art performance across multiple IF tasks, including medical image fusion, infrared-visible image fusion, multi-focus image fusion, and multi-exposure image fusion.

Code — https://github.com/fengjiawei123/SigFusion.

Introduction

Image fusion (IF) integrates complementary information from multiple source images into a unified representation (Karim et al. 2023). Common IF tasks include medical image fusion, infrared-visible fusion (collectively multi-modal image fusion, MMIF), multi-focus, and multi-exposure fusion (digital photography image fusion, DPIF) (Zhang et al. 2021). Effective IF significantly benefits downstream tasks, such as object detection and medical segmentation (Liu et al. 2020; Hermessi, Mourali, and Zagrouba 2021).

However, a key bottleneck remains the scarcity of large-scale, real-world training datasets (Zhang and Ma 2021; Wang et al. 2022; Jung et al. 2020; Li and Wu 2018), due to constraints such as patient privacy (medical imaging), security concerns (remote sensing), and stringent pre-registration requirements. This challenge is further amplified by the success of large models, highlighting the critical role of data volume in achieving strong performance (Zhang et al. 2024; Li et al. 2024b, 2025; Zou et al. 2025).

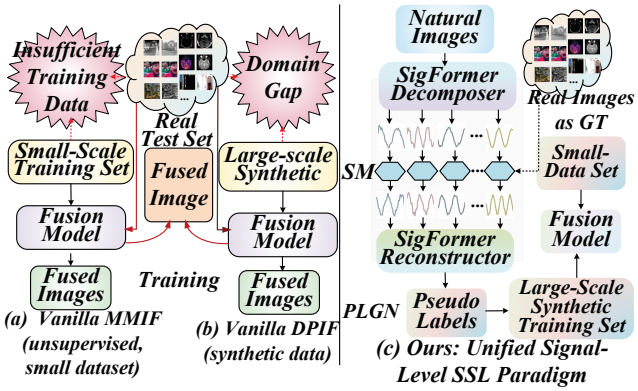

Figure 1: Existing IF Paradigm *vs* Ours. Our model synthesizes more realistic multi-source images and pseudo-labels at the signal level from large-scale unlabeled natural images.

Existing methods address data scarcity in two primary ways (see Fig. 1(a,b)): unsupervised learning on limited real datasets, and synthetic dataset creation tailored to specific tasks. Unsupervised methods, prevalent in MMIF (Xu et al. 2020a; Cheng, Xu, and Wu 2023; Tang, Li, and Ma 2025; Tang et al. 2024), often hit performance ceilings despite sophisticated architectures and carefully designed loss functions (Liu et al. 2023a; Xu et al. 2020a; Tang et al. 2025a; Wang et al. 2025a,b). The primary reason lies in the scarcity of training data, which causes models to overfit to the limited scene distributions and modality-specific characteristics present in the training set, thereby impairing generalization and real-world robustness. Conversely, DPIF methods utilize synthesized datasets (e.g., simulated multi-exposure and multi-focus images via brightness adjustments and localized blurs (Wang et al. 2023; Kaur and Singh 2023; Qu et al. 2023; Zou et al. 2024; Liu et al. 2025b)). Although promising, synthetic data suffer from significant domain gaps and task specificity. This motivates the central question of our

*Corresponding author: Haiyu Song

work: can we develop a unified paradigm for dataset synthesis that minimizes the domain gap between synthetic data and real multi-source images across diverse IF tasks?

To answer this, we propose a unified, self-supervised learning (SSL) paradigm (Fig. 1(c)). Following the SSL paradigm, our model operates in two stages: a pretext task stage, where high-quality representations are pre-trained on synthesized datasets with pseudo-labels; and a downstream task stage, fine-tuning these representations on real IF data. A core innovation is our signal-level Pseudo-Label Generation Network (PLGN). Unlike existing pixel- or feature-level methods, PLGN synthesizes various image fusion datasets from abundant natural images by learning signal-level characteristics from real multi-source data. By modeling images as simplified 1D signals characterized by amplitude, frequency, and phase, it overcomes variations in modality, resolution, and dynamic range, and significantly reduces domain gaps. Two essential challenges arise: capturing diverse signal patterns from multi-source images and mapping these patterns accurately into natural images.

To this end, we design two novel components for PLGN: a learnable 1D Signal Modulator (SM) and SigFormer. SM explicitly learns implicit signal-level features from real multi-source images, embedding these patterns into synthetic datasets to approximate real-world distributions. SigFormer integrates traditional signal decomposition-reconstruction techniques with Transformer. Its hierarchical, coarse-to-fine decomposition adaptively identifies optimal signal representation spaces, providing robust multi-level features.

Interestingly, our approach serves dual functions: SigFormer with SM synthesizes realistic datasets and pseudo-labels; without SM, SigFormer functions as a robust fusion model. Additionally, we develop a dedicated training loss supervising both pixel- and signal-level consistency.

Our contributions are summarized as follows:

• We propose the first unified, signal-level self-supervised learning paradigm for IF, unifying dataset synthesis and fusion tasks within one framework, suitable for diverse IF tasks.

• We propose PLGN and its central component SM, narrowing the domain gap by embedding learned signal-level characteristics from real images into synthetic data. A custom loss function supervises training at both pixel and signal levels.

• We design SigFormer, a cascaded Transformer architecture that integrates signal decomposition and reconstruction, serving as a versatile backbone for both dataset synthesis and IF.

• We synthesize and will release large-scale datasets for IF tasks (VIF, MIF, MFIF, and MEF), showing significant performance improvements across various IF models.

Related Work

Deep Learning for Image Fusion. Due to the scarcity of large-scale real-world training datasets (Zhang and Ma 2021; Wang et al. 2022; Karim et al. 2023; Liu et al. 2024a), multimodal image fusion (MMIF) and digital photography image fusion (DPIF) have evolved along distinct paths. MMIF typically employs unsupervised methods trained on small-scale real-world datasets. U2Fusion (Xu et al. 2020a) pioneered modular unsupervised fusion with task-specific losses. CoCoNet (Liu et al. 2024b) introduced a coupled contrastive framework to capture shared and modality-specific features. FILM (Zhao et al. 2024b) enhanced unsupervised fusion via auxiliary textual supervision. Nevertheless, all methods face performance ceilings due to the limited size of datasets. DPIF mitigates data scarcity via synthesis. For instance, (Wang et al. 2022, 2023) generate multi-focus datasets using random masks, and (Qu et al. 2023) simulate exposure changes via brightness curve adjustments to synthesize multi-exposure image sets. Yet, domain gaps between synthetic and real images constrain the performance of the model in real scenarios. Our method enables signal-level self-supervised training on abundant natural images, narrowing domain gaps and unifying multiple fusion tasks under a shared learning paradigm.

Self-Supervised Learning (SSL). SSL consists of two stages: pretext tasks and downstream tasks. The former learns representations from unlabeled data by exploiting inherent data structures (Gui et al. 2024), while the latter fine-tunes the learned weights on task-specific real datasets. Recent SSL-based IF methods follow this paradigm. DeFusion (Liang et al. 2022) trains an image decomposer via inpainting-based reconstruction. EMMA (Zhao et al. 2024a) generates pseudo-labels using pixel-level equivariance to supervise fusion models. However, both approaches rely on pixel- or patch-level generation methods. Currently, no dedicated signal-level generation method exists for IF. We propose a signal-level PLGN for pretext tasks. It generates image pairs with source signal features as datasets and PL from unlabeled natural images.

Signal Decomposition and Reconstruction. Signal decomposition has served as a foundation for IF, separating source images into interpretable components—such as low- and high-frequency signals—for selective merging and reconstruction (Zhu et al. 2018, 2021). This framework enhances edge preservation and noise robustness while offering interpretability. However, traditional signal-based methods are limited by their non-adaptive design, relying on fixed filters or hand-crafted decomposition rules, which struggle to generalize across diverse scenes or modalities. To overcome this, we propose SigFormer, a hybrid framework that combines the benefits of signal decomposition with neural networks to learn adaptive signal representations.

Methodology

Fig. 2 shows a unified signal-level self-supervised learning framework for IF, which consists of two stages: a pretext task and a downstream task. In the pretext stage, the Pseudo-Label Generation Network (PLGN)—comprising the Signal Modulator (SM) and SigFormer—generates synthetic datasets and pseudo-labels from natural images. These synthesized data are used to pre-train the fusion network, which shares SigFormer as its backbone. In the downstream stage, the fusion network is fine-tuned on real multi-source data, ensuring domain adaptation.

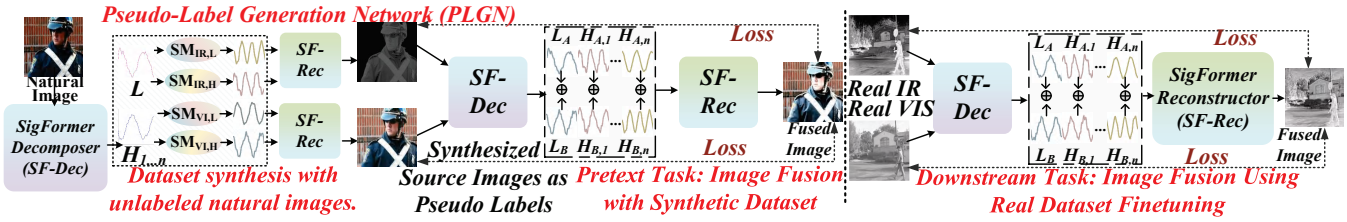

Figure 2: Schematic diagram of our model. It follows a two-stage SSL paradigm: a pretext task and a downstream task. In Stage I, PLGN synthesizes multi-source image pairs from single unlabeled natural images, treating them as pseudo-labels for the fusion task. Given the abundance of natural images, this effectively constructs a large-scale fusion training set, enabling the generation of high-quality pre-trained weights. Stage II fine-tunes these weights on a real fusion dataset.

Pretext Task

Formulation. Given a set of unlabeled natural images $D = \{X_i\}_{i=1}^N$, we employ pseudo-label generation networks $G(\cdot)$ to produce two pseudo-labels $G_1(X_i)$ and $G_2(X_i)$ for each image X_i, mimicking the dual-input requirement of fusion networks. These pseudo-labels are then used as inputs to the image fusion network $FN(\cdot)$, which generates the fused image F. In this process, the pseudo-labels serve as supervision without the need for manual annotations, enabling the network to learn meaningful image fusion representations:

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} L(FN_\theta(G_1(X_i), G_2(X_i)), G_1(X_i), G_2(X_i)). \quad (1)$$

The analysis of G is in Sec. 3.3. Here, we analyze FN and L.

Fusion Network FN**.** The fusion network leverages SigFormer's decomposer (defined by Eq. (8)-(10)) to transform image $G1$ and $G2$ into frequency signal sets $\{L_n^{G1}, H_1^{G1}, H_2^{G1}, \ldots, H_n^{G1}\}$ and $\{L_n^{G2}, H_1^{G2}, H_2^{G2}, \ldots, H_n^{G2}\}$, where L and H denote high- and low-frequency subbands. These subbands are fused by channel-wise concatenation and reconstructed by SigFormer's reconstructor (Eq. (14)) to obtain fused images.

Loss Function L**.** Loss functions differ slightly between MMIF and DPIF due to task-specific characteristics. For MMIF, we use:

$$L_{MMIF} = \alpha_1 L_{mse} + \beta_1 L_{ssim} + \gamma_1 (L_{L1}^{Max} + L_{L1}^{Grad}), \quad (2)$$

where L_{L1}^{Max} and L_{L1}^{Grad} represent L_1 metrics in the spatial pixel maximum (selecting the larger pixel value between source images as GT) and gradient domain, respectively. For DPIF tasks, the loss simplifies to:

$$L_{DPIF} = \alpha_2 L_{ssim} + \beta_2 (L_{L1}^{Max} + L_{L1}^{Grad}), \quad (3)$$

where α_2 and β_2 are hyperparameters. $G1$ and $G2$ serve as both inputs and pseudo-labels.

Pseudo-Label Generation Network (PLGN)

SigFormer with SM forms PLGN, which automatically generates pseudo-labels from unlabeled natural images without manual annotations. As the pseudo-labels generated by PLGN also serve as inputs to the fusion network (Eq. 1), PLGN effectively functions as a dataset synthesis network. PLGN's training and inference are shown in Fig. 3.

Inference. Taking the synthesis of VIS-IR images as an example, SigFormer decomposes a natural image I into a series of frequency signals $\{L_n^I, H_1^I, H_2^I, \ldots, H_n^I\}$. These signals are modulated by task-specific SM modules into IR and VIS characteristic signals, denoted by $\{L_n^{IR}, H_1^{IR}, H_2^{IR}, \ldots, H_n^{IR}\}$ and $\{L_n^{VIS}, H_1^{VIS}, H_2^{VIS}, \ldots, H_n^{VIS}\}$. SigFormer then reconstructs these into synthetic images I_{IR} and I_{VIS}.

Training. Training PLGN is challenging due to the scarcity of appropriate training datasets. We address this by using fused images from multiple fusion models as inputs, with real source images as ground truth (GT). First, we feed the fused image F and the GT image I_{GT} into SigFormer's decomposer ξ, obtaining two sets of frequency signals: $A_F = \{L_n^F, H_1^F, H_2^F, \ldots, H_n^F\}$ and $A_{GT} = \{L_n^{GT}, H_1^{GT}, H_2^{GT}, \ldots, H_n^{GT}\}$, as in Eq. (4).

$$A_F = \xi(F), \quad A_{GT} = \xi(GT). \quad (4)$$

SM then maps the frequency signals from A_F to A_{GT} as closely as possible, described by Eq. (5).

$$L_{c,i}^F = \delta_l(L_i^F), \quad H_{c,i}^F = \delta_h(H_i^F), \quad (5)$$

where δ_l and δ_h denote the SM that processes low- and high-frequency signals, resp. Finally, the modulated signals are fed into reconstructor ξ^T to obtain the output image S:

$$S = \xi^T(L_{c,n}^F, H_{c,1}^F, H_{c,2}^F, ..., H_{c,n}^F). \quad (6)$$

To optimize PLGN, we design a custom two-level loss function. It measures the differences between the input img. and GT at both the signal and pixel levels, as shown in Eq. (7):

$$\min_{\theta} l_{\text{signal}}(L_n^{GT}, \delta_s(L_{c,n}^F; \theta)) + l_{\text{signal}}(H_i^{GT}, \delta_s(H_i^{GT}; \theta)) + l_{\text{pixel}}(I_{GT}, G(F; \theta)), \quad (7)$$

where L_n^{GT} and H_i^{GT} represent the corresponding frequency sub-bands of source images, used as signal-level GT, and l_{signal} and l_{pixel} represent the signal and pixel-level loss functions, both measured using MSE. G refers to PLGN, and θ represents the trainable neural weights. Transforming fused images back to source images during training enhances the network's sensitivity to source features, allowing signal characteristic injection into natural images. The involvement of multiple fusion model results and the design of decomposing images into multiple signal subbands provide

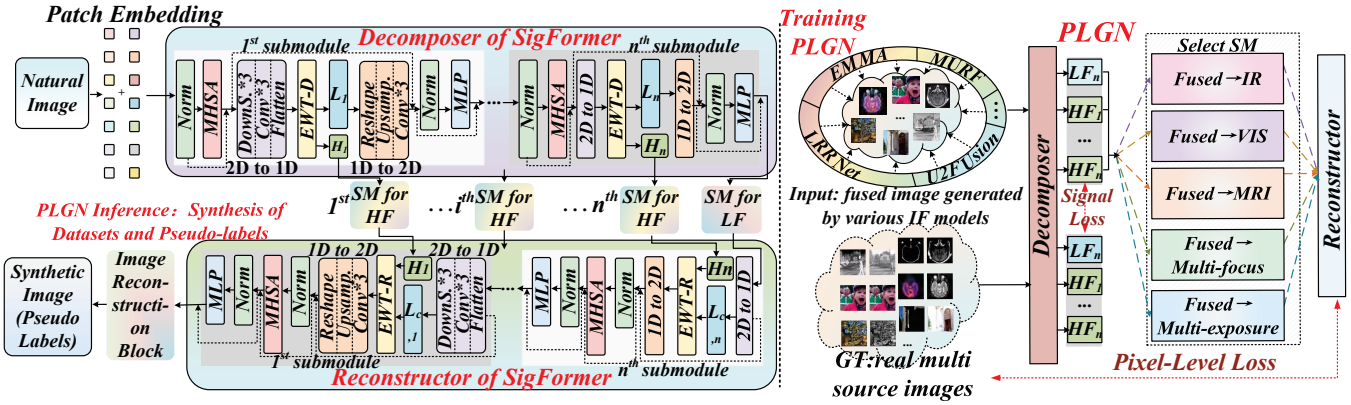

Figure 3: Schematic diagram of PLGN. The left part illustrates dataset and pseudo-label synthesis using PLGN, while the right part shows its training process.

PLGN with a sufficient training dataset. With M source image pairs, N fusion models, and each image decomposed into Z frequency subbands, the training set used by PLGN comprises $2M{\times}N{\times}Z$ pairs of 1D signals.

Signal Modulator (SM)

SM is a key component of PLGN, injecting real multi-source signal characteristics into natural images. Multiple SM modules embedded within SigFormer independently modulate frequency signals for specific image modalities (e.g., IR, VIS, CT, MRI, PET, multi-focus, multi-exposure).

Training. We use fused images from various fusion models as the training images, with real multi-source images as GT. The variability in design and techniques across these fusion models introduces considerable diversity in the fused images, enhancing the robustness of the training dataset. Detailed training procedures are in Section 3.3.

Architecture. We use a Transformer for low-frequency signals and an MLP for high-frequency signals. Transformer captures long-range dependencies, making it ideal for broad contextual processing in low-frequency components, while MLP efficiently modulates sparse high-frequency subbands. Detailed architecture is in the Appendix.

SigFormer

To explore appropriate signal frequency spaces that facilitates SM's learning, we design SigFormer.

Architecture SigFormer consists of a highly symmetrical structure with a decomposer and a reconstructor, as shown in Fig. 3. The decomposer utilizes a Transformer and the 1D signal decomposition method, Empirical Wavelet Transform (EWT) (Gilles 2013), consisting of N decomposition submodules. Taking the i-th submodule as an example, first, the multi-head self-attention (MHSA) establishes global long-range dependencies on the patch embedding from the output of the previous submodule:

$$X_i' = \Psi(X_i) + X_i. \tag{8}$$

Here, X_i is the input to the i-th submodule, and X_i' is the result after being processed by MHSA Ψ. The enhanced 2D signals x_i' are transformed into 1D representations:

$$X_{i,1D}' = \Upsilon(X_i'), \tag{9}$$

where Υ denotes down-sampling, $1{\times}1$ convolution and flatten operations. Then, EWT decomposes $x_{i,1D}'$ into a low-frequency L_i and two high-frequency signals H_i^1 and H_i^2:

$$L_i, H_i^1, H_i^2 = \mathrm{EWT}(X_{i,1D}'). \tag{10}$$

To perform subsequent operations, L_i is reshaped into 2D by applying 1*1 convolution and reshape operations Υ^T:

$$L_{i,2D} = \Upsilon^T(L_i) + X_i'. \tag{11}$$

Finally, L_i is mapped to appropriate feature spaces by MLP to prevent hindering network's learning:

$$L_i' = \mathrm{MLP}(L_{i,2D}) + L_{i,2D}. \tag{12}$$

L_i' is the output of the i-th submodule and is fed into the $(i+1)$-th submodule to repeat the process:

$$X_{i+1} = L_i'. \tag{13}$$

These submodules progressively decompose low-frequency signals while preserving high-frequency signals at each layer, enhancing hierarchical feature learning.

The reconstructor mirrors the decomposer, replacing EWT decomposition with EWT reconstruction to progressively restore input signals through hierarchical levels. In the i-th submodule, it uses L_{i+1}, H_{i+1}^1, and H_{i+1}^2 from the $(i+1)$-th submodule to recover L_i' via EWT reconstruction.

$$L_i' = \mathrm{EWT}^T(L_{i+1}, H_{i+1}^1, H_{i+1}^2). \tag{14}$$

The other steps of reconstructor are similar to decomposer. Repeat the above submodules, we obtain the final patch embedding from 1st submodule. Finally, an reconstruction block converts these patch embeddings into result images.

Dual Usage and Pretraining. SigFormer can be used with SM for PLGN or independently for fusion tasks. To ensure robust decomposition and reconstruction, it is pretrained in an unsupervised manner on large-scale natural images, learning to reconstruct each input I from itself:

$$\min_{\theta} \sum_{i=1}^{N} \frac{1}{N} ||I_i - f_\theta(I_i)||^2, \tag{15}$$

where θ denotes the trained weights; N denotes the number of samples in the natural image dataset.

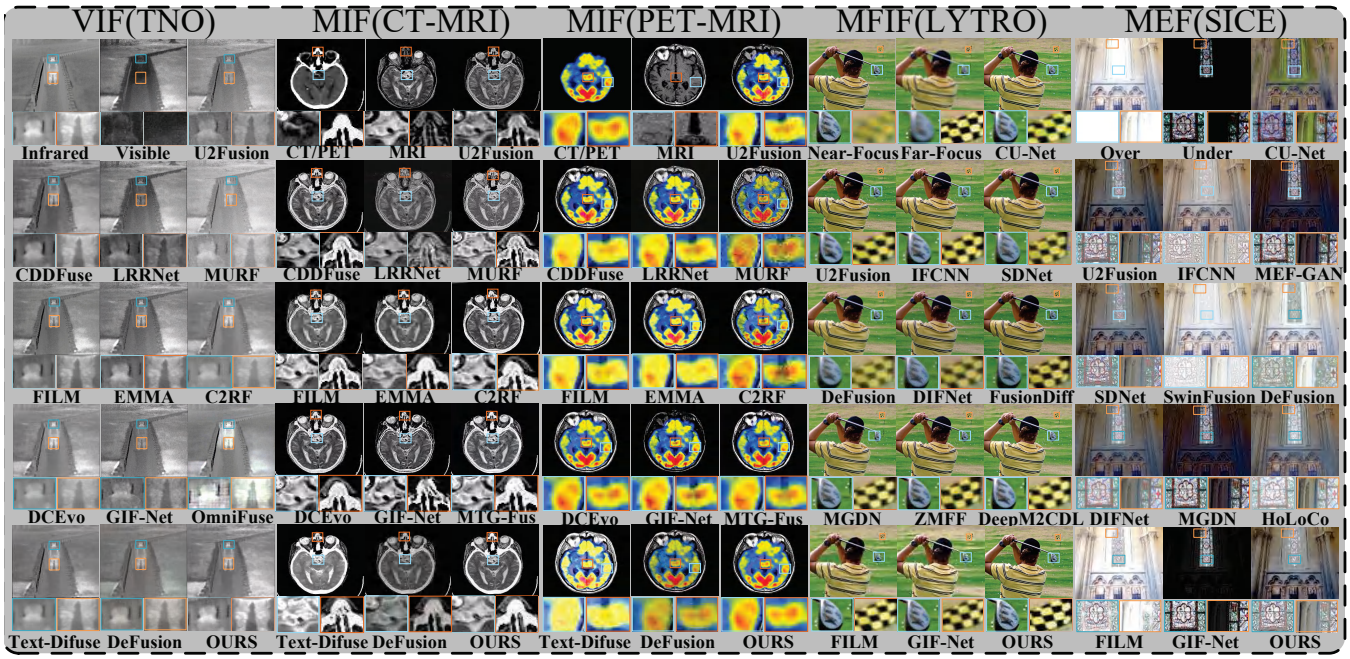

Figure 4: Results of IF models on four tasks. Assessing guidelines: VIF: IR brightness, VIS color fidelity and edges; MIF: CT bone, PET color, MRI edges; MFIF: detail retention from sharper images; MEF: brightness balance and edges.

Stage II: Fine-tuning Image Fusion Network

Despite reducing the domain gap, residual differences still hinder full adaptation to real multi-source images. Therefore, we fine-tune the pre-trained fusion network using real multi-source image pairs with the same loss functions used in pretraining (Eq. (2)-(3)):

$$\min_{\theta} \sum_{i=1}^{N} L(FN_{\theta}(I_A, I_B); I_A, I_B). \tag{16}$$

As real IF datasets lack fused GT, source images serve as supervision, aligning with standard IF training practices. To inherit pre-trained weights, the network retains the SigFormer.

Experimental Results

Settings. Hyper-parameters are set as follows: in Eq. (2), $\alpha_1 = 2$, $\beta_1 = 1$, $\gamma_1 = 10$; in Eq. (3), $\alpha_2 = 1$, $\beta_2 = 10$. The EWT decomposer uses 2 sub-bands (two submodules). We train with Adam (learning rate 1×10^{-4}, 100 epochs, batch size 16) on a workstation with an i9-14900k CPU, RTX 4090 GPU, and 128 GB RAM, using PyTorch with CUDA 12.1.

Datasets. For VIS-IR, we train on MSRS (Tang et al. 2022) and test on M^{3}FD (Liu et al. 2022) and TNO (Toet 2017). For MIF, 334 pairs from the Harvard Medical website[1] are split into $300/34$ train/val, and the test set contains 21 MRI-CT and 42 MRI-PET pairs (metrics as in VIS-IR). For MFIF, we train on RealMFF (Zhang et al. 2020a) (710 pairs) and test on LYTRO (Nejati, Samavi, and Shirani 2015) and MFFW (Xu et al. 2020b). For MEF, we use the SICE (Cai, Gu, and Zhang 2018) training set for training, and the SICE test set and MEFB (Zhang 2021) for testing. For the pretext task, Flickr25k (Wang et al. 2008) (25,000 images) is used to synthesize four variants per image, yielding four benchmark sets ($4 \times 25{,}000$ images).

Metrics. We selected Q_G (Li, Kang, and Hu 2013), Q_M (Wang and Liu 2008), Q_P (Zhao, Laganiere, and Liu 2007), MI (Wang and Bovik 2002), SD (Zhao et al. 2023) and $VIFF$ (Wang et al. 2004). Except for the no-reference metric SD, all reference-based metrics use the two source images as supervision, following the mainstream practice in IF.

Comparison with State-of-the-art Methods

Infrared-Visible Image Fusion (VIF). Comparison Models: U2Fusion (Xu et al. 2020a), DeFusion (Liang et al. 2022), CDDFuse (Zhao et al. 2023), LRRNet (Li et al. 2023), MURF (Xu, Yuan, and Ma 2023), EMMA (Zhao et al. 2024a), Text-Difuse (Zhang, Cao, and Ma 2024), FILM (Zhao et al. 2024b), C2RF (Tang et al. 2025b), GIFNet (Cheng et al. 2025), DCEvo (Liu et al. 2025a), and Omini-Fuse (Zhang et al. 2025) are used for comparison. **Qualitative Comparison:** The first three columns of Fig. 4 show qualitative results on two test image pairs. Our model preserves salient thermal targets from infrared images and clear backgrounds from visible images. CDDFuse is competitive but shows lower contrast than our model. LRRNet and DCEvo retain more visible details, but key infrared features are weakened. **Quantitative Comparison:** As in Table 1, our model performs best on most metrics.

Medical Image Fusion (MIF). Comparison Models: U2Fusion, DeFusion, CDDFuse, LRRNet, MURF, EMMA, Text-Difuse, FILM, C2RF, GIFNet, DCEvo, and MTG-Fusion (Wang et al. 2025c) are used. **Qualitative Compar-**

[1]https://www.med.harvard.edu/AANLIB/home.html

			Dataset:M^{3}FD						Dataset:TNO					
Task	Method	Pub/Year	$Q_G\uparrow$	$Q_M\uparrow$	$Q_P\uparrow$	$MI\uparrow$	$SD\uparrow$	$VIFF\uparrow$	$Q_G\uparrow$	$Q_M\uparrow$	$Q_P\uparrow$	$MI\uparrow$	$SD\uparrow$	$VIFF\uparrow$
VIF	U2Fusion	TPAMI 20	0.521	0.488	0.416	2.361	33.02	0.672	0.452	0.530	0.293	1.785	31.49	0.575
	DeFusion	ECCV 22	0.310	0.371	0.370	2.421	30.51	0.549	0.339	0.461	0.285	2.006	30.99	0.553
	CDDFuse	CVPR 23	0.572	0.808	0.455	3.873	41.29	0.781	0.466	0.643	0.379	3.069	44.80	0.730
	LRRNet	TPAMI 23	0.398	0.415	0.371	2.812	30.13	0.565	0.364	0.466	0.230	2.387	38.57	0.538
	MURF	TPAMI 23	0.216	0.269	0.079	2.435	38.31	0.401	0.348	0.373	0.251	1.623	33.69	0.491
	EMMA	CVPR 24	0.505	0.512	0.460	3.771	38.00	0.760	0.468	0.504	0.355	2.909	46.65	0.704
	Text-Difuse	NIPS 24	0.217	0.340	0.232	2.054	**47.48**	0.667	0.254	0.326	0.064	1.680	**46.66**	0.234
	FILM	ICML 24	0.528	0.980	0.436	3.608	41.19	**0.806**	0.490	0.803	0.364	2.851	43.77	0.725
	C2RF	IJCV 25	0.276	0.287	0.099	2.498	38.05	0.320	0.335	0.505	0.185	2.064	40.34	0.461
	GIF-Net	CVPR 25	0.396	0.319	0.319	2.529	42.69	0.557	0.343	0.383	0.217	1.929	42.86	0.500
	DCEvo	CVPR 25	0.569	0.961	**0.469**	4.007	39.85	0.801	**0.570**	0.785	0.328	3.105	40.78	0.726
	OminiFuse	TPAMI 25	0.305	0.362	0.277	3.208	39.31	0.557	0.249	0.403	0.191	2.287	40.46	0.529
	OURS	-	**0.588**	**0.995**	0.467	**4.575**	43.24	**0.806**	0.552	**0.879**	**0.394**	**3.240**	39.42	**0.790**
	Method	Pub/Year	Dataset:MRI-CT						Dataset:MRI-PET					
MIF	U2Fusion	TPAMI 20	0.624	0.140	0.281	2.929	51.68	0.366	0.667	0.125	0.333	3.452	56.41	0.439
	DeFusion	ECCV 22	0.625	0.123	0.293	3.168	66.17	0.465	0.750	0.151	0.317	2.312	63.46	0.521
	CDDFuse	CVPR 23	0.485	0.109	0.030	2.321	78.99	0.089	0.831	0.575	0.445	2.736	74.23	0.670
	LRRNet	TPAMI 23	0.582	0.117	0.207	2.757	37.24	0.351	0.713	0.158	0.381	**3.663**	51.77	0.476
	MURF	TPAMI 23	0.663	0.193	0.279	3.065	75.58	0.398	0.653	0.118	0.336	2.222	62.03	0.392
	EMMA	CVPR 24	0.557	0.153	0.316	3.279	79.60	0.495	0.599	0.143	0.318	2.448	78.41	0.536
	Text-Difuse	NIPS 24	0.534	0.110	0.320	3.178	74.53	0.490	0.216	0.144	0.312	2.331	**84.35**	0.443
	FILM	ICML 24	0.589	**0.284**	0.323	3.344	79.09	0.496	0.680	0.505	0.496	2.751	74.09	0.666
	C2RF	IJCV 25	0.539	0.171	0.131	2.953	76.69	0.328	0.623	0.216	0.284	2.183	74.66	0.617
	GIF-Net	CVPR 25	0.607	0.128	0.236	3.067	70.22	0.345	0.631	0.105	0.245	2.299	60.51	0.411
	DCEvo	CVPR 25	0.608	0.173	0.323	3.280	80.44	0.503	0.724	0.627	0.479	2.862	74.30	0.749
	MTG-Fusion	IJCV 25	0.215	0.155	0.264	3.159	80.70	0.480	0.564	0.255	0.398	2.378	67.69	0.525
	OURS	-	**0.685**	0.230	**0.366**	**3.491**	**81.74**	**0.574**	**0.852**	**0.661**	**0.497**	3.050	74.28	**0.820**
	Method	Pub/Year	Dataset:LYTRO						Dataset:MFFW					
MFIF	CUNet	TPAMI 20	0.526	0.553	0.696	5.441	58.70	1.022	0.482	0.455	0.552	4.593	56.33	0.847
	U2Fusion	TPAMI 20	0.580	0.480	0.742	5.677	58.37	1.086	0.537	0.405	0.611	4.876	55.26	0.825
	IFCNN	INFFUS 20	**0.663**	0.947	0.818	6.914	57.55	1.259	0.589	0.658	0.667	5.528	55.94	1.018
	SDNet	IJCV 21	0.599	0.570	0.769	6.217	56.88	1.128	0.432	0.333	0.501	4.945	55.22	0.954
	DeFusion	ECCV 22	0.455	0.325	0.660	5.984	54.39	1.028	0.418	0.296	0.518	5.137	51.55	0.876
	DIFNet	CVPR 22	0.437	0.325	0.688	5.774	49.67	1.032	0.422	0.309	0.577	4.867	46.66	0.890
	FusionDiff	ESWA 23	0.629	0.821	0.783	6.554	56.13	1.188	0.545	0.602	0.659	5.334	53.27	0.993
	MGDN	ACMMM 23	0.662	0.901	0.810	6.655	56.88	1.226	0.606	0.650	0.663	5.558	54.47	1.020
	ZMFF	INFFUS 23	0.631	0.600	0.785	6.235	57.06	1.175	0.552	0.487	0.635	5.092	54.38	0.990
	DeepM^{2}CDL	TPAMI 24	0.639	1.004	0.810	6.441	58.05	1.267	0.582	0.805	0.679	5.372	56.01	1.064
	FILM	ICML 24	0.619	0.567	0.782	6.758	**59.15**	1.283	0.498	0.436	0.544	5.254	**57.10**	0.919
	GIF-Net	CVPR 25	0.442	0.310	0.546	5.515	68.45	0.910	0.367	0.242	0.386	4.775	56.09	0.725
	OURS	-	**0.663**	**1.231**	**0.829**	**6.987**	58.76	**1.311**	**0.644**	**1.004**	**0.709**	**6.000**	56.18	**1.126**
	Method	Pub/Year	Dataset:MEFB						Dataset:SICE					
MEF	CUNet	TPAMI 20	0.384	0.367	0.530	2.822	53.86	1.177	0.383	0.367	0.463	2.021	41.20	1.257
	U2Fusion	TPAMI 20	0.492	0.431	0.658	5.012	53.43	1.239	0.540	0.408	0.611	4.524	41.13	1.257
	IFCNN	INFFUS 20	0.558	0.511	0.683	4.715	61.52	1.254	0.581	0.478	0.643	4.236	47.29	1.282
	MEF-GAN	TIP 20	0.226	0.253	0.153	3.681	64.86	0.997	0.262	0.240	0.202	3.255	47.83	0.854
	SDNet	IJCV 21	0.302	0.291	0.498	4.243	54.15	1.061	0.526	0.634	0.594	4.661	34.32	0.608
	SwinFusion	JAS 22	0.535	0.511	0.621	4.618	63.31	1.201	0.568	0.458	0.554	3.820	47.01	1.158
	DeFusion	ECCV 22	0.417	0.382	0.485	3.934	51.21	0.973	0.406	0.353	0.363	2.993	39.42	0.893
	DIFNet	CVPR 22	0.431	0.373	0.670	5.729	43.68	1.220	0.426	0.331	0.557	4.698	32.31	1.070
	MGDN	ACMMM 23	0.505	0.450	0.637	3.490	52.69	1.281	0.496	0.404	0.546	2.971	40.28	1.285
	HoLoCo	INFFUS 23	0.424	0.373	0.559	3.886	53.89	1.215	0.306	0.259	0.265	3.016	42.71	1.061
	FILM	ICML 24	0.679	0.984	0.716	6.028	**68.82**	1.521	0.600	0.469	0.641	5.449	**54.20**	1.591
	GIF-Net	CVPR 25	0.280	0.268	0.456	4.602	21.57	0.343	0.291	0.368	0.466	4.376	24.25	0.544
	OURS	-	**0.699**	**1.290**	**0.717**	**6.358**	65.85	**1.417**	**0.731**	**0.560**	**0.671**	**5.816**	51.02	**1.615**

Table 1: Average quantitative results of various fusion models on VIF, MIF, MFIF, and MEF tasks.

VIF (TNO)	$Q_G\uparrow$	$Q_M\uparrow$	$Q_P\uparrow$	$MI\uparrow$	$SD\uparrow$	$VIFF\uparrow$	MFIF (LYTRO)	$Q_G\uparrow$	$Q_M\uparrow$	$Q_P\uparrow$	$MI\uparrow$	$SD\uparrow$	$VIFF\uparrow$
w/o-all SM	0.4018	0.6173	0.2568	2.4025	28.345	0.4751	w/o-all SM	0.4018	0.6173	0.6088	4.7025	41.896	0.9751
w/o-high SM	0.4368	0.7125	0.3653	2.6872	35.294	0.5520	w/o-high SM	0.4268	0.7344	0.6953	5.6872	52.319	1.1977
w/o-low SM	0.4111	0.7204	0.3453	2.6584	33.158	0.5970	w/o-low SM	0.4118	0.7284	0.6553	5.5441	47.628	1.0297
w/o-EWT	0.4230	0.7652	0.3258	2.6796	29.876	0.5166	w/o-EWT	0.4237	0.7331	0.7134	5.6920	55.472	1.0211
Trans→CNN	0.4592	0.7582	0.2742	2.6814	37.621	0.5795	Trans→CNN	0.5592	0.6982	0.7042	5.6814	50.234	1.0795
EWT→DWT	0.4237	0.7331	0.3834	2.6920	36.912	0.6211	EWT→DWT	0.5830	0.7652	0.8058	6.0796	43.571	1.1166
w/o-Pretrained	0.4781	0.7844	0.3732	2.7018	31.427	0.4783	w/o-Pretrained	0.5781	0.7344	0.7132	5.7018	48.930	1.0783
w/o-PLGN	0.4823	0.7436	0.3122	2.6757	38.791	0.5523	w/o-PLGN	0.5823	0.7036	0.7022	5.6757	57.204	1.0523
Default	**0.5516**	**0.8794**	**0.3936**	**3.2423**	**39.422**	**0.7904**	Default	**0.6631**	**1.2306**	**0.8285**	**6.9870**	**58.763**	**1.3105**

Table 2: Quantitative results of ablation study on VIF and MFIF tasks. Results on other tasks are in supplementary material.

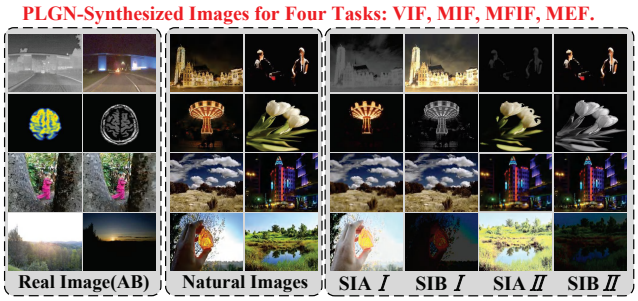

Figure 5: Real source images and PLGN-synthesized images. PLGN synthesizes images consistent with real multi-source data. SIA and SIB denote the synthesized modality-A/B images generated from natural images.

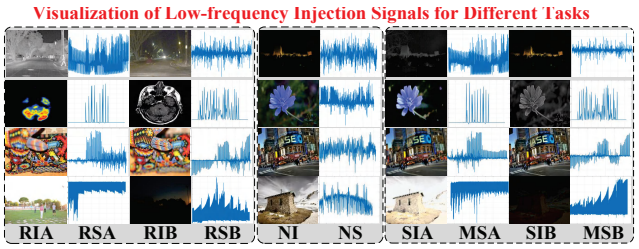

Figure 6: Comparison of real, natural, and modulated signals. RIA/RIB are real modality A/B images with signals RSA/RSB; NI/NS denote the natural image and its signal; SIA/SIB are synthesized modality A/B images with modulated signals MSA/MSB.

ison: Columns 4 to 9 of Fig. 4 show CT-MRI and PET-MRI fusion results. Our model captures MRI soft tissue, key white areas (bones) from CT, and color from PET, benefiting from synthetic datasets that preserve source characteristics. Other methods show different levels of color distortion. **Quantitative Comparison:** Results in Table 1 show that our model outperforms others on most metrics. FILM and MURF are competitive.

Multi-Focus Image Fusion (MFIF). Comparison Models: CU-Net (Deng and Dragotti 2020), U2Fusion, IFCNN (Zhang et al. 2020b), SDNet (Zhang and Ma 2021), DeFusion, DIFNet (Jung et al. 2020), FusionDiff (Li et al. 2024a), MGDN (Guan et al. 2023), ZMFF (Hu et al. 2023), DeepM^{2}CDL (Deng et al. 2023), FILM, and GIF-Net. **Qualitative Comparison:** Columns 10–12 of Fig. 4 show that our model better preserves focused edge details, while others suffer defocus spread near boundaries, benefiting from the pretext stage capturing defocus blur. **Quantitative Comparison:** In Table 1, our model ranks first on 7/8 metrics and second on SD; IFCNN remains competitive due to its MFIF-specific design.

Multi-Exposure Image Fusion (MEF). Comparison Models: CU-Net, U2Fusion, IFCNN, MEF-GAN(Xu, Ma, and Zhang 2020), SDNet, SwinFusion (Ma et al. 2022), DeFusion, DIFNet, MGDN, HoLoCo (Liu et al. 2023b), FILM, and GIF-Net. **Qualitative Comparison:** Columns 13–15 of Fig. 4 show our model achieves more balanced exposure while preserving underexposed details/edges; others exhibit uneven exposure (SwinFusion, MGDN, MFF-GAN), detail loss (FILM, DeFusion), or color distortion. **Quantitative Comparison:** Table 1 over 12 models indicates SigFusion is competitive in exposure regulation and complementary feature extraction.

Ablation Study

PLGN combines SM and SigFormer. Key results are shown here; more experiments are in the Appendix.

Effect of SM. SMs separately process high- and low-frequency signals. We evaluate four variants: (1) removing all SMs, (2) removing only high-frequency SMs, (3) removing only low-frequency SMs, and (4) keeping all. Table 2 shows that any removal degrades performance, confirming SMs are crucial for modulating natural-image signals.

Effect of SigFormer. SigFormer couples a Transformer with EWT. We compare (1) Transformer only (no EWT), (2) a CNN in place of the Transformer, (3) DWT in place of EWT, and (4) the default design. Table 2 shows the default SigFormer performs best, demonstrating the benefit of combining Transformer and EWT.

Effect of Pretext Task (PT). PT is central to our framework. We study three variants: (1) no PT, (2) PT without PLGN (natural images without explicit signal characteristics), and (3) the default model. Table 2 shows PT pretraining clearly improves performance, highlighting the value of our dataset synthesis and pseudo-label generation.

Synthetic Dataset Visualization

To validate PLGN, Fig. 5 compares natural images with versions infused with signals from different source modalities, showing distinct changes while remaining visually consistent with the target modality (e.g., infrared intensities, weakened edges, defocus blur). Fig. 6 further shows that modulation drives natural-image signals toward the target modality, confirming the effectiveness of signal-level feature injection.

Broader Impact of PLGN

Our PLGN synthesizes large-scale training datasets from natural images. These datasets can refine various IF models' performance. To validate this, we used the synthetic datasets for pretraining, then fine-tuned these models on their original datasets without altering default settings. Results are presented in supplementary materials, indicating the significance of the proposed PLGN for IF.

Conclusion

We present SigFusion, a unified signal-level self-supervised learning paradigm. To our knowledge, it is the first SSL-based IF method focused on the signal level. It addresses the scarcity of large-scale real-world training data in IF. Using SigFormer, SM, and custom loss functions, natural images are infused with signal characteristics from multi-source images, enabling effective pretext training. Besides, SigFormer also serves as a backbone for IF. Experiments show that SigFusion yields SOTA performance on four IF tasks.

Acknowledgments

This work was supported by the National Natural Science Foundation of China [No. 62401097]; the Natural Science Foundation of Liaoning Province (Doctoral Research Start-up Project) [No. 2024-BS-028]; Fundamental Research Funds for Central Universities, Dalian Minzu University [No. 0854-53].

References

Cai, J.; Gu, S.; and Zhang, L. 2018. Learning a deep single image contrast enhancer from multi-exposure images. *IEEE Transactions on Image Processing*, 27(4): 2049–2062.

Cheng, C.; Xu, T.; Feng, Z.; Wu, X.; Tang, Z.; Li, H.; Zhang, Z.; Atito, S.; Awais, M.; and Kittler, J. 2025. One Model for ALL: Low-Level Task Interaction Is a Key to Task-Agnostic Image Fusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 28102–28112.

Cheng, C.; Xu, T.; and Wu, X.-J. 2023. MUFusion: A general unsupervised image fusion network based on memory unit. *Information Fusion*, 92: 80–92.

Deng, X.; and Dragotti, P. L. 2020. Deep convolutional neural network for multi-modal image restoration and fusion. *IEEE transactions on pattern analysis and machine intelligence*, 43(10): 3333–3348.

Deng, X.; Xu, J.; Gao, F.; Sun, X.; and Xu, M. 2023. DeepM 2 CDL: Deep Multi-scale Multi-modal Convolutional Dictionary Learning Network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Gilles, J. 2013. Empirical wavelet transform. *IEEE transactions on signal processing*, 61(16): 3999–4010.

Guan, Y.; Xu, R.; Yao, M.; Wang, L.; and Xiong, Z. 2023. Mutual-guided dynamic network for image fusion. In *Proceedings of the 31st ACM International Conference on Multimedia*, 1779–1788.

Gui, J.; Chen, T.; Zhang, J.; Cao, Q.; Sun, Z.; Luo, H.; and Tao, D. 2024. A Survey on Self-supervised Learning: Algorithms, Applications, and Future Trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Hermessi, H.; Mourali, O.; and Zagrouba, E. 2021. Multimodal medical image fusion review: Theoretical background and recent advances. *Signal Processing*, 183: 108036.

Hu, X.; Jiang, J.; Liu, X.; and Ma, J. 2023. ZMFF: Zero-shot multi-focus image fusion. *Information Fusion*, 92: 127–138.

Jung, H.; Kim, Y.; Jang, H.; Ha, N.; and Sohn, K. 2020. Unsupervised deep image fusion with structure tensor representations. *IEEE Transactions on Image Processing*, 29: 3845–3858.

Karim, S.; Tong, G.; Li, J.; Qadir, A.; Farooq, U.; and Yu, Y. 2023. Current advances and future perspectives of image fusion: A comprehensive review. *Information Fusion*, 90: 185–217.

Kaur, R.; and Singh, S. 2023. Multi-focus Image Fusion Methods: A Review. In *International Conference on Advanced Computing, Machine Learning, Robotics and Internet Technologies*, 112–125. Springer.

Li, H.; and Wu, X.-J. 2018. DenseFuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing*, 28(5): 2614–2623.

Li, H.; Xu, T.; Wu, X.-J.; Lu, J.; and Kittler, J. 2023. Lrrnet: A novel representation learning guided fusion network for infrared and visible images. *IEEE transactions on pattern analysis and machine intelligence*, 45(9): 11040–11052.

Li, M.; Pei, R.; Zheng, T.; Zhang, Y.; and Fu, W. 2024a. FusionDiff: Multi-focus image fusion using denoising diffusion probabilistic models. *Expert Systems with Applications*, 238: 121664.

Li, S.; Kang, X.; and Hu, J. 2013. Image fusion with guided filtering. *IEEE Transactions on Image processing*, 22(7): 2864–2875.

Li, X.; Liu, J.; Chen, Z.; Zou, Y.; Ma, L.; Fan, X.; and Liu, R. 2024b. Contourlet residual for prompt learning enhanced infrared image super-resolution. In *European Conference on Computer Vision*, 270–288. Springer.

Li, X.; Wang, Z.; Zou, Y.; Chen, Z.; Ma, J.; Jiang, Z.; Ma, L.; and Liu, J. 2025. Difiisr: A diffusion model with gradient guidance for infrared image super-resolution. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 7534–7544.

Liang, P.; Jiang, J.; Liu, X.; and Ma, J. 2022. Fusion from decomposition: A self-supervised decomposition approach for image fusion. In *European Conference on Computer Vision*, 719–735. Springer.

Liu, J.; Fan, X.; Huang, Z.; Wu, G.; Liu, R.; Zhong, W.; and Luo, Z. 2022. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5802–5811.

Liu, J.; Li, S.; Liu, H.; Dian, R.; and Wei, X. 2023a. A lightweight pixel-level unified image fusion network. *IEEE Transactions on Neural Networks and Learning Systems*.

Liu, J.; Li, X.; Wang, Z.; Jiang, Z.; Zhong, W.; Fan, W.; and Xu, B. 2024a. PromptFusion: Harmonized semantic prompt learning for infrared and visible image fusion. *IEEE/CAA Journal of Automatica Sinica*.

Liu, J.; Lin, R.; Wu, G.; Liu, R.; Luo, Z.; and Fan, X. 2024b. Coconet: Coupled contrastive learning network with multi-level feature ensemble for multi-modality image fusion. *International Journal of Computer Vision*, 132(5): 1748–1775.

Liu, J.; Wu, G.; Luan, J.; Jiang, Z.; Liu, R.; and Fan, X. 2023b. HoLoCo: Holistic and local contrastive learning network for multi-exposure image fusion. *Information Fusion*, 95: 237–249.

Liu, J.; Zhang, B.; Mei, Q.; Li, X.; Zou, Y.; Jiang, Z.; Ma, L.; Liu, R.; and Fan, X. 2025a. DCEvo: Discriminative Cross-Dimensional Evolutionary Learning for Infrared and Visible Image Fusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2226–2235.

Liu, Y.; Wang, L.; Cheng, J.; Li, C.; and Chen, X. 2020. Multi-focus image fusion: A survey of the state of the art. *Information Fusion*, 64: 71–91.

Liu, Y.; Zou, Y.; Li, X.; Zhu, X.; Han, K.; Jiang, Z.; Ma, L.; and Liu, J. 2025b. Toward a Training-Free Plug-and-Play Refinement Framework for Infrared and Visible Image Registration and Fusion. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 1268–1277.

Ma, J.; Tang, L.; Fan, F.; Huang, J.; Mei, X.; and Ma, Y. 2022. SwinFusion: Cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA Journal of Automatica Sinica*, 9(7): 1200–1217.

Nejati, M.; Samavi, S.; and Shirani, S. 2015. Multi-focus image fusion using dictionary-based sparse representation. *Information Fusion*, 25: 72–84.

Qu, L.; Liu, S.; Wang, M.; and Song, Z. 2023. Rethinking multi-exposure image fusion with extreme and diverse exposure levels: A robust framework based on Fourier transform and contrastive learning. *Information Fusion*, 92: 389–403.

Tang, L.; Deng, Y.; Yi, X.; Yan, Q.; Yuan, Y.; and Ma, J. 2024. DRMF: Degradation-robust multi-modal image fusion via composable diffusion prior. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 8546–8555.

Tang, L.; Li, C.; and Ma, J. 2025. Mask-DiFuser: A Masked Diffusion Model for Unified Unsupervised Image Fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–18.

Tang, L.; Wang, Y.; Cai, Z.; Jiang, J.; and Ma, J. 2025a. ControlFusion: A Controllable Image Fusion Framework with Language-Vision Degradation Prompts. *Advances in Neural Information Processing Systems*.

Tang, L.; Yan, Q.; Xiang, X.; Fang, L.; and Ma, J. 2025b. C2RF: Bridging Multi-modal Image Registration and Fusion via Commonality Mining and Contrastive Learning. *International Journal of Computer Vision*, 1–19.

Tang, L.; Yuan, J.; Zhang, H.; Jiang, X.; and Ma, J. 2022. PIAFusion: A progressive infrared and visible image fusion network based on illumination aware. *Information Fusion*, 83: 79–92.

Toet, A. 2017. The TNO multiband image data collection. *Data in brief*, 15: 249–251.

Wang, P.-w.; and Liu, B. 2008. A novel image fusion metric based on multi-scale analysis. In *2008 9th international conference on signal processing*, 965–968. IEEE.

Wang, X.; Zhang, L.; Li, X.; and Ma, W.-Y. 2008. Annotating images by mining image search results. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(11): 1919–1932.

Wang, Z.; and Bovik, A. C. 2002. A universal image quality index. *IEEE signal processing letters*, 9(3): 81–84.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Wang, Z.; Li, X.; Duan, H.; and Zhang, X. 2022. A self-supervised residual feature learning model for multifocus image fusion. *IEEE Transactions on Image Processing*, 31: 4527–4542.

Wang, Z.; Li, X.; Zhao, L.; Duan, H.; Wang, S.; Liu, H.; and Zhang, X. 2023. When multi-focus image fusion networks meet traditional edge-preservation technology. *International Journal of Computer Vision*, 131(10): 2529–2552.

Wang, Z.; Zhang, J.; Guan, T.; Zhou, Y.; Li, X.; Dong, M.; and Liu, J. 2025a. Efficient Rectified Flow for Image Fusion. *Advances in Neural Information Processing Systems*.

Wang, Z.; Zhang, J.; Song, H.; Ge, M.; Wang, J.; and Duan, H. 2025b. Highlight What You Want: Weakly-Supervised Instance-Level Controllable Infrared-Visible Image Fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12637–12647.

Wang, Z.; Zhao, L.; Zhang, J.; Song, R.; Song, H.; Meng, J.; and Wang, S. 2025c. Multi-text guidance is important: Multi-modality image fusion via large generative vision-language model. *International Journal of Computer Vision*, 1–23.

Xu, H.; Ma, J.; Jiang, J.; Guo, X.; and Ling, H. 2020a. U2Fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1): 502–518.

Xu, H.; Ma, J.; and Zhang, X.-P. 2020. MEF-GAN: Multi-exposure image fusion via generative adversarial networks. *IEEE Transactions on Image Processing*, 29: 7203–7216.

Xu, H.; Yuan, J.; and Ma, J. 2023. Murf: Mutually reinforcing multi-modal image registration and fusion. *IEEE transactions on pattern analysis and machine intelligence*, 45(10): 12148–12166.

Xu, S.; Wei, X.; Zhang, C.; Liu, J.; and Zhang, J. 2020b. MFFW: A new dataset for multi-focus image fusion. *arXiv preprint arXiv:2002.04780*.

Zhang, H.; Cao, L.; and Ma, J. 2024. Text-DiFuse: An interactive multi-modal image fusion framework based on text-modulated diffusion model. *Advances in Neural Information Processing Systems*, 37: 39552–39572.

Zhang, H.; Cao, L.; Zuo, X.; Shao, Z.; and Ma, J. 2025. OmniFuse: Composite Degradation-Robust Image Fusion with Language-Driven Semantics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhang, H.; and Ma, J. 2021. SDNet: A versatile squeeze-and-decomposition network for real-time image fusion. *International Journal of Computer Vision*, 129(10): 2761–2785.

Zhang, H.; Xu, H.; Tian, X.; Jiang, J.; and Ma, J. 2021. Image fusion meets deep learning: A survey and perspective. *Information Fusion*, 76: 323–336.

Zhang, J.; Huang, J.; Jin, S.; and Lu, S. 2024. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhang, J.; Liao, Q.; Liu, S.; Ma, H.; Yang, W.; and Xue, J.-H. 2020a. Real-MFF: A large realistic multi-focus image dataset with ground truth. *Pattern Recognition Letters*, 138: 370–377.

Zhang, X. 2021. Benchmarking and comparing multi-exposure image fusion algorithms. *Information Fusion*, 74: 111–131.

Zhang, Y.; Liu, Y.; Sun, P.; Yan, H.; Zhao, X.; and Zhang, L. 2020b. IFCNN: A general image fusion framework based on convolutional neural network. *Information Fusion*, 54: 99–118.

Zhao, J.; Laganiere, R.; and Liu, Z. 2007. Performance assessment of combinative pixel-level image fusion based on an absolute feature measurement. *Int. J. Innov. Comput. Inf. Control*, 3(6): 1433–1447.

Zhao, Z.; Bai, H.; Zhang, J.; Zhang, Y.; Xu, S.; Lin, Z.; Timofte, R.; and Van Gool, L. 2023. Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5906–5916.

Zhao, Z.; Bai, H.; Zhang, J.; Zhang, Y.; Zhang, K.; Xu, S.; Chen, D.; Timofte, R.; and Van Gool, L. 2024a. Equivariant multimodality image fusion. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 25912–25921.

Zhao, Z.; Deng, L.; Bai, H.; Cui, Y.; Zhang, Z.; Zhang, Y.; Qin, H.; Chen, D.; Zhang, J.; Wang, P.; et al. 2024b. Image Fusion via Vision-Language Model. *arXiv preprint arXiv:2402.02235*.

Zhu, R.; Li, X.; Zhang, X.; and Wang, J. 2021. HID: the hybrid image decomposition model for MRI and CT fusion. *IEEE Journal of Biomedical and Health Informatics*, 26(2): 727–739.

Zhu, Z.; Yin, H.; Chai, Y.; Li, Y.; and Qi, G. 2018. A novel multi-modality image fusion method based on image decomposition and sparse representation. *Information Sciences*, 432: 516–529.

Zou, Y.; Chen, Z.; Zhang, Z.; Li, X.; Ma, L.; Liu, J.; Wang, P.; and Zhang, Y. 2025. Contourlet refinement gate framework for thermal spectrum distribution regularized infrared image super-resolution. *International Journal of Computer Vision*.

Zou, Y.; Li, X.; Jiang, Z.; and Liu, J. 2024. Enhancing neural radiance fields with adaptive multi-exposure fusion: A bilevel optimization approach for novel view synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7882–7890.