When Integral Meets Decomposition: A Signal-Level Self-Supervised Feature Decompose Paradigm for Multi-Modal Image Fusion

Zeyu Wang[1] **Jiayu Wang**[1, 2] **Haiyu Song**[1, *] **Haoran Duan**[3, *]
[1]College of Computer Science and Engineering, Dalian Minzu University, Dalian, China
[2]School of Artificial Intelligence and Robotics, Hunan University, Changsha, China
[3] Department of Automation, Tsinghua University, Beijing, China
{wangzeyu, shy}@dlnu.edu.cn jiayuwang@hnu.edu.cn haoran.duan@ieee.org

Abstract

Multimodal image fusion (MMIF) aims to integrate complementary information from different modalities into a high-quality fused image and support downstream tasks. Recently, feature decomposition has become an important paradigm by separating source images into common and modality-specific unique features. However, existing methods lack clear supervision because ground-truth (GT) decomposition feature maps are unavailable. They usually combine multiple image-level metrics as losses, which are inherently incomplete and may conflict since each pixel couples attributes such as texture, edge, and contour. To address this, we propose a 1D signal-level self-supervised feature decomposition paradigm. Our core insight is to reformulate feature decomposition from unclear 2D image-level supervision into an integral-driven 1D signal-level optimization problem. This objective-level reformulation uses the 1D signal form to compute the integral constraint. The decomposer is optimized by the integral area between common and original signals, enabling more stable optimization with a clear optimization objective. Our model follows a two-stage SSL framework. Stage I designs dual pretext tasks for integral-driven decomposition at the signal level and structure-preserving reconstruction at the image level. Stage II fuses unique features and combines them with common features to reconstruct the fused image. Experiments on representative MMIF tasks show state-of-the-art (SOTA) performance. Code: github.com/Wangjiayu0512/SIDFusion.

1 Introduction

Multimodal image fusion (MMIF) integrates complementary information from different modalities to produce fused images [32, 18, 14, 10]. Typical tasks include medical image fusion (MIF) and infrared–visible fusion (VIF) [19, 52, 44, 21, 40], where modalities capture complementary cues, such as thermal radiation in infrared images and textures in visible images [12, 11, 13]. MMIF also benefits downstream tasks, including medical image segmentation [22] and object detection [20, 23].

Currently, feature decomposition paradigm has become a key method [15]. It decomposes source images into two complementary feature types: common features representing shared scene information and unique features reflecting modality-specific details. This strategy facilitates the interaction and fusion of modality-specific information while preserving complete scene structure.

However, existing 2D image-level methods face a key challenge that decomposer training heavily depends on handcrafted or empirical loss design caused by the lack of ground truth (GT) com-

*Corresponding authors

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

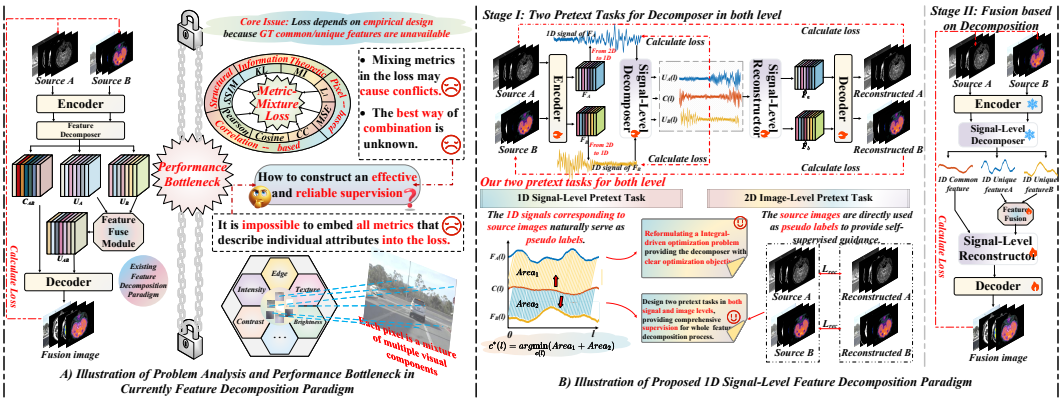


Figure 1: Existing decomposition paradigms *vs* Ours. We introduce signal-level decomposition and image-level reconstruction pretext tasks for accurate decomposition and high-quality fusion.

mon/unique feature maps as supervision. Existing methods have to empirically embed multiple metrics into the loss function [49, 1, 2, 5, 9]. It is well known that each image pixel is a coupling of multiple visual attributes, such as edges, textures, contrast, light and contour. Obviously, it is almost impossible to incorporate all metrics that describe specific visual attributes into the loss, which makes an intrinsic difficulty. More importantly, we have no prior knowledge of whether these metrics are mutually conflicting, nor can we determine which combinations lead to better performance. This raises a central question: **How can we construct an effective and reliable supervisory objective?**

To address this, we propose a 1D signal-level self-supervised feature decomposition paradigm. Our core insight is to reformulate feature decomposition from an unclear supervision problem requiring simultaneous optimization of multiple visual attributes into an integral-driven mathematical problem. This reformulation shifts the focus from architectural design to the construction of a clear supervisory objective for feature decomposition. Accordingly, we design a tailored loss. In this formulation, one only needs to optimize the integral area between the common and source signals. It avoids conflicts from metric-mixing losses and enables disentanglement of coupled information. Under this integral-driven objective, the decomposer achieves more stable and accurate optimization.

Holistically, our model follows a two-stage self-supervised learning (SSL) framework [30]. In Stage I, we design dual pretext tasks at both the signal- and image-level. For decomposition, motivated by the reformulated problem, we design a novel signal-level pretext task. Features from source images are transformed into 1D signals and decomposed into low- and high-frequency components via DWT [27]. The network then acts as an integral solver, estimating common features from paired low-frequency bands and extracting unique features from high-frequency bands of each modality. Paired source signals directly provide pseudo labels for strict mathematical constraints. For reconstruction, we return to the image-level, where original images supervise spatial structure preservation. In Stage II, we first fuse the unique feature signals, then combine them with the common signals, and finally apply IDWT to reconstruct fused images with complete information. Experimental results demonstrate that our method achieves state-of-the-art (SOTA) performance, validating the proposed paradigm.

Our contributions are summarized as follows:

- We propose a novel self-supervised feature decomposition paradigm at 1D signal-level to enable accurate feature decomposition.
- We reformulate the feature decomposition problem from an image-level supervision problem with unclear objectives into a signal-level integral optimization problem with an explicit mathematical objective. This clear optimization target reduces the ambiguity caused by empirical metric-mixture losses and enables more accurate feature decomposition.
- We design complementary dual pretext tasks at both levels. The signal-level task offers strong mathematical constraints for decomposition, while the image-level task ensures reliable spatial structures. They achieve both precise decomposition and faithful reconstruction.

2 Related Work

Feature Decomposition-based MMIF. Feature decomposition-based fusion methods decompose multimodal images into two complementary types of features: common features and unique features. This strategy not only promotes sufficient interaction and fusion of cross-modal complementary information, but also helps preserve the overall scene structure. Several methods adopt this idea. For instance, DIDFuse [51] designs a dedicated feature decomposition module to explicitly separate common and unique features in an end-to-end manner; FD-Fuse [5] imposes indirect constraints on the decomposition process via pretext tasks; CU-Net [6] formulates feature decomposition as a mathematical optimization problem and uses sparse convolutions to solve a hand-crafted objective function. Furthermore, external prior knowledge from large language models is used to guide the decomposition process. MTG-Fusion [42] leverages a large language model to generate multiple descriptive texts and uses their embeddings to guide visual feature decomposition. However, since GT common and unique features are unavailable, these decomposers still lack reliable supervision. We address this issue by reformulating feature decomposition over signalized features as an integral-driven optimization problem, which provides an explicit objective for accurate feature separation.

Self-Supervised Learning. Self-supervised learning constructs pretext tasks to generate supervision from unlabeled data, usually through pretraining and downstream adaptation [54]. Due to the lack of GT in MMIF, many fusion methods adopt SSL strategies. DeFusion [15] learns cross-modal semantic relations via masked image restoration. SMFuse [24] adopts a self-supervised scheme to generate accurate masks for multi-focus image fusion. Wang et al. use image super-resolution as a pretext task, treating high-resolution images as pseudo labels for low-resolution inputs [39]. EMMA [50] exploits image equivariance and generates supervision signals from geometrically transformed natural images. However, these methods are not specifically designed for the feature decomposition paradigm and cannot fundamentally resolve the absence of GT for common and unique features. Therefore, we design an SSL paradigm with dual pretext tasks. We transform the feature decomposition problem into an integral solving problem defined at the 1D signal-level, which naturally yields a pretext task with a clear mathematical objective, using 1D signals of the source images to provide direct supervision. At the image-level, we use the source images as supervision for a reconstruction task, thus providing comprehensive supervision for the entire feature decomposition process.

Signal Modeling-based Image Fusion. Recently, SigFusion [38] proposes a unified self-supervised paradigm from the signal-level perspective, integrating dataset synthesis and image fusion into one framework. However, it is fundamentally different from our work. SigFusion uses signal-level modeling mainly to learn modality-specific signal distributions and synthesize large-scale training data, aiming to alleviate the lack of realistic training pairs. Our method uses signalization to reformulate the decomposition problem itself, focusing on the lack of direct supervision for common and unique features and providing the decomposer with an explicit integral-driven objective.

3 Method

This section describes the workflow of our proposed model. As shown in Fig. 2, our model consists of an encoder ζ, a signal-level feature decomposer $SD(\cdot)$, an MLP based feature fusion module δ and a decoder ξ. Both the encoder and decoder are composed of Restormer blocks [49]. Following the self-supervised learning (SSL) paradigm, our model adopts a two-stage paradigm. Stage I performs signal-level decomposition and reconstruction pretraining to obtain accurate common and unique features. Stage II conducts cross-modal fusion on the decomposed unique features and combines them with the common features to generate the final fused image.

3.1 Problem Formulation and Modeling

In MMIF, mainstream methods based on feature decomposition paradigm typically follow a two-stage pipeline. Given a pair of source images (I_A, I_B), a decomposer $D(\cdot)$ first extracts the common feature C and unique features (U_A, U_B):

$$C, U_A, U_B = D(I_A, I_B). \tag{1}$$

Then, these features are fused to obtain fused image I_f:

$$I_f = F(C, U_A, U_B). \tag{2}$$

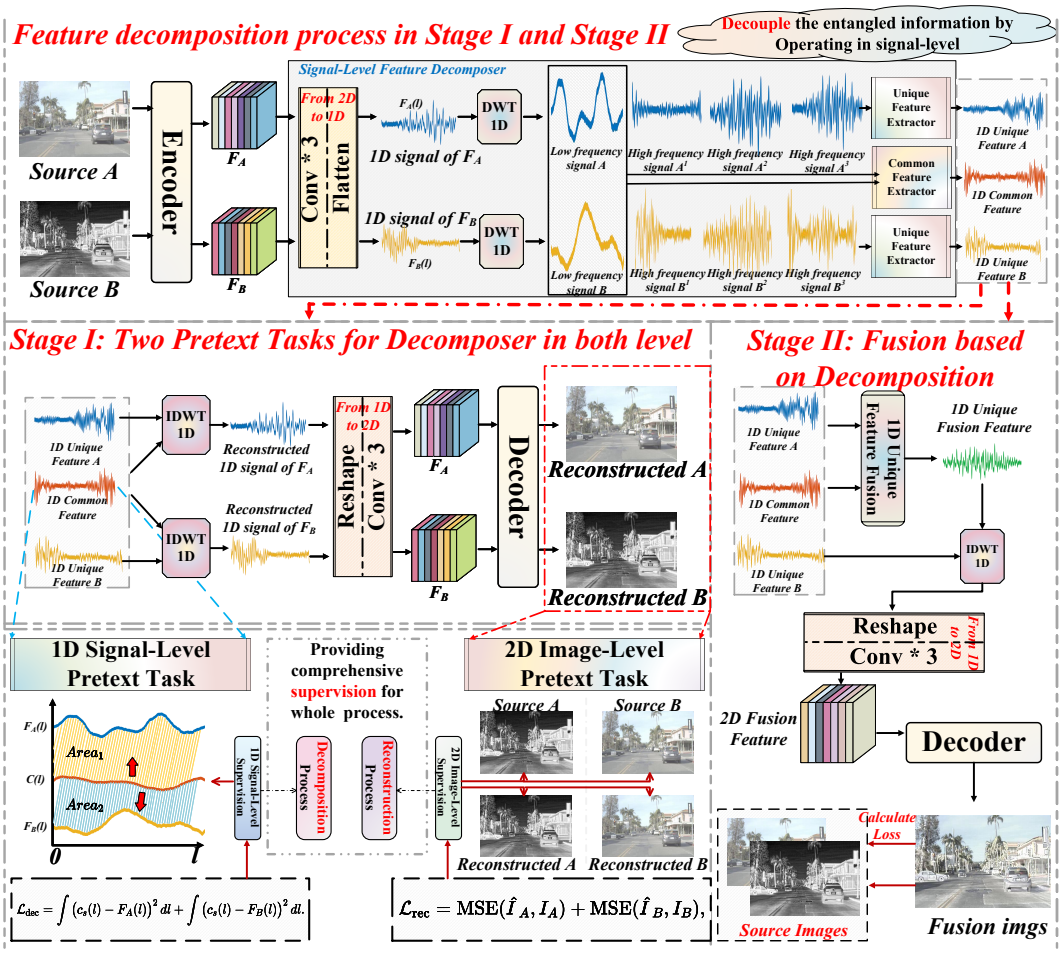


Figure 2: Schematic of the proposed model. It adopts a two-stage self-supervised paradigm. Stage I: feature decomposition via 1D signal-level and 2D image-level pretext tasks; Stage II: fuses the decomposed features to reconstruct the final fused image.

However, these methods suffer from a fundamental limitation in supervision. Because ground-truth common and unique feature maps are unavailable, the training objective is inevitably constructed in a handcrafted or empirical manner, typically by mixing multiple attribute metrics. Yet each pixel couples diverse visual attributes such as edges, textures and contours, making it impractical to cover all attributes with metrics. Consequently, the final objective is often incomplete and may even introduce conflicts, which prevents accurate and stable decomposition. Therefore, it is crucial to design a decomposition paradigm with a clear and reliable supervisory objective.

Reformulating the Feature Decomposition Problem. To address these issues, we propose a signal-level feature decomposition paradigm that reformulates the decomposition problem from an ill-posed 2D image-level problem into a well-defined optimization problem in the 1D signal domain.

Given a source image pair (I_A, I_B) and their 1D signal representations $(s_A(l), s_B(l))$, we optimize the common signal $c_s(l)$ by minimizing the accumulated integral area induced by its deviation from the two original source signals:

$$c_s^*(l) = \arg\min_{c_s} \int \big(c_s(l) - s_A(l)\big)^2 dl + \int \big(c_s(l) - s_B(l)\big)^2 dl. \tag{3}$$

In this formulation, the common signal is encouraged to lie close to both source signals in the integral sense, so that it can capture the component consistently supported by the two modalities. Compared

with image-level metric mixing, this objective provides a more direct criterion for common feature estimation and reduces the influence of local fluctuations or modality-specific disturbances.

3.2 Stage I: Self-supervised Feature Decomposition with Dual Pretext Tasks

In Stage I, we design two complementary pretext tasks for self-supervised common and unique feature extraction. The signal-level task provides an explicit objective for feature decomposition, while the image-level task preserves spatial structures through reconstruction. Procedure is as follows.

Feature Decomposition based on the Signal-level Pretext Task.
Given a pair of source images I_A, I_B, we first extract high-level feature maps via an encoder ζ:

$$F_A, F_B = \zeta(I_A), \zeta(I_B), \tag{4}$$

These features are then transformed into 1D signal-level representations:

$$s_A(l), s_B(l) = \mathcal{H}\big(F_A, F_B\big), \tag{5}$$

where $\mathcal{H}$ denotes the transformation with a 1×1 convolution followed by flattening and l indexes the 1D positions corresponding to spatial locations. To enable more fine-grained feature decomposition, we decompose each signal into low- and high-frequency sub-bands via DWT 1D:

$$L_m(l), \{H_m^i(l)\}_{i=1}^N = \mathrm{DWT}\big(s_m(l)\big), \quad m \in \{A, B\}. \tag{6}$$

Low-frequency components mainly encode global structure shared across modalities, while high-frequency components capture unique details such as textures and edges.

Based on these characteristics, we design a signal decomposer $SD(\cdot)$ with two paths: the common-path takes the low-frequency pair $(L_A(l), L_B(l))$ as input and estimates the common signal $c_s(l)$:

$$c_s(l) = SD\big(L_A(l), L_B(l)\big). \tag{7}$$

The unique-path takes high-frequency sub-bands of each modality to estimate unique signals $u_{s^A}(l)$ and $u_{s^B}(l)$:

$$u_{s^A}(l), u_{s^B}(l) = SD\big(\{H_A^i(l)\}_{i=1}^N, \{H_B^i(l)\}_{i=1}^N\big). \tag{8}$$

To optimize the common feature, we adopt a tailored integral-driven loss $\mathcal{L}_{\text{dec}}$:

$$\mathcal{L}_{\text{dec}} = \int \big(c_s(l) - s_A(l)\big)^2 dl + \int \big(c_s(l) - s_B(l)\big)^2 dl, \tag{9}$$

where the 1D signals $s_A(l)$ and $s_B(l)$ serve as pseudo labels. This loss enforces global consistency along the entire signal, suppressing local noise and guiding accurate estimation of $c_s(l)$.

Image Reconstruction based on the Image-level Pretext Task.
The reconstruction process mirrors the decomposition and uses the decomposed features to recover the original images, providing an image-level constraint.

First, $c_s(l)$ and $u_{s^m}(l)$ are fed into the inverse DWT, denoted as DWT^T, to reconstruct the 1D signal:

$$\hat{s}_m(l) = \mathrm{DWT}^T\big(c_s(l), u_{s^m}(l)\big), \quad m \in \{A, B\}. \tag{10}$$

Then, the reconstructed signals are mapped back to 2D feature maps by the inverse transformation $\mathcal{H}^T$ and reconstructed to the source images by a decoder ξ:

$$\hat{I}_m = \xi\big(\mathcal{H}^T(\hat{s}_m(l))\big), \quad m \in \{A, B\}. \tag{11}$$

In this reconstruction process, we employ I_A and I_B act as pseudo labels to preserve spatial structure:

$$\mathcal{L}_{\text{rec}} = \mathrm{MSE}(\hat{I}_A, I_A) + \mathrm{MSE}(\hat{I}_B, I_B), \tag{12}$$

Loss Function of Stage I. The overall loss function of Stage I is defined as Eq. 13:

$$\mathcal{L}_{\text{stage1}} = \alpha\mathcal{L}_{\text{dec}} + \beta\mathcal{L}_{\text{rec}}, \tag{13}$$

where α and β are hyperparameters. Through these dual-pretext tasks, Stage I provides a clear optimization target for decomposition at 1D signal-level and enforces structural correctness in 2D image-level, producing accurate common and unique features for the subsequent fusion stage.

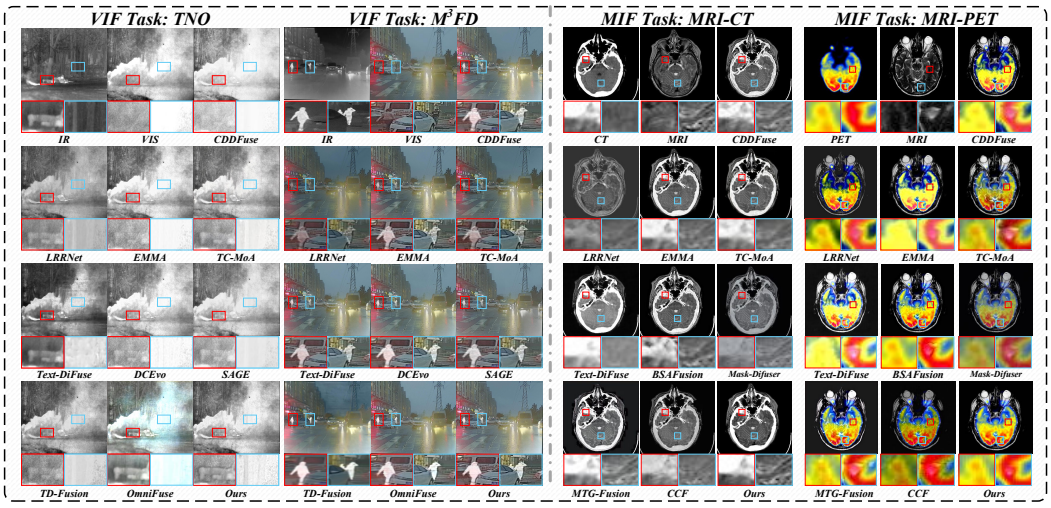


Figure 3: Qualitative comparison of various fusion models.

3.3 Stage II: Feature Decomposition-based Image Fusion

In Stage II, our goal is to generate a high-quality fused image based on the accurate decomposed features obtained from Stage I. Training and inference processes are as follows.

Training. We freeze the encoder ζ and the signal decomposer SD pretrained in Stage I. Given a source image pair (I_A, I_B), we obtain the common signal $c_s(l)$ and unique signals $u_{s^A}(l)$ and $u_{s^B}(l)$ via the Stage I modules. We then fuse the unique features via an MLP-based feature fusion module δ:

$$u_{\mathrm{f}}(l) = \delta\big(u_{s^A}(l), u_{s^B}(l)\big), \tag{14}$$

Next, the fused specific signal $u_{\mathrm{f}}(l)$ and the common signal $c_s(l)$ are passed through the inverse DWT to obtain the fused 1D signal:

$$f_s(l) = \mathrm{DWT}^{\mathrm{T}}\big(c_s(l), u_{\mathrm{f}}(l)\big). \tag{15}$$

This signal is mapped back to a 2D feature map by $\mathcal{H}^T$ and decoded to the fused image:

$$I_f = \xi(\mathcal{H}^T\big(f_s(l)\big)\big). \tag{16}$$

Loss Function of Stage II. Following mainstream practice, we use the source images as supervision to guide fusion. The Stage II loss is defined as:

$$\mathcal{L}_{\text{stage2}} = \gamma\mathcal{L}_{\text{SSIM}}(I_f; I_A, I_B) + \lambda\mathcal{L}_{L1}(I_f; I_A, I_B) + \omega\mathcal{L}_{\text{Grad}}(I_f; I_A, I_B), \tag{17}$$

where $\mathcal{L}_{\text{SSIM}}$ is the SSIM-based structural similarity loss, $\mathcal{L}_{L1}$ is the spatial L_1 loss, $\mathcal{L}_{\text{Grad}}$ is the gradient-domain L_1 loss, and γ, λ, ω are hyperparameters.

Inference. During inference, only the Stage II fusion pipeline is executed. Given a registered source image pair (I_A, I_B), the model $\mathcal{G}$ directly produces the fused image I_f in a single forward pass:

$$I_f = \mathcal{G}(I_A, I_B). \tag{18}$$

4 Experimental Results

4.1 Setting

We adopt the AdamW optimizer with an initial learning rate of 3×10^{-4}, batch size 4, and no weight decay. The learning rate is halved every 20 epochs. All experiments are conducted on Ubuntu 22.04 with an Intel i9-14900K CPU, RTX 4090 GPU, 64GB RAM, and PyTorch 2.3.0. In Stage I, we set α = 1, β = 2 for 200 epochs. In Stage II, γ, λ and ω are set as 1, 1 and 15.6 for 200 epochs, respectively.

Table 1: Quantitative comparison of various fusion models on the VIF and MIF tasks. Best is **bold**; second best is underlined.

VIF Task		M³FD							MSRS							TNO						
Methods	Pub/Year	$Q_{MI}\uparrow$	$Q_{NICE}\uparrow$	$Q_P\uparrow$	$Q_{CB}\uparrow$	$MI\uparrow$	$VIF_p\uparrow$	$Q_Y\uparrow$	$Q_{MI}\uparrow$	$Q_{NICE}\uparrow$	$Q_P\uparrow$	$Q_{CB}\uparrow$	$MI\uparrow$	$VIF_p\uparrow$	$Q_Y\uparrow$	$Q_{MI}\uparrow$	$Q_{NICE}\uparrow$	$Q_P\uparrow$	$Q_{CB}\uparrow$	$MI\uparrow$	$VIF_p\uparrow$	$Q_Y\uparrow$
CDDFuse	CVPR 23	0.5741	0.8123	0.4554	0.4746	3.8730	0.4410	0.8698	0.7576	0.8229	0.5399	0.5669	4.9089	0.5106	0.8267	0.4654	0.8090	0.3789	0.4504	3.0691	0.4428	0.7874
LRRNet	TPAMI 23	0.6353	0.8073	0.3709	0.4346	2.8124	0.3550	0.7209	0.6608	0.8082	0.3372	0.3928	2.8662	0.2933	0.5067	0.3686	0.8062	0.2303	0.4831	2.3868	0.3685	0.7011
EMMA	CVPR 24	0.5564	0.8118	0.4602	0.4775	3.7712	0.4690	0.7979	0.6413	0.8161	0.4961	0.5432	4.1637	0.5230	0.7840	0.4356	0.8081	0.3550	0.5112	2.9089	**0.4847**	0.8072
TC-MoA	CVPR 24	0.5011	0.8096	0.5133	0.4934	3.3330	0.4603	0.8538	0.7446	0.8119	0.4833	0.5558	3.4925	0.4924	0.8825	0.4068	0.8071	0.4020	0.5124	2.5747	0.4312	0.8664
Text-DiFuse	NeurIPS 24	0.4262	0.8057	0.4323	0.3398	2.0535	0.1574	0.2566	0.7252	0.8032	0.5378	0.3660	1.3095	0.1570	0.2122	0.2977	0.8047	0.1167	0.4296	1.8773	0.2638	0.4410
DCEvo	CVPR 25	0.6085	0.8135	0.4692	0.4801	4.0074	0.4485	0.8987	0.6246	0.8170	0.4967	0.5627	4.0493	0.5107	0.8650	**0.5739**	0.8125	0.4280	0.5089	3.6054	0.4605	**0.9060**
SAGE	CVPR 25	0.4566	0.8080	0.4338	0.4647	2.8508	0.4257	0.8324	0.5153	0.8101	0.4348	0.4775	2.9895	0.3455	0.7732	0.3751	0.8060	0.3323	0.4715	2.2033	0.3963	0.7855
TD-Fusion	CVPR 25	0.4494	0.8081	0.4987	**0.5108**	3.0200	0.4629	0.8519	0.7462	0.8082	0.4170	0.5268	2.8434	0.4407	0.7922	0.3640	0.8062	0.3512	0.5288	2.4507	0.4317	0.8183
OmniFuse	TPAMI 25	0.4882	0.8092	0.2770	0.4776	3.2075	0.4432	0.6543	0.7208	0.8079	0.5174	0.4211	2.6641	0.4121	0.4927	0.3814	0.8062	0.1908	0.4374	2.2871	0.3933	0.6406
C2RF	IJCV 25	0.4082	0.8075	0.4989	0.3882	2.4979	0.2241	0.4293	0.6928	0.8039	0.4922	0.4246	1.7292	0.1589	0.3892	0.3535	0.8059	0.1847	0.4362	2.0637	0.2975	0.6155
SigFusion	AAAI 26	**0.7123**	0.8186	0.4665	0.4796	**4.5751**	0.4478	**0.9239**	0.8034	0.8256	**0.5492**	0.5643	4.9123	0.4432	0.8772	0.5296	0.8114	0.3936	0.5064	3.2423	0.4455	0.8911
Ours	-	0.6608	**0.8209**	**0.5735**	0.4998	4.1050	**0.4694**	0.8988	**0.8405**	**0.8290**	0.5483	**0.5798**	**5.4547**	**0.5253**	**0.8936**	0.5711	**0.8128**	**0.4318**	**0.5346**	**3.8015**	0.4768	0.8450
MIF Task		**MRI-CT**							**MRI-PET**							**MRI-SPECT**						
Methods	Pub/Year	$Q_{MI}\uparrow$	$Q_{NICE}\uparrow$	$Q_P\uparrow$	$Q_{CB}\uparrow$	$MI\uparrow$	$VIF_p\uparrow$	$Q_Y\uparrow$	$Q_{MI}\uparrow$	$Q_{NICE}\uparrow$	$Q_P\uparrow$	$Q_{CB}\uparrow$	$MI\uparrow$	$VIF_p\uparrow$	$Q_Y\uparrow$	$Q_{MI}\uparrow$	$Q_{NICE}\uparrow$	$Q_P\uparrow$	$Q_{CB}\uparrow$	$MI\uparrow$	$VIF_p\uparrow$	$Q_Y\uparrow$
CDDFuse	CVPR 23	0.8403	0.8098	0.4122	0.6911	3.7522	0.3735	0.9047	0.7856	0.8064	0.4448	**0.7213**	2.7359	0.3872	**0.9125**	0.8778	0.8068	0.6289	**0.7305**	2.7296	0.4480	0.9118
LRRNet	TPAMI 23	0.6273	0.8066	0.2073	0.2205	2.7570	0.2295	0.3285	0.6427	0.8051	0.3809	0.2129	**3.6627**	0.2807	0.3122	0.7355	0.8056	0.3580	0.1839	**3.8263**	0.2607	0.2763
EMMA	CVPR 24	0.7032	0.8078	0.3160	0.5627	3.2790	0.3575	0.7590	0.6648	0.8055	0.3177	0.5606	2.4481	0.3577	0.7433	0.7086	0.8057	0.3368	0.4834	2.3611	0.3660	0.6633
TC-MoA	CVPR 24	0.7279	0.8080	0.3686	0.5838	3.3169	0.3751	0.7782	0.6758	0.8055	0.4692	0.6129	2.4498	0.3900	0.8394	0.7397	0.8059	0.5564	0.5426	2.3331	0.4422	0.7659
Text-DiFuse	NeurIPS 24	0.6456	0.8075	0.2643	0.3239	3.1594	0.3150	0.3891	0.5235	0.8051	0.3118	0.2606	2.3310	0.2973	0.2630	0.5638	0.8053	0.3649	0.1976	2.2253	0.3261	0.2550
CCF	NeurIPS 24	0.7361	0.8079	0.2825	0.3479	3.2844	0.2968	0.4329	0.6439	0.8050	0.2733	0.2637	2.4926	0.2854	0.3114	0.7520	0.8058	0.3911	0.2372	2.9410	0.3356	0.3177
BSAFusion	AAAI 25	0.6929	0.8079	0.3229	0.3905	3.2802	0.3435	0.4578	0.7470	0.8068	0.4788	0.3446	2.8615	0.4197	0.4200	0.8150	0.8072	**0.6393**	0.3106	2.7563	**0.4550**	0.3766
Mask-Difuser	TPAMI 25	0.7171	0.8081	0.3595	0.3443	3.2990	0.3088	0.4206	0.6911	0.8062	0.4399	0.3035	2.7060	0.3676	0.3590	0.7334	0.8062	0.5022	0.2620	2.5270	0.3677	0.3300
MTG-fusion	IJCV 25	0.6610	0.8076	0.3896	0.3466	3.1782	0.2873	0.4343	0.5958	0.8053	0.3977	0.2651	2.3778	0.3043	0.3453	0.6129	0.8053	0.3987	0.2673	2.3096	0.2844	0.3084
C2RF	IJCV 25	0.6374	0.8072	0.1306	0.3464	2.9530	0.2188	0.3220	0.5827	0.8050	0.2842	0.3543	2.1825	0.2205	0.2686	0.6760	0.8055	0.5087	0.3513	2.2566	0.3184	0.2887
SigFusion	AAAI 26	0.8176	0.8086	**0.6689**	0.3560	3.4913	0.3819	0.9063	0.8762	0.8074	0.4965	0.7157	3.0495	**0.4610**	0.4273	0.8310	**0.8083**	0.5162	0.6621	2.8941	0.4418	0.8798
Ours	-	**0.9369**	**0.8108**	0.4034	**0.7008**	**4.0184**	**0.4034**	**0.9439**	**0.8843**	**0.8074**	**0.4967**	0.4740	3.0547	0.4577	0.5808	**0.9778**	**0.8083**	0.5507	0.6697	3.0655	0.4520	**0.9149**

Table 2: Quantitative results of ablation experiments on VIF and MIF tasks.

VIF Task									MIF Task								
Task	Case	Configurations	$Q_{MI}\uparrow$	$Q_{NICE}\uparrow$	$Q_P\uparrow$	$Q_{CB}\uparrow$	$MI\uparrow$	$VIF_p\uparrow$	$Q_Y\uparrow$	Task	Case	Configurations	$Q_{MI}\uparrow$	$Q_{NICE}\uparrow$	$Q_P\uparrow$	$Q_{CB}\uparrow$	$MI\uparrow$	$VIF_p\uparrow$	$Q_Y\uparrow$
VIF	I	w/o signal-level Dec.	0.6282	0.8181	0.5350	0.4716	3.8631	0.4391	0.8729	MIF	I	w/o signal-level Dec.	0.9012	0.8071	0.3627	0.6715	3.6129	0.3689	0.9127
		w/o DWT	0.6425	0.8192	0.5551	0.4860	3.9812	0.4534	0.8847			w/o DWT	0.9185	0.8090	0.3825	0.6880	3.8310	0.3897	0.9312
		w/ IDWT ← learnable Rec.	0.6531	0.8198	0.5631	0.4925	4.0200	0.4594	0.8911			w/ IDWT ← learnable Rec.	0.9253	0.8096	0.3928	0.6905	3.9025	0.3954	0.9361
		w/ signal- ← image-level Dec.	0.6380	0.8185	0.5460	0.4793	3.9022	0.4468	0.8781			w/ signal- ← image-level Dec.	0.9108	0.8086	0.3755	0.6810	3.7420	0.3798	0.9275
		Default setups (ours)	**0.6608**	**0.8209**	**0.5735**	**0.4998**	**4.1050**	**0.4694**	**0.8988**			Default setups (ours)	**0.9369**	**0.8108**	**0.4034**	**0.7008**	**4.0184**	**0.4034**	**0.9439**
	II	w/o signal-level pretext task	0.6315	0.8188	0.5420	0.4868	3.9033	0.4512	0.8799		II	w/o signal-level pretext task	0.9144	0.8093	0.3786	0.6869	3.7562	0.3831	0.9293
		w/o image-level pretext task	0.6495	0.8195	0.5525	0.4820	3.9800	0.4553	0.8850			w/o image-level pretext task	0.9228	**0.8294**	0.3866	0.6765	3.8041	0.3912	0.9335
		w/ $L_{\text{dec}} \leftarrow L_1$	0.6463	0.8195	0.5482	0.4817	3.9521	0.4520	0.8832			w/ $L_{\text{dec}} \leftarrow L_1$	0.9202	0.8091	0.3814	0.6851	3.7820	0.3886	0.9318
		Default setups (ours)	**0.6608**	**0.8209**	**0.5735**	**0.4998**	**4.1050**	**0.4694**	**0.8988**			Default setups (ours)	**0.9369**	0.8108	**0.4034**	**0.7008**	**4.0184**	**0.4034**	**0.9439**

4.2 Datasets and Metrics

Datasets. For VIF, we train on 400 image pairs from MSRS [25]. Evaluation uses 361 pairs from MSRS, 100 pairs from M^{3}FD, and 25 pairs from TNO. For MIF, we train on 453 registered medical image pairs from the Harvard Medical website [34]. For testing, we select 21 MRI–CT pairs [49], 42 MRI–PET pairs [49], and 73 MRI–SPECT pairs [49]. All test sets are disjoint from the training set.
Metrics. We adopt widely used image quality metrics following mainstream practice, including Q_{MI} [28], Q_{NICE} [37], Q_P [45], Q_{CB} [41], MI [28], VIF_P [29], and Q_Y [46].

4.3 Comparison with SOTA

Comparison Algorithms. For VIF task, we select CDDFuse [49], LRRNet [8], EMMA [50], TC-MoA [53], Text-DiFuse [47], DCEvo [20], SAGE [43], TDFusion [1], OmniFuse [48], C2RF [33] and SigFusion[38]. For MIF task, we select CDDFuse, LRRNet, EMMA, TC-MoA, Text-DiFuse, CCF [4], BSA-Fusion [7], Mask-Difuser [31], MTG-Fusion [42], C2RF [33] and SigFusion[38].
Qualitative Comparison. Fig. 3 shows the qualitative comparison results of 12 fusion models on four test sets. More results are shown in Appendix D.1. For VIF task, the first 2 groups show results on TNO [36] and M^{3}FD [17]. Benefiting from the 1D signal-level decomposition, our fused images better preserve both the common structure and modality-specific cues. For MIF task, the last two groups are the MRI-CT and MRI-PET datasets results. Our dual-pretext task design provides accurate and comprehensive supervision for both levels, yielding fused images with improved color fidelity.
Quantitative Comparison. Table 1 reports the results of various models on seven metrics across VIF and MIF. Our model achieves superior results in most metrics. We attribute these gains to the signal-level decomposer and dual-pretext design, which provide clear supervision for the common component and enhance cross-modal interaction, yielding a high-quality representation for fusion.

4.4 Ablation Study

To comprehensively verify the effectiveness of each key component, we conduct a systematic ablation study. More results are in Appendix F.
Effect of Signal-Level Decomposer. Our signal-level decomposer transfers the decomposition problem into 1D signal-level. To assess its impact, we conduct the following ablations: (1) removing the signal-level decomposer(Dec.); (2) removing frequency decomposition; (3) replacing IDWT with a learnable reconstructor(Rec.); (4) substituting an image-level decomposer; (5) default setup. The results are presented in Table 2 Case I, which demonstrate the effectiveness of our paradigm.

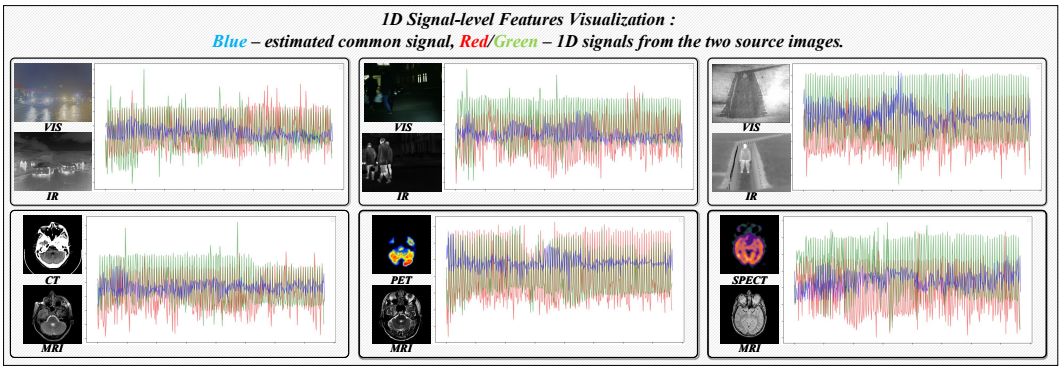


Figure 4: Visualization of intermediate 1D signals. The learned common signal $c_s(l)$ follows the shared trend of the source signals $s_A(l)$ and $s_B(l)$.

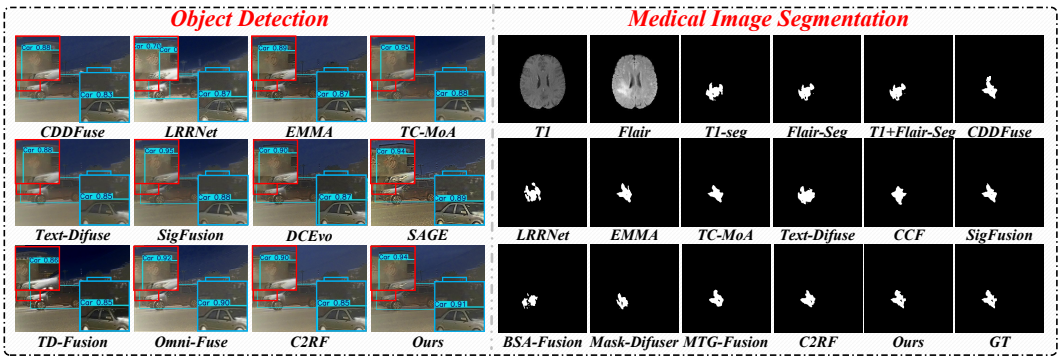


Figure 5: Qualitative results on downstream tasks. Left: object detection performance using fused images generated by different fusion methods. Right: medical image segmentation. Our fused images provide more accurate results.

Effect of Dual Pretext Tasks. To evaluate the dual pretext tasks at both levels, we perform the following ablations: (1) removing the signal-level pretext task; (2) removing the image-level pretext task; (3) replacing the proposed L_{dec} with vanilla L_1 only; (4) default configuration. The results are shown in Table 2 Case II. Apparently, removing either pretext task degrades performance, highlighting the necessity to provide decomposer a clear supervision.

Sensitivity to Key Hyperparameters. We analyze the sensitivity of key hyperparameters. Full analysis and ablation results are provided in Appendix F.

4.5 Signal-Level Feature Visualizations

To further validate the proposed signal-level formulation, we visualize the source signals $s_A(l)$ and $s_B(l)$ and the learned common signal $c_s(l)$ in Fig. 4. The red and green curves denote the two source signals, while the blue curve denotes the estimated common signal. It can be observed that $c_s(l)$ generally lies between $s_A(l)$ and $s_B(l)$ and follows their shared trend, indicating that it captures the component consistently shared by both modalities.

4.6 Downstream Tasks

Object Detection for VIF Task. We evaluate detection on fused images produced by various VIF models. A pre-trained YOLOv12 without fine-tuning[35] is used to ensure fairness. Qualitative results are given in Fig. 5 (left) and quantitative results see Appendix Table 4. Compared with other methods, our model achieves stronger detection performance, suggesting that it better preserves task-relevant complementary cues and provides more favorable inputs for downstream object detection.

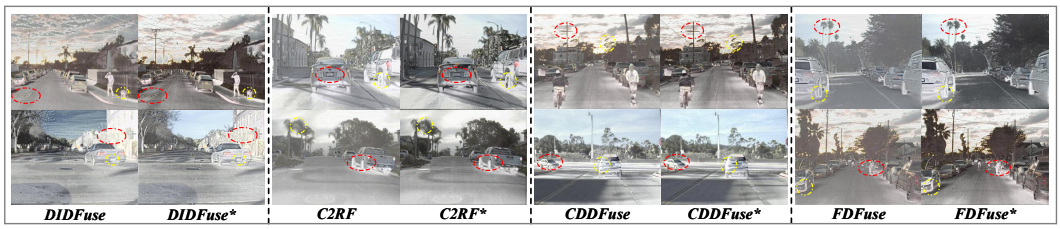


Figure 6: Broader impact of our paradigm. Refined variants (denoted by "*") obtain fused images with clearer structures and more salient target details, as highlighted by the red and yellow circles.

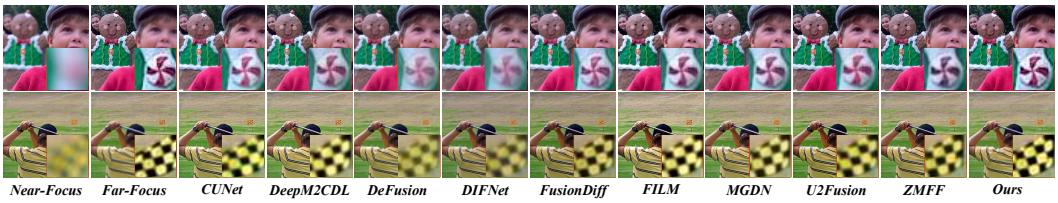


Figure 7: Qualitative results on multi-focus image fusion, showing that the proposed signal-level decomposition paradigm can be naturally transferred to other image fusion tasks.

Medical Image Segmentation for MIF Task. We evaluate medical image segmentation using a pre-trained UniverSeg without fine-tuning[3]. Qualitative results are given in Fig. 5 (right) and quantitative results see Appendix Table 5. Compared with single-modality inputs, fused images provide more informative representations, leading to predictions closer to the ground truth. Our method also outperforms other models, indicating that the proposed strategy helps preserve common anatomical or lesion structures while retaining modality-specific details for boundary delineation.

4.7 Broader Impact

Refining decomposition-based MMIF paradigms. Our signal-level decomposition paradigm provides a clearer objective for common and unique feature separation and can be plugged into existing decomposition-based MMIF models. To validate this, we insert it into DIDFuse, CDDFuse, FD-Fuse [5], and C2RF without changing their default settings. As shown in Fig. 6, the enhanced models preserve more complete structures and richer details. Quantitative results in Appendix Table 6 further confirm its effectiveness in improving existing decomposition-based MMIF methods.

Extension to other image fusion tasks. Our paradigm is not limited to MMIF. We further select multi-focus image fusion as a representative task, where common scene structures and source-specific focused details need to be separated. As shown in Fig. 7, the proposed paradigm can also produce clearer focus boundaries and more faithful structural details on this task. Detailed quantitative results are provided in the Appendix Table 7, suggesting its applicability to broader image fusion problems.

4.8 Model Complexity and Overhead Analysis

We analyze the computational overhead including parameters, GFLOPs, and inference time. Results are in Appendix Table 9. Our model remains computationally competitive. Although the DWT operation introduces a slight additional overhead, the overall complexity is still acceptable.

5 Conclusion

In this paper, we presented a signal-level feature decomposition paradigm that reformulates feature decomposition into an integral-driven optimization at the 1D signal-level. This reformulation yields a matched loss, providing a clear learning objective. Based on this, our two-stage SSL pipeline first performs a decomposition and reconstruction in Stage I, with dual pretext tasks at both levels supervising the entire decomposition and reconstruction process. Stage II then carries out fusion. Extensive experiments show advanced performance, validating the effectiveness of our paradigm.

Acknowledgments

This work was supported by the National Natural Science Foundation of China [No.62401097, 62601484]; Fundamental Research Funds for Central Universities, Dalian Minzu University [No.0854-53]; Liaoning Province Applied Basic Research Program [2026JH2/101300163]; Liaoning Province Science and Technology Joint Plan (2024JH2/102600113).

Competing Interests

The authors declare no competing interests.

References

[1] H. Bai, J. Zhang, Z. Zhao, Y. Wu, L. Deng, Y. Cui, T. Feng, and S. Xu. Task-driven image fusion with learnable fusion loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7457–7468, 2025.

[2] H. Bai, Z. Zhao, J. Zhang, Y. Wu, L. Deng, Y. Cui, B. Jiang, and S. Xu. Refusion: Learning image fusion from reconstruction with learnable loss via meta-learning. *International Journal of Computer Vision*, 133(5):2547–2567, 2025.

[3] V. I. Butoi*, J. J. G. Ortiz*, T. Ma, M. R. Sabuncu, J. Guttag, and A. V. Dalca. Universeg: Universal medical image segmentation. *International Conference on Computer Vision*, 2023.

[4] B. Cao, X. Xu, P. Zhu, Q. Wang, and Q. Hu. Conditional controllable image fusion. *Advances in Neural Information Processing Systems*, 37:120311–120335, 2024.

[5] M. Cheng, H. Huang, X. Liu, H. Mo, S. Wu, and X. Zhao. Fdfuse: Infrared and visible image fusion based on feature decomposition. *IEEE Transactions on Instrumentation and Measurement*, 2025.

[6] X. Deng and P. L. Dragotti. Deep convolutional neural network for multi-modal image restoration and fusion. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2020.

[7] H. Li, D. Su, Q. Cai, and Y. Zhang. Bsafusion: A bidirectional stepwise feature alignment network for unaligned medical image fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 4725–4733, 2025.

[8] H. Li, T. Xu, X.-J. Wu, J. Lu, and J. Kittler. 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, 2023.

[9] W. Li, B. Li, H. Song, P. Wang, and Z. Wang. Infrared and visible image fusion via iterative feature decomposition and deep balanced fusion. *Pattern Recognition*, 174:113022, 2026.

[10] X. Li, S. Du, Y. Zou, H. Xu, Z. Jiang, and J. Liu. Unifusion: A unified image fusion framework with robust representation and source-aware preservation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 33869–33880, 2026.

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

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

[13] X. Li, H. Xu, S. Li, X. Chen, Z. Jiang, and J. Liu. Drfusion: Drift-resilient temporally consistent infrared-visible video fusion. *arXiv preprint arXiv:2605.25775*, 2026.

[14] X. Li, H. Xu, X. Zhu, J. Ma, Y. Zou, Z. Jiang, and J. Liu. Uncertainty-aware spatial-frequency registration and fusion for infrared and visible images. *arXiv preprint arXiv:2605.13049*, 2026.

[15] P. Liang, J. Jiang, X. Liu, and J. Ma. Fusion from decomposition: A self-supervised decomposition approach for image fusion. In *European conference on computer vision*, pages 719–735. Springer, 2022.

[16] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755, 2014.

[17] J. Liu, X. Fan, Z. Huang, G. Wu, R. Liu, W. Zhong, and Z. Luo. 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*, pages 5802–5811, 2022.

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

[19] J. Liu, G. Wu, Z. Liu, D. Wang, Z. Jiang, L. Ma, W. Zhong, X. Fan, and R. Liu. Infrared and visible image fusion: From data compatibility to task adaption. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4):2349–2369, 2025.

[20] J. Liu, B. Zhang, Q. Mei, X. Li, Y. Zou, Z. Jiang, L. Ma, R. Liu, and X. Fan. Dcevo: Discriminative cross-dimensional evolutionary learning for infrared and visible image fusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2226–2235, 2025.

[21] R. Liu, Z. Liu, J. Liu, X. Fan, and Z. Luo. A task-guided, implicitly-searched and meta-initialized deep model for image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(10):6594–6609, 2024.

[22] Y. Liu, Y. Shi, F. Mu, J. Cheng, C. Li, and X. Chen. Multimodal mri volumetric data fusion with convolutional neural networks. *IEEE Transactions on Instrumentation and Measurement*, 71:1–15, 2022.

[23] Z. Liu, J. Liu, G. Wu, L. Ma, X. Fan, and R. Liu. Bi-level dynamic learning for jointly multi-modality image fusion and beyond. *arXiv preprint arXiv:2305.06720*, 2023.

[24] J. Ma, Z. Le, X. Tian, and J. Jiang. Smfuse: Multi-focus image fusion via self-supervised mask-optimization. *IEEE Transactions on Computational Imaging*, 7:309–320, 2021.

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

[26] F. Milletari, N. Navab, and S.-A. Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016.

[27] M. Nechyba. Introduction to the discrete wavelet transform (dwt). *University of Florida, February*, 2004.

[28] G. Qu, D. Zhang, and P. Yan. Information measure for performance of image fusion. *Electronics letters*, 38(7):313–315, 2002.

[29] H. R. Sheikh and A. C. Bovik. Image information and visual quality. *IEEE Transactions on image processing*, 15(2):430–444, 2006.

[30] O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa, F. Massa, D. Haziza, L. Wehrstedt, J. Wang, T. Darcet, T. Moutakanni, L. Sentana, C. Roberts, A. Vedaldi, J. Tolan, J. Brandt, C. Couprie, J. Mairal, H. Jégou, P. Labatut, and P. Bojanowski. DINOv3, 2025.

[31] L. Tang, C. Li, and J. Ma. Mask-difuser: A masked diffusion model for unified unsupervised image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–18, 2025.

[32] L. Tang, Y. Wang, Z. Cai, J. Jiang, and J. Ma. Controlfusion: A controllable image fusion framework with language-vision degradation prompts. *Advances in Neural Information Processing Systems*, 2025.

[33] L. Tang, Q. Yan, X. Xiang, L. Fang, and J. Ma. C2rf: Bridging multi-modal image registration and fusion via commonality mining and contrastive learning. *International Journal of Computer Vision*, 133:5262–5280, 2025.

[34] W. Tang, F. He, Y. Liu, and Y. Duan. Matr: Multimodal medical image fusion via multiscale adaptive transformer. *IEEE Transactions on Image Processing*, 31:5134–5149, 2022.

[35] Y. Tian, Q. Ye, and D. Doermann. Yolov12: Attention-centric real-time object detectors. *arXiv preprint arXiv:2502.12524*, 2025.

[36] A. Toet. The tno multiband image data collection. *Data in Brief*, 15:249–251, 2017.

[37] Q. Wang, Y. Shen, and J. Q. Zhang. A nonlinear correlation measure for multivariable data set. *Physica D: Nonlinear Phenomena*, 200(3-4):287–295, 2005.

[38] Z. Wang, J. Feng, J. Wang, P. Wang, and H. Song. Sigfusion: Unified signal-level self-supervised learning paradigm for image fusion. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(12):10385–10393, Mar. 2026.

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

[40] Z. Wang, J. Zhang, T. Guan, Y. Zhou, X. Li, M. Dong, and J. Liu. Efficient rectified flow for image fusion. *Advances in Neural Information Processing Systems*, 2025.

[41] Z. Wang, J. Zhang, H. Song, M. Ge, J. Wang, and H. Duan. Highlight what you want: Weakly-supervised instance-level controllable infrared-visible image fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12637–12647, 2025.

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

[43] G. Wu, H. Liu, H. Fu, Y. Peng, J. Liu, X. Fan, and R. Liu. Every sam drop counts: Embracing semantic priors for multi-modality image fusion and beyond. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17882–17891, 2025.

[44] X. Wu, Z.-H. Cao, T.-Z. Huang, L.-J. Deng, J. Chanussot, and G. Vivone. Fully-connected transformer for multi-source image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3):2071–2088, 2025.

[45] C. S. Xydeas and V. Petrovic. Objective image fusion performance measure. *Electronics letters*, 36(4):308–309, 2000.

[46] C. Yang, J.-Q. Zhang, X.-R. Wang, and X. Liu. A novel similarity based quality metric for image fusion. *Information Fusion*, 9(2):156–160, 2008.

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

[48] H. Zhang, L. Cao, X. Zuo, Z. Shao, and J. Ma. Omnifuse: Composite degradation-robust image fusion with language-driven semantics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(9):7577–7595, 2025.

[49] Z. Zhao, H. Bai, J. Zhang, Y. Zhang, S. Xu, Z. Lin, R. Timofte, and L. Van Gool. 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*, pages 5906–5916, 2023.

[50] Z. Zhao, H. Bai, J. Zhang, Y. Zhang, K. Zhang, S. Xu, D. Chen, R. Timofte, and L. Van Gool. Equivariant multi-modality image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25912–25921, 2024.

[51] Z. Zhao, S. Xu, C. Zhang, J. Liu, P. Li, and J. Zhang. Didfuse: Deep image decomposition for infrared and visible image fusion. *arXiv preprint arXiv:2003.09210*, 2020.

[52] M. Zhou, J. Huang, K. Yan, D. Hong, X. Jia, J. Chanussot, and C. Li. A general spatial-frequency learning framework for multimodal image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(7):5281–5298, 2025.

[53] P. Zhu, Y. Sun, B. Cao, and Q. Hu. Task-customized mixture of adapters for general image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7099–7108, June 2024.

[54] Y. Zong, O. M. Aodha, and T. M. Hospedales. Self-supervised multimodal learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(7):5299–5318, 2025.

[55] K. H. Zou, S. K. Warfield, A. Bharatha, C. M. Tempany, M. R. Kaus, S. J. Haker, W. M. Wells, F. A. Jolesz, and R. Kikinis. Statistical validation of image segmentation quality based on a spatial overlap index1: scientific reports. *Academic Radiology*, 11(2):178–189, 2004.

A Motivation of the Signal-level Formulation

The core motivation of adopting the signal-level formulation is to redefine the optimization objective for feature decomposition. For decomposition-based MMIF methods, GT feature maps for common and unique features are unavailable. Existing methods usually rely on a combination of image-level metrics, such as intensity and gradient, to indirectly constrain feature decomposition. However, these metrics only describe partial image attributes and are difficult to directly guide the decomposer to explicitly learn common feature and unique features. Therefore, the feature decomposition process still suffers from an unclear supervision target.

We reformulate this supervision-ambiguous feature decomposition problem as a signal-level integral optimization problem. The integral objective describes the global accumulated relation along the signal sequence, making the 1D signal space a natural domain for this optimization form. In this domain, the common signal can be constrained by the integral area between itself and the two source signals, which provides the decomposer with an explicit, computable, and decomposition-oriented learning target. In this way, the signalized representation becomes a key form to support the integral-driven objective, giving the decomposer a clearer optimization objective.

Please note that, the signal-level formulation does not discard spatial information. The 1D signal is mainly used to construct a clearer decomposition objective, while the decomposed signals are mapped back to 2D feature maps and further constrained by the image-level reconstruction task. Therefore, the signal-level and image-level tasks are complementary: the former provides an explicit objective for common and unique feature separation, while the latter preserves spatial structural consistency during reconstruction.

B Ambiguous Supervision of Pixel-Level Metric-Mixture Paradigm.

In current decomposition-based MMIF methods, ground-truth common and unique feature maps are unavailable. As a result, existing methods usually constrain the decomposition process by imposing losses between the learned representation and the source images. However, each image pixel does not correspond to a single visual property. Instead, it simultaneously couples multiple attributes, such as intensity, contrast, edge, texture, contour, structure, and salient modality-dependent responses. To make this issue more explicit, we summarize representative visual attributes coupled in image pixels and their commonly used constraints in Table 3.

Table 3: Representative visual attributes coupled in image pixels and their commonly used constraints in MMIF.

Visual Attribute	Pixel-level Manifestation	Commonly Used Constraints
Intensity / Brightness	Absolute pixel magnitude, local luminance	L_1 loss, L_2 loss, reconstruction loss
Contrast	Relative intensity difference between a pixel and its local neighborhood, or between foreground and background	SSIM contrast term, local contrast loss, entropy-based or information-based constraints
Edge	Sharp local intensity variation around object boundaries and structural transitions	Gradient loss, Sobel loss, Laplacian loss
Texture	fine details, and high-frequency responses	Gradient loss, frequency-domain loss
Boundary	Continuous object outlines and region boundaries	SSIM loss, gradient loss, boundary-aware loss
Structural Layout	Spatial arrangement of objects, organs, roads, targets, or anatomical regions	SSIM loss, perceptual loss
Color	Channel-wise appearance, pseudo-color distribution, and modality-related color responses	Color consistency loss, histogram loss

As shown in Table 3, a single pixel simultaneously reflects multiple visual attributes, whereas each loss term only constrains part of them from a specific perspective. Obviously, it is almost impossible to embed all these attribute-related constraints into a single loss formulation. More importantly, there is no prior knowledge about which combination of these metrics is optimal for feature decomposition. Different losses may even impose potentially conflicting optimization preferences. Therefore, existing supervision based on empirical metric mixtures is inherently incomplete and ambiguous. This is exactly why we reformulate feature decomposition into a signal-level integral optimization problem, which provides the decomposer with a clearer objective.

Algorithm 1 Stage I: Self-supervised signal-level feature decomposition with dual pretext tasks

Input: Training set $\mathcal{D} = \{(I_A^k, I_B^k)\}_{k=1}^N$
Modules: Encoder ζ, signal-level decomposer SD, decoder ξ
Hyper-parameters: α, β for Stage I.

Initialize parameters of ζ, SD, ξ
for epoch $= 1$ to E_1 **do**
  for each mini-batch $(I_A, I_B) \subset \mathcal{D}$ **do**
    $F_A, F_B = \zeta(I_A), \zeta(I_B)$
    $s_A(l), s_B(l) = H(F_A, F_B)$
    $L_A(l), \{H_A^i(l)\}_{i=1}^N = \mathrm{DWT}(s_A(l))$
    $L_B(l), \{H_B^i(l)\}_{i=1}^N = \mathrm{DWT}(s_B(l))$
    // Signal-level decomposer: common & unique signals
    $c_s(l) = SD_{\text{common}}(L_A(l), L_B(l))$
    $u_s^A(l), u_s^B(l) = SD_{\text{specific}}(\{H_A^i(l)\}, \{H_B^i(l)\})$
    // Signal-level pretext: common-signal optimization with L_{dec}
    $L_{\mathrm{dec}} = \text{IntegralSquaredError}(c_s(l), s_A(l), s_B(l))$
    // Image-level pretext: reconstruct each source image
    $\hat{s}_A(l) = \mathrm{DWT}^{-1}(c_s(l), u_s^A(l))$
    $\hat{s}_B(l) = \mathrm{DWT}^{-1}(c_s(l), u_s^B(l))$
    $\hat{F}_A = H^\top(\hat{s}_A(l)), \quad \hat{F}_B = H^\top(\hat{s}_B(l))$
    $\hat{I}_A = \xi(\hat{F}_A), \quad \hat{I}_B = \xi(\hat{F}_B)$
    $L_{\mathrm{rec}} = \mathrm{MSE}(\hat{I}_A, I_A) + \mathrm{MSE}(\hat{I}_B, I_B)$
    // Dual-pretext loss for decomposer
    $L_{\text{stage1}} = \alpha L_{\mathrm{dec}} + \beta L_{\mathrm{rec}}$
    Update parameters of ζ, SD, ξ by back-propagating L_{stage1}
  end for
end for
Output: Pretrained encoder ζ and signal-level decomposer SD

C Two-Stage Training and Inference Pipeline

For clarity, we summarize the overall pipeline of the proposed method in a two-stage manner, as shown in Algorithm 1 and Algorithm 2.

C.1 Stage I: Self-supervised signal-level decomposition.

In the first stage (Algorithm 1), we focus on learning a stable signal-level decomposer in a fully self-supervised fashion. Given paired source images (I_A, I_B), we first extract high-level features via the encoder ζ, and then project the 2D feature maps into 1D signals. These 1D signals are decomposed by a 1D DWT into low- and high-frequency components, which are further processed by the signal-level decomposer SD to obtain a common signal $c_s(l)$ and modality-specific unique signals $u_s^A(l)$ and $u_s^B(l)$. To train SD, we introduce a dual-pretext objective: a signal-level loss L_{dec} that directly supervises the decomposition in the 1D domain, and an image-level reconstruction loss L_{rec} that reconstructs each source image via inverse DWT and the decoder ξ. The decomposer and encoder are optimized jointly with the combined loss $L_{\text{stage1}} = \alpha L_{\mathrm{dec}} + \beta L_{\mathrm{rec}}$, yielding a pretrained encoder–decomposer pair that can reliably separate common and modality-specific information.

C.2 Stage II: Fusion based on signal-level decomposition.

In the second stage (Algorithm 2), we freeze the encoder ζ and decomposer SD learned in Stage I, and only train the fusion head δ together with the decoder ξ. For each training pair, we reuse the signal-level decomposition to obtain the common signal $c_s(l)$ and unique signals $u_sA(l)$ and $u_sB(l)$, and then feed the unique signals into the learnable fusion module δ to produce a fused unique signal $u_f(l)$. Combining $c_s(l)$ and $u_f(l)$ through inverse DWT gives a fused 1D signal, which is mapped back to 2D feature space and decoded into the fused image I_f. The fusion-specific loss $\mathcal{L}$ is computed between I_f and the source images, encouraging the fused output to preserve structural

Algorithm 2 Stage II: Fusion based on signal-level decomposition

Require: Training set $\mathcal{D} = \{(I_A^k, I_B^k)\}_{k=1}^N$ // *registered multimodal image pairs*
Require: Pretrained encoder ζ and signal-level decomposer SD from Stage I // *fixed feature extractor and decomposer*
Ensure: Fused image generator $G(I_A, I_B)$ // *for both MIF and VIF tasks*
Freeze parameters of ζ and SD // *use Stage I as a fixed decomposer*
Initialize fusion module δ and decoder ξ // *learnable fusion head and image reconstructor*
for each training epoch **do**
  for each mini-batch $(I_A, I_B) \subset \mathcal{D}$ **do**
    $F_A, F_B = \zeta(I_A), \zeta(I_B)$
    $s_A(l), s_B(l) = H(F_A, F_B)$
    $L_A(l), \{H_A^i(l)\} = \mathrm{DWT}(s_A(l))$
    $L_B(l), \{H_B^i(l)\} = \mathrm{DWT}(s_B(l))$
    $c_s(l) = SD_{\mathrm{common}}(L_A(l), L_B(l))$
    $u_s^A(l), u_s^B(l) = SD_{\mathrm{specific}}(\{H_A^i(l)\}, \{H_B^i(l)\})$ // *modality-specific unique signals*
    $u_f(l) = \delta(u_s^A(l), u_s^B(l))$ // *fuse unique signals by the learnable fusion head*
    $f_s(l) = \mathrm{DWT}^{-1}(c_s(l), u_f(l))$ // *reconstruct fused 1D signal with inverse DWT*
    $F_f = H^\top(f_s(l))$
    $I_f = \xi(F_f)$ // *decode fused feature into fused image*
    $L_{\mathrm{fuse}} = \mathcal{L}(I_f; I_A, I_B)$
    Update parameters of δ and ξ using gradient of L_{fuse} // *optimize fusion head and decoder*
  end for
end for
$I_f = G(I_A, I_B)$ for a given test pair (I_A, I_B) // *inference: one forward pass through the Stage II pipeline*

details, intensity information, and gradient cues from both modalities. After training, inference reduces to a single forward pass through this Stage II pipeline, i.e., $I_f = G(I_A, I_B)$, for both MIF and VIF tasks.

D Extended Experimental Results and Visualizations

We adopt the same evaluation as in the main paper. Following mainstream practice, we use a set of widely used image quality metrics, including Q_{MI} [28], Q_{NICE} [37], Q_P [45], Q_{CB} [41], MI [28], VIF_P [29], and Q_Y [46]. For fair comparison, we also keep the same state-of-the-art baselines as in the main paper: for the VIF task we consider CDDFuse [49], LRRNet [8], EMMA [50], TC-MoA [53], Text-DiFuse [47], SigFusion [38], DCEvo [20], SAGE [43], TDFusion [1], OmniFuse [48], and C2RF [33]; for the MIF task we adopt CDDFuse [49], LRRNet [8], EMMA [50], TC-MoA [53], Text-DiFuse [47], CCF [4], SigFusion [38], BSA-Fusion [7], Mask-Difuser [31], MTG-Fusion [42], and C2RF [33]. Due to page constraints in the main paper, only a subset of results can be shown there; this section reports extended quantitative results and richer visualizations across all datasets and modalities to more comprehensively demonstrate the behavior and advantages of our method.

D.1 Full Qualitative Comparisons

We provide visual comparisons against all competing methods on six datasets. For each dataset, we show the source images, the fused results generated by representative competitors, and the output of our method. The extended visualizations are presented in Fig. 8, where the first three groups correspond to MIF cases (MRI-CT, MRI-PET, MRI-SPECT) and the last three groups correspond to VIF cases (TNO, M^3FD, MSRS). Within each row, the first two columns display the source image pairs, while the remaining columns show the fused results of different comparison methods, with our fusion result placed in the rightmost column.

Benefiting from the proposed 1D signal-level decomposer and the integral-driven loss L_{dec}, our method yields fused images that retain both the global structures and modality-specific details. On the MIF datasets, our results preserve sharp boundaries, bone regions and texture from MRI and CT, while maintaining the functional patterns of PET and SPECT without color distortion. On the VIF datasets,

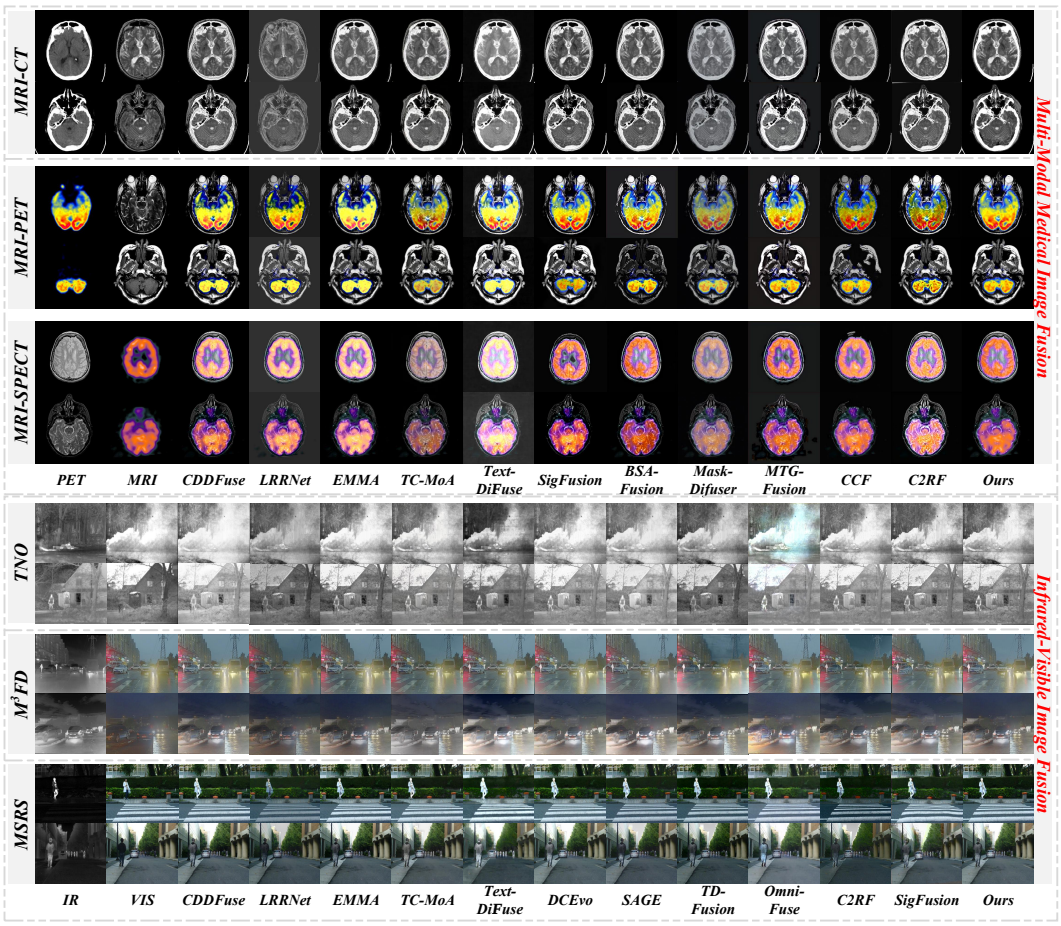


Figure 8: Full qualitative comparisons on six datasets. The top three groups show MIF task and the bottom three groups show VIF cases. In each row, the first two columns are the source image pairs and the remaining columns are fused results of different methods, with our method in the rightmost column, yielding sharper structures and richer complementary details.

our fused images simultaneously preserve the fine textures and contrast of the visible modality and the salient thermal targets from infrared, leading to clearer edges, more stable illumination than competing methods. These visual results are consistent with our signal-level loss which provides a clearer optimization objective with the dual pretext tasks, jointly enabling more accurate feature decomposition and, consequently, more advanced fusion results across all six datasets.

D.2 Downstream Task Evaluation

Object Detection for VIF Task. To evaluate whether the fused images can better support downstream visual understanding, we conduct object detection on the VIF task. Specifically, fused images generated by different VIF methods are fed into the same pre-trained YOLOv12 detector [35], without modifying the detector or fine-tuning it on any specific fusion result. This setting ensures that the comparison mainly reflects the influence of different fusion outputs on detection performance.

As shown in Fig. 5(left), the detection results produced from our fused images are generally more complete and accurate than those from other fusion methods. In challenging scenes, competing methods may miss small targets, generate inaccurate bounding boxes, or confuse objects with weak contrast. In contrast, our method better preserves complementary infrared and visible cues, which helps the detector localize objects more reliably. Table 4 reports the per-class AP [16] and mAP results. Compared with other VIF methods, our method achieves the best overall mAP and competitive

Table 4: AP for object detection on source images and fused images from different methods. mAP is the mean of the six category APs.

Method	People	Car	Bus	Motorcycle	Lamp	Truck	mAP
IR	0.512	0.581	0.447	0.468	0.403	0.452	0.477
VIS	0.563	0.638	0.483	0.521	0.436	0.494	0.522
CDDFuse	0.662	0.726	0.618	0.634	0.573	0.602	0.636
LRRNet	0.641	0.707	0.597	0.612	0.551	0.582	0.615
EMMA	0.651	0.719	0.612	0.631	0.562	0.593	0.628
TC-MoA	0.642	0.715	0.606	0.623	0.559	0.601	0.624
Text-DiFuse	0.598	0.689	0.542	0.588	0.521	0.552	0.582
C2RF	0.611	0.701	0.553	0.599	0.515	0.561	0.590
SigFusion	0.688	**0.763**	0.653	0.649	<u>0.579</u>	0.622	0.659
DCEvo	<u>0.689</u>	0.735	**0.661**	0.653	**0.583**	0.612	0.655
SAGE	0.629	0.703	0.571	0.609	0.542	0.571	0.604
TD-Fusion	0.661	0.731	0.632	0.643	**0.583**	0.611	0.643
OmniFuse	0.649	0.721	0.601	0.621	0.571	0.603	0.628
Ours	**0.691**	<u>0.737</u>	<u>0.658</u>	**0.651**	0.577	**0.623**	<u>0.656</u>

Table 5: Quantitative comparison for medical image segmentation.

Metrics	T1	Flair	CDDFuse	LRRNet	EMMA	TC-MoA	Text-DiFuse
IoU	0.615	0.657	0.762	0.731	0.743	0.689	0.685
Dice	0.627	0.727	<u>0.850</u>	0.836	0.841	0.829	0.751
Metrics	CCF	SigFusion	BSAFusion	Mask-Difuser	MTG-Fusion	C2RF	Ours
IoU	0.661	0.735	0.745	0.723	<u>0.775</u>	0.772	**0.790**
Dice	0.752	0.770	0.832	0.829	0.846	0.848	**0.856**

AP on most object categories, demonstrating that the proposed fusion strategy can provide more task-favorable inputs for downstream detection.

Medical Image Segmentation for MIF Task. We further evaluate the downstream utility of our method on medical image segmentation. In this experiment, the fused images generated by different MIF methods are used as inputs to UniverSeg [3], and the segmentation results are compared with the ground-truth masks. This evaluation examines whether the fused images can preserve lesion-related structures and boundary information that are useful for medical analysis.

As shown in Fig. 5(right), segmentation results based on our fused images are closer to the ground truth, especially around lesion boundaries and structurally complex regions. Other fusion methods may produce incomplete lesion regions or less accurate boundaries due to insufficient preservation of cross-modal complementary information. In contrast, our method better maintains common anatomical structures while retaining modality-specific details, which provides more discriminative cues for segmentation. Table 5 reports the IoU [55] and Dice [26] results. Our method obtains the best performance among the compared fusion methods, confirming that the proposed decomposition strategy can improve not only visual fusion quality but also downstream medical segmentation performance.

D.3 Visualization of Common and Unique Features

To better understand our feature decomposition mechanism, we compare the estimated common and unique features of our method with those of several representative decomposition-based baselines. For each source pair (I_A, I_B), we visualize the common feature maps and the corresponding modality-unique feature maps.

As shown in Fig. 9, existing image-level decomposers typically produce two common features. In practice, these two features are not strictly shared as they often contain modality-specific patterns, which contradicts the assumption that there should be a single common component underlying both modalities. In contrast, our signal-level decomposer explicitly models a single common feature

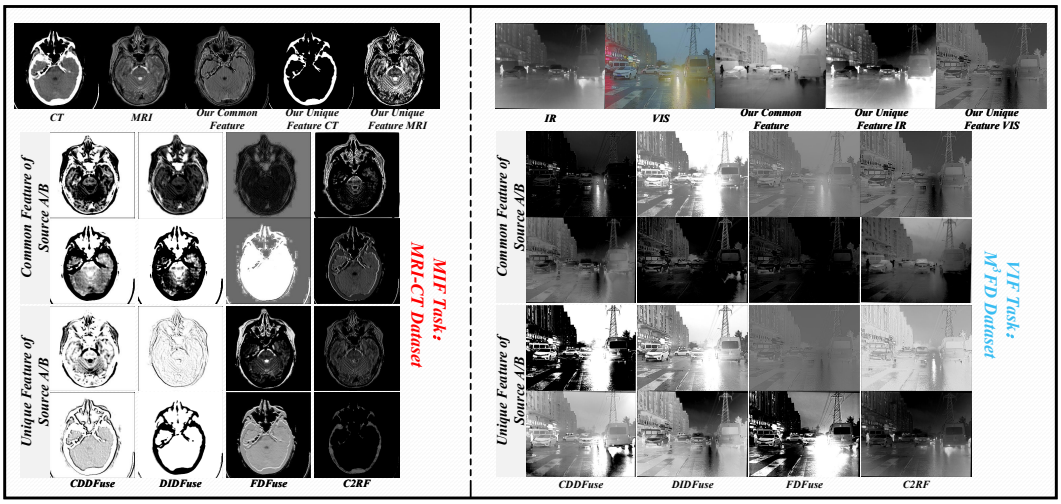


Figure 9: Visualization of common and unique feature maps produced by different decomposition strategies. For each source pair, we display the estimated common features and the modality-unique features. Compared with existing methods that generate two inconsistent common features, our signal-level decomposer yields a single common map and better-separated unique maps.

map C_{AB} together with two unique feature maps for I_A and I_B, leading to a more reasonable and identifiable decomposition form.

The visualizations further reveal that, for our method, the common feature mainly captures shared semantic content and global structures, while the unique feature concentrate on modality-specific cues. These observations align with our signal-level formulation and support that the proposed decomposer separates common and modality-specific information in a more structured and disentangled way.

D.4 t-SNE Visualization of Frequency and Decomposed Features

To further analyze the relationship between frequency components and the learned decomposed features, we visualize their feature distributions using t-SNE. Specifically, we project the low-frequency components, high-frequency components, learned common features, and modality-specific unique features into a two-dimensional space. As shown in Fig. 10, the learned common features are mainly distributed around the low-frequency components of both modalities, indicating that low-frequency signals contain more modality-shared structural information. In contrast, the learned unique features are closer to the corresponding high-frequency components and form more modality-specific distributions. This observation is consistent with our design motivation that common information can be better extracted from low-frequency components, while modality-specific details are more related to high-frequency components.

These results provide intuitive evidence for the proposed decomposition strategy. By exploiting the complementary roles of low- and high-frequency components, our method can more effectively separate common structures and modality-specific details.

D.5 Visualization of Training Curves

To analyze the optimization behavior of the proposed signal-level formulation, we compare it with a conventional 2D image-domain mixed-loss strategy. As shown in Fig. 11, our signal-level objective converges faster and reaches a lower final loss, indicating a clearer and more stable optimization target.

We further visualize the similarity curves during training, including SSIM, Pearson correlation, and cosine similarity. Compared with the 2D mixed-loss strategy, our method shows higher and more stable similarity between the learned common features and source signals. These results provide intuitive evidence that the proposed signal-level objective can better guide common feature learning.

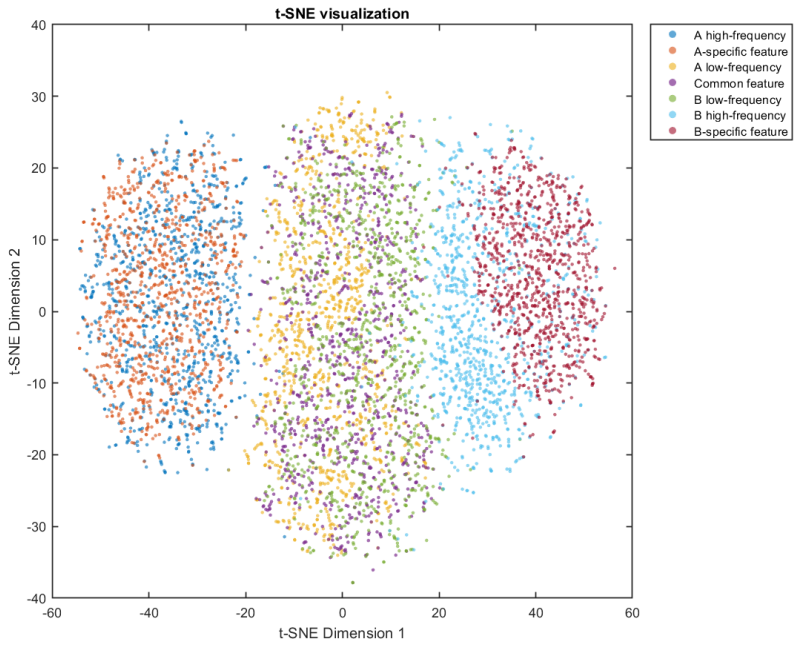
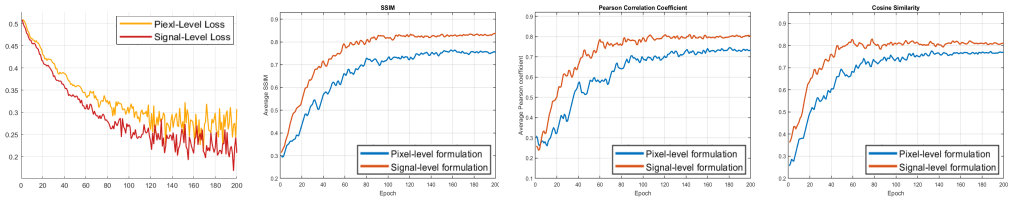


Figure 10: t-SNE visualization of low-/high-frequency components and learned decomposed features. The common features are close to the low-frequency components of both modalities, while the unique features are closer to their corresponding high-frequency components, supporting the design of extracting common information from low frequencies and unique information from high frequencies.



(a) Loss Curve *(b) SSIM Curve* *(c) Pearson Curve* *(d) Cosine Curve*

Figure 11: Visualization of training curves. We compare the proposed signal-level objective with a conventional 2D image-domain mixed-loss strategy in terms of training loss and feature similarity, including SSIM, Pearson correlation, and cosine similarity. The results show that our signal-level objective converges faster and achieves a lower final loss, while maintaining higher and more stable similarity between the learned common features and source signals.

E Additional Results for Broader Impact

In this section, we provide the detailed quantitative results corresponding to the broader impact analysis in the main paper. We mainly consider two aspects: refining existing decomposition-based MMIF paradigms and extending the proposed signal-level decomposition paradigm to other image fusion tasks.

E.1 Refining Decomposition-based MMIF Paradigms

To further validate the compatibility of the proposed signal-level decomposition paradigm, we insert it into several representative decomposition-based MMIF models. For a fair comparison, we keep the original training strategies, network settings, and evaluation protocols of these models unchanged, and only replace or enhance their feature decomposition process with our signal-level decomposition paradigm. The enhanced variants are denoted by "*".

Table 6 reports the quantitative results. Compared with the original models, the enhanced variants achieve consistent improvements on most evaluation metrics. These results indicate that the proposed

Table 6: The refinement effects results of our paradigm. Original vs. upgraded models (* denotes our paradigm inserted).

Method	$Q_{MI}\uparrow$	$Q_{NICE}\uparrow$	$Q_P\uparrow$	$Q_{CB}\uparrow$	$MI\uparrow$	$VIF_p\uparrow$	$Q_Y\uparrow$
DIDFuse	0.6130	0.8083	0.3570	0.4830	3.2370	0.3610	0.9020
DIDFuse*	0.7466	0.8196	0.4198	0.5299	3.9038	0.4596	0.9101
Gain	21.8%↑	1.4%↑	17.6%↑	9.7%↑	20.6%↑	27.3%↑	0.9%↑
CDDFuse	0.6610	0.8107	0.4010	0.6940	3.9820	0.3990	0.9380
CDDFuse*	0.7582	0.8277	0.4523	0.7051	4.3563	0.4736	0.9446
Gain	14.7%↑	2.1%↑	12.8%↑	1.6%↑	9.4%↑	18.7%↑	0.7%↑
FDFuse	0.6420	0.8086	0.3890	0.6860	3.9050	0.3920	0.9330
FDFuse*	0.8044	0.8215	0.4820	0.7320	4.5532	0.5088	0.9433
Gain	25.3%↑	1.6%↑	23.9%↑	6.7%↑	16.6%↑	29.8%↑	1.1%↑
C2RF	0.6360	0.8071	0.3660	0.6740	3.8510	0.3730	0.9290
C2RF*	0.8243	0.8224	0.4663	0.7535	4.8022	0.4566	0.9383
Gain	29.6%↑	1.9%↑	27.4%↑	11.8%↑	24.7%↑	22.4%↑	1.0%↑

Table 7: Quantitative comparison on the MFIF task against 9 competing methods using 6 evaluation metrics on the LYTRO and MFFW datasets. Although MFIF may provide fused ground truth for evaluation, the decomposition of common and unique features still lacks direct supervision. Our method remains applicable in this setting and achieves strong overall performance.

Method	Pub/Year	Dataset: LYTRO						Dataset: MFFW					
		$Q_G\uparrow$	$Q_M\uparrow$	$Q_P\uparrow$	$MI\uparrow$	$SD\uparrow$	$VIFF\uparrow$	$Q_G\uparrow$	$Q_M\uparrow$	$Q_P\uparrow$	$MI\uparrow$	$SD\uparrow$	$VIFF\uparrow$
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
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
DeepM2CDL	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
Ours	-	0.659	**1.031**	**0.811**	6.733	58.66	**1.292**	**0.611**	**0.994**	**0.682**	**6.002**	56.18	**1.112**

paradigm is not limited to our specific network architecture. Instead, by providing a clearer objective for separating common and unique features, it can serve as a generally effective refinement for existing decomposition-based MMIF methods.

E.2 Extension to Multi-focus Image Fusion

We further examine whether the proposed signal-level decomposition paradigm can be transferred to other image fusion tasks beyond MMIF. In this section, we select multi-focus image fusion as a representative task. Different from MMIF, multi-focus image fusion aims to integrate multiple images of the same scene with different focus regions. Although the imaging setting is different, this task also requires the model to preserve common scene structures while extracting source-specific focused details. Therefore, multi-focus image fusion provides a suitable testbed for evaluating the broader applicability of our decomposition paradigm.

Table 7 reports the quantitative results. The proposed paradigm achieves competitive or superior performance across multiple metrics. These results are consistent with the qualitative observations in the main paper, where our method produces clearer focus boundaries and more faithful structural details. This further demonstrates that the signal-level decomposition paradigm can be naturally extended to other image fusion tasks and has potential value for more general fusion problems.

F Extended Ablation Study

To comprehensively verify the effectiveness of each key component, we conduct an extended ablation study on all six datasets for both VIF and MIF tasks. The quantitative results of all variants are summarized in Table 8, and representative qualitative comparisons for typical cases are visualized in Fig. 12, complementing the main-paper results.

Table 8: Quantitative results of multiple ablation experiments for VIF and MIF tasks on six datasets.

Configuration settings	VIF task $Q_{MI}\uparrow$	$Q_{NICE}\uparrow$	$Q_P\uparrow$	$Q_{CB}\uparrow$	$MI\uparrow$	$VIF_p\uparrow$	$Q_Y\uparrow$	MIF task $Q_{MI}\uparrow$	$Q_{NICE}\uparrow$	$Q_P\uparrow$	$Q_{CB}\uparrow$	$MI\uparrow$	$VIF_p\uparrow$	$Q_Y\uparrow$
I. Effect of Signal-Level Decomposer														
	VIF: M^3FD							*MIF: MRI-CT*						
w/o signal-level decomposer	0.6282	0.8181	0.5350	0.4716	3.8631	0.4391	0.8729	0.9012	0.8071	0.3627	0.6715	3.6129	0.3689	0.9127
w/o DWT	0.6425	0.8192	0.5551	0.4860	3.9812	0.4534	0.8847	0.9185	0.8090	0.3825	0.6880	3.8310	0.3897	0.9312
w/ IDWT ← learnable reconstructor	0.6531	0.8198	0.5631	0.4925	4.0200	0.4594	0.8911	0.9253	0.8096	0.3928	0.6905	3.9025	0.3954	0.9361
w/ signal-level ← image-level decomposer	0.6380	0.8185	0.5460	0.4793	3.9022	0.4468	0.8781	0.9108	0.8086	0.3755	0.6810	3.7420	0.3798	0.9275
Default setups (ours)	**0.6608**	**0.8209**	**0.5735**	**0.4998**	**4.1050**	**0.4694**	**0.8988**	**0.9369**	**0.8108**	**0.4034**	**0.7008**	**4.0184**	**0.4034**	**0.9439**
	VIF: MSRS							*MIF: MRI-PET*						
w/o signal-level decomposer	0.8154	0.8230	0.5112	0.4667	4.8120	0.4932	0.8791	0.8022	0.8053	0.4613	0.4442	2.8931	0.4351	0.5620
w/o DWT	0.8260	0.8268	0.5304	0.5551	5.0921	0.5085	0.8897	0.8191	0.8061	0.4780	0.4557	2.9874	0.4467	0.5734
w/ IDWT ← learnable reconstructor	0.8335	0.8279	0.5411	0.5645	5.1982	0.5166	0.8935	0.8318	0.8070	0.4876	0.4621	3.0315	0.4520	0.5826
w/ signal-level ← image-level decomposer	0.8213	0.8250	0.5283	0.5512	5.0455	0.5033	0.8860	0.8244	0.8066	0.4762	0.4520	2.9634	0.4447	0.5692
Default setups (ours)	**0.8405**	**0.8290**	**0.5483**	**0.5798**	**5.4547**	**0.5253**	**0.8936**	**0.8843**	**0.8074**	**0.4967**	**0.4740**	**3.0547**	**0.4577**	**0.5808**
	VIF: TNO							*MIF: MRI-SPECT*						
w/o signal-level decomposer	0.5410	0.8105	0.4023	0.5115	3.7412	0.4544	0.8322	0.9621	0.8072	0.5260	0.6521	2.9435	0.4410	0.8998
w/o DWT	0.5567	0.8116	0.4195	0.5237	3.7952	0.4658	0.8390	0.9699	0.8079	0.5361	0.6615	2.9857	0.4478	0.9047
w/ IDWT ← learnable reconstructor	0.5630	0.8121	0.4264	0.5298	3.8434	0.4722	0.8424	0.9755	0.8080	0.5445	0.6659	3.0213	0.4511	0.9101
w/ signal-level ← image-level decomposer	0.5489	0.8113	0.4152	0.5189	3.7680	0.4607	0.8361	0.9682	0.8075	0.5324	0.6582	2.9880	0.4482	0.9036
Default setups (ours)	**0.5711**	**0.8128**	**0.4318**	**0.5346**	**3.8015**	**0.4768**	**0.8450**	**0.9778**	**0.8083**	**0.5507**	**0.6697**	**3.0655**	**0.4520**	**0.9149**
II. Effect of Dual Pretext Tasks														
	VIF: M^3FD							*MIF: MRI-CT*						
w/o signal-level pretext task	0.6315	0.8188	0.5420	0.4868	3.9033	0.4512	0.8799	0.9144	0.8093	0.3786	0.6869	3.7562	0.3831	0.9293
w/o image-level pretext task	0.6495	0.8195	0.5525	0.4820	3.9800	0.4553	0.8850	0.9228	**0.8294**	0.3866	0.6765	3.8041	0.3912	0.9335
w/ $L_{\text{dec}} \leftarrow L_1$	0.6463	0.8195	0.5482	0.4817	3.9521	0.4520	0.8832	0.9202	0.8091	0.3814	0.6851	3.7820	0.3886	0.9318
Default setups (ours)	**0.6608**	**0.8209**	**0.5735**	**0.4998**	**4.1050**	**0.4694**	**0.8988**	**0.9369**	0.8108	**0.4034**	**0.7008**	**4.0184**	**0.4034**	**0.9439**
	VIF: MSRS							*MIF: MRI-PET*						
w/o signal-level pretext task	0.8128	0.8220	0.5287	0.5652	5.1053	0.5188	0.8865	0.8230	0.8062	**0.5025**	0.4604	2.9634	0.4461	0.5742
w/o image-level pretext task	**0.8467**	0.8265	0.5361	0.5661	5.1980	0.5134	0.8922	0.8322	0.8070	0.4860	0.4583	3.0258	0.4523	**0.5831**
w/ $L_{\text{dec}} \leftarrow L_1$	0.8223	0.8241	0.5338	0.5594	5.1647	0.5110	0.8895	0.8285	0.8067	0.4822	0.4562	3.0043	0.4495	0.5796
Default setups (ours)	0.8405	**0.8290**	**0.5483**	**0.5798**	**5.4547**	**0.5253**	**0.8936**	**0.8843**	**0.8074**	0.4967	**0.4740**	**3.0547**	**0.4577**	0.5808
	VIF: TNO							*MIF: MRI-SPECT*						
w/o signal-level pretext task	0.5468	0.8105	0.4140	0.5235	3.7845	0.4620	0.8367	0.9759	0.8075	0.5422	0.6631	3.0176	0.4505	0.9085
w/o image-level pretext task	0.5590	0.8118	0.4256	0.5212	3.8210	0.4701	0.8412	0.9758	0.8080	0.5480	0.6610	3.0435	**0.4526**	0.9120
w/ $L_{\text{dec}} \leftarrow L_1$	**0.5727**	0.8112	0.4203	0.5194	3.7958	0.4672	0.8389	0.9747	0.8078	0.5451	0.6598	3.0299	0.4516	0.9103
Default setups (ours)	0.5711	**0.8128**	**0.4318**	**0.5346**	**3.8015**	**0.4768**	**0.8450**	**0.9778**	**0.8083**	**0.5507**	**0.6697**	**3.0655**	0.4520	**0.9149**

Effect of Signal-Level Decomposer. Our signal-level decomposer transfers the decomposition problem into 1D signal-level, effectively disentangling image-level coupling and enabling more accurate separation of features. To assess its impact, we conduct the following ablations: (1) removing the signal-level decomposer; (2) removing frequency decomposition; (3) replacing IDWT with a learnable reconstructor; (4) substituting an image-level decomposer; (5) default setup. The corresponding quantitative results are given in the upper part of Table 8. Across all six datasets and seven metrics, the default setup consistently achieves the advanced performance, while removing or weakening any part of the signal-level decomposer leads to clear drops, which further demonstrates the effectiveness of our paradigm. In Fig. 12 (left), we visualize typical VIF and MIF examples under different configurations: the variants without a complete signal-level decomposer tend to produce blurred structures, noisy background or loss of weak details, whereas the default setup preserves sharper organ boundaries, clearer edges and richer complementary cues from both modalities.

Effect of Dual Pretext Tasks. To evaluate the dual pretext tasks at both levels, we perform the following ablations: (1) removing the signal-level pretext task; (2) removing the image-level pretext task; (3) replacing the proposed L_{dec} with vanilla L_1 only; (4) default configuration. The results are reported in the lower part of Table 8. Removing either pretext task consistently degrades performance on all datasets, and using a naive L_1 loss also weakens the results, highlighting the necessity to provide the decomposer a clear and properly designed supervision. These tendencies are also clearly reflected in Fig. 12 (right), where the default configuration yields the most faithful results.

Sensitivity to Key Hyperparameters. We further analyze the sensitivity of key hyperparameters and design choices. The corresponding curves and additional ablation plots are provided in Fig. 14 of the Appendix.

G Reproducibility Verification

We further conduct robustness checks with respect to random initialization. Keeping all architectural settings unchanged, we rerun our model with multiple random seeds. As shown in Fig. 13, the results under different seeds are tightly clustered, and the quantitative metrics exhibit small standard deviations, indicating that our method is stable.

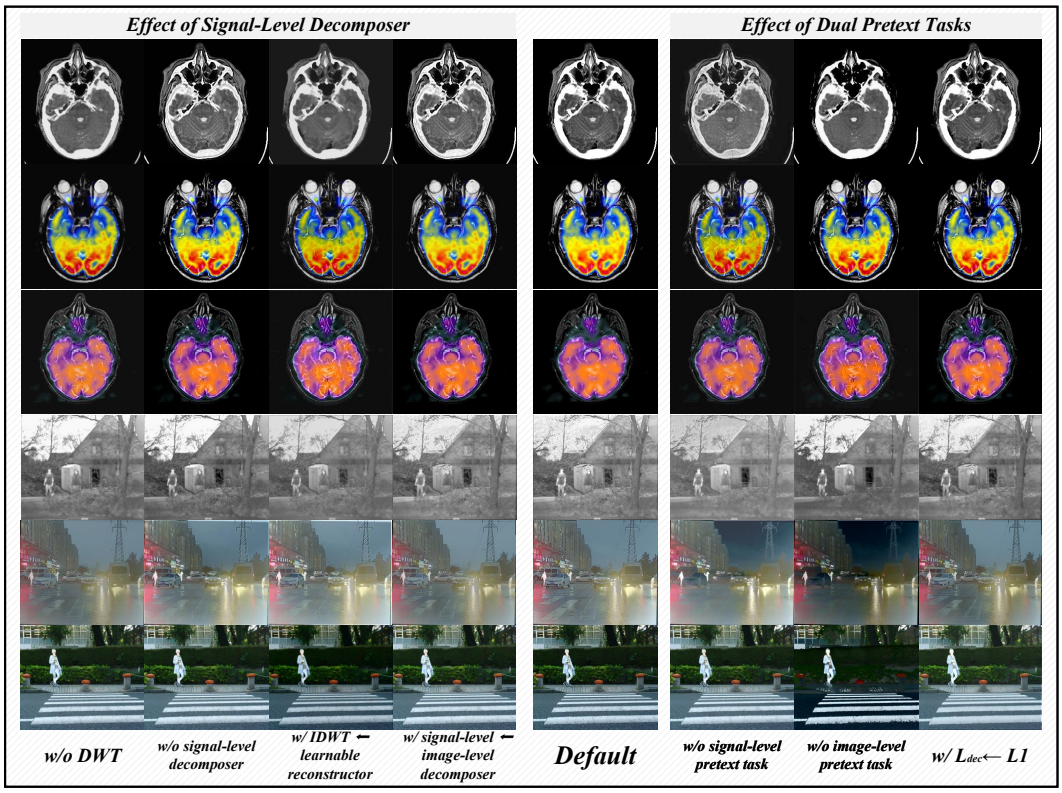


Figure 12: Visualization of ablation results for different configurations on representative VIF and MIF cases.

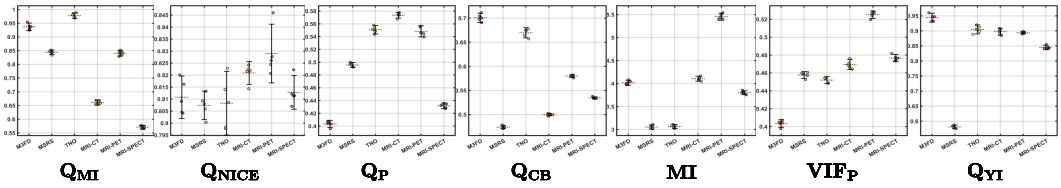


Figure 13: Reproducibility verification under different random seeds. For each metric, markers denote independent runs with different initializations, and their tight clustering with small standard deviations indicates that our method is robust to random initialization.

H Model Complexity and Overhead Analysis

We analyze the computational overhead, including the number of parameters, GFLOPs, and inference time as shown in Table 9. Compared with representative baselines, our model remains computationally competitive. Although the use of DWT introduces some additional overhead, the resulting performance gains make this cost well justified.

I Limitations

While our algorithm demonstrates promising performance, it still has several limitations. First, our method currently relies on pre-registered image pairs to achieve optimal fusion results. This requirement, which is also shared by many mainstream image fusion models, may restrict its applicability in scenarios where accurate pre-registration is challenging or expensive. Second, since our method operates in the signal domain, larger images, when flattened, produce longer 1D signals

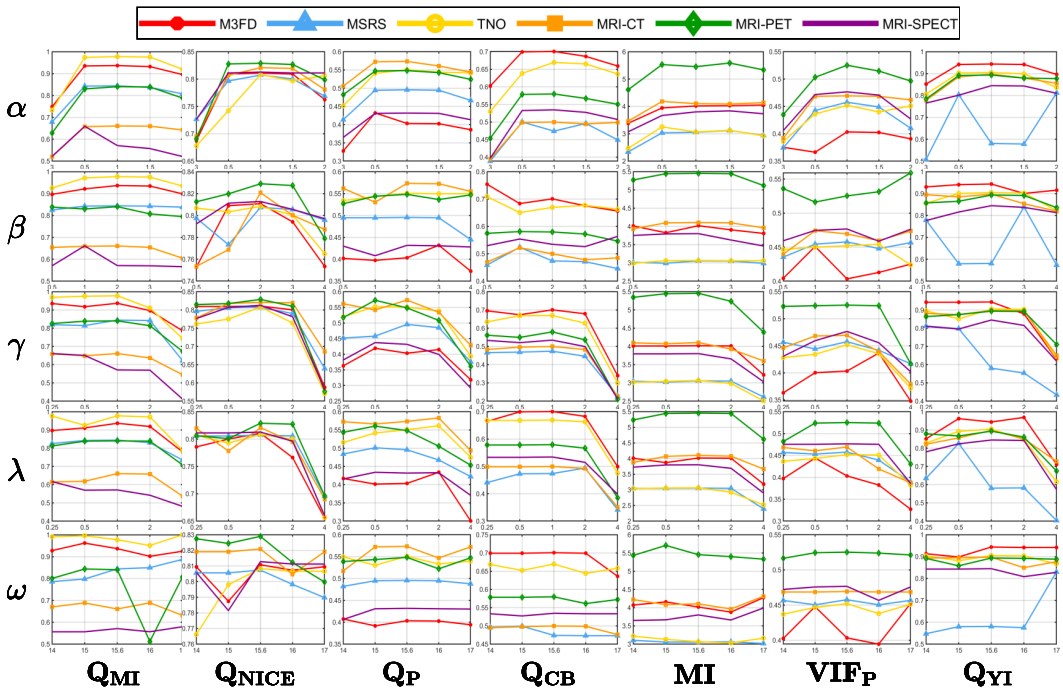


Figure 14: Sensitivity to key hyperparameters.

Table 9: Comparison of runtime, computational cost (GFLOPs), and model size (Params) on VIF and MIF tasks. All values are averaged per image.

VIF				MIF			
Method	**Time (s)**	**GFLOPs (G)**	**Params (M)**	**Method**	**Time (s)**	**GFLOPs (G)**	**Params (M)**
CDDFuse	0.006	116.851	1.186	CDDFuse	0.025	116.851	1.186
LRRNet	0.001	0.001	0.049	LRRNet	0.001	0.001	0.049
EMMA	0.058	8.861	1.516	EMMA	0.009	8.861	1.516
TC-MoA	0.543	61.000	340.580	TC-MoA	0.161	61.000	340.580
Text-DiFuse	23.818	18513	119.460	Text-DiFuse	24.424	2742.5	119.460
SigFusion	0.533	41.761	0.177	CCF	31.876	1114.000	552.810
DCEvo	0.521	15.000	2.000	SigFusion	0.037	41.761	0.177
SAGE	0.020	29.285	0.136	BSA-Fusion	0.127	13.437	9.694
TD-Fusion	0.053	3.886	0.059	Mask-Difuser	0.503	995.706	171.262
OmniFuse	2.614	172.923	78.320	MTG-Fusion	0.053	121.411	4.273
C2RF	0.089	123.869	1.325	C2RF	0.112	123.869	1.325
Ours	0.059	4.441	0.610	Ours	0.011	4.441	0.610

and thus lead to relatively slower processing. This indicates that image size has an impact on the computational efficiency of our algorithm, which suggests a promising direction for future optimization.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: see in our abstract.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Appendix.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: Our paper does not include theoretical results

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide full details for reproducing the proposed method in Section 3 and the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Our code and data will be publicly available.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide complete training and testing settings in Section 4 and Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: See it in Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See Section 4.1 and Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper conducted in this paper complies with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See it in Section 4.7 and appendix.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: There is no such risks in our paper.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: In Section 4.2, we introduce all datasets and metrics we used.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: Our paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.