RELATION MATTERS: LEARNING LOCAL RELATIONS FOR BINARY INFRARED–VISIBLE IMAGE FUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Infrared–Visible Image Fusion (IVIF) integrates thermal targets with spatial details, but full-precision models require substantial computation and storage. Binarization reduces these costs, yet binary IVIF faces three connected challenges: binary representation, where binarization can obscure local structural differences; feature propagation, where binary processing can degrade the magnitude information needed for reconstruction; and capacity-compatible training, where the binary network must learn complementary fusion patterns within its limited representational capacity. To address these challenges, we propose SDRTNet, a binary IVIF framework that jointly addresses binary representation, feature propagation, and capacity-compatible training. Sliding Directional Relation Transform (SDRT) makes local directional differences explicit by decomposing overlapping 2×2 patches into a base component and four directional relation components for separate binary processing. Local Statistical Alignment (LSA) then uses the local means and variances of raw SDRT features to calibrate binary convolution outputs before spatial reconstruction. During training, Binary Vector-Quantized Knowledge Distillation (BVQKD) organizes a full-precision teacher's fusion knowledge into a shared codebook and guides the binary student through codeword-assignment distributions, without adding inference cost. Experiments on four fusion benchmarks and two downstream tasks show competitive fusion performance with reduced computational and storage costs.

1 INTRODUCTION

Infrared–Visible Image Fusion (IVIF) integrates thermal targets from infrared images with spatial details from visible images(Tang et al. (2023); Zhang (2022)). Full-precision (FP) models achieve promising results, but their computational and storage demands can limit deployment on resource-constrained devices(Zhang et al. (2023)). Binary Neural Networks (BNNs) reduce these costs by quantizing weights and activations to one-bit values. However, information loss makes it difficult to retain local structures and magnitude information needed for reconstruction(Yang et al. (2017)).

Existing BNN methods address binarization loss through several strategies. XNOR-Net uses scaling factors to improve the numerical approximation of binary weights and activations(Rastegari et al. (2016)). Bi-Real Net preserves full-precision information through real-valued shortcuts(Liu et al. (2020a)). ReActNet adjusts activation distributions using learnable transformations(Liu et al. (2020b)). For low-level vision, BTM introduces a batch-normalization-free training strategy for binary image super-resolution(Jiang et al. (2021)). These methods improve the numerical scale, information pathway, activation response, or optimization of binary networks. However, image fusion requires local structures and magnitude information to remain useful through binary processing. The network must also learn to combine complementary information from two modalities within its limited representational capacity.

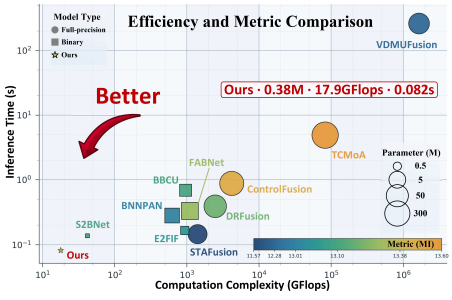


Figure 1: Efficiency and metric comparison between SDRTNet and existing methods under resolution of 1024×768.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

The difficulty of binary IVIF extends from preserving useful information to using it during feature propagation and training. Point-wise binarization can map different local patterns to the same signs, making their structures difficult to distinguish. Yet retaining distinguishable structures is only the first step. These structures must also pass through feature transformations and contribute to the features used for spatial reconstruction. Local feature levels and variation strengths matter in this process because they affect the reconstructed image content and contrast. Binary processing limits the numerical information available to these transformations, making the quality of propagated features important beyond structural representation. Even when useful information is better retained, the network must still learn to combine complementary cues from both modalities within its limited representational capacity. Accordingly, we consider three connected aspects: **Binary Representation**, **Feature Propagation**, and **Capacity-Compatible Training**. Binary representation should retain useful distinctions among local structures; feature propagation should preserve local information needed for reconstruction; and capacity-compatible training should help the network learn effective fusion patterns within its limited capacity. Together, these aspects connect the information captured by binary features, the features formed during propagation, and the fusion patterns learned during training.

Based on this analysis, we propose SDRTNet, a binary IVIF framework built around Sliding Directional Relation Transform (SDRT), Local Statistical Alignment (LSA), and Binary Vector-Quantized Knowledge Distillation (BVQKD). For **Binary Representation**, SDRT maps each overlapping 2×2 patch into a base component and four directional relation components. Independent binary branches process these components, making local directional differences available to binary activations. ISDRT reconstructs spatial features from the processed components through a fixed least-squares pseudoinverse and overlap averaging. For **Feature Propagation**, LSA uses the local means and variances of raw SDRT features to calibrate the features produced by binary convolution before ISDRT. These statistics provide a reference for local feature values and their variation that relation signs alone do not retain. A learnable gated residual controls how much this reference contributes to the features in each branch. For **Capacity-Compatible Training**, BVQKD organizes the fusion knowledge of a full-precision teacher into a learned codebook. Inspired by vector-quantized representation learning(van den Oord et al. (2017)), the codebook is trained with image reconstruction and gradient constraints so that its codewords reflect the teacher's fusion output. The binary student then learns from codeword-assignment distributions over this shared codebook, without directly matching each continuous teacher feature value. The fusion loss remains the main objective, and codeword supervision provides additional guidance. BVQKD is used only during training and adds no inference cost. Fig. 1 compares SDRTNet with existing methods in fusion performance and efficiency. The main contributions of this work are as follows:

- To mitigate the information degradation caused by binarization in IVIF, we propose SDRTNet, a unified binary framework that addresses the connected difficulties of binary representation, feature propagation, and capacity-compatible training.
- In SDRTNet, we propose three components, SDRT for preserving local directional relations, LSA for carrying local magnitude information through statistical calibration, and BVQKD for guiding the learning of fusion patterns through codebook.
- Comprehensive experiments demonstrate that our method achieves satisfactory results in all fusion tasks and downstream tasks.

2 Related Work

Infrared–Visible Image Fusion. Deep learning has substantially advanced Infrared–Visible Image Fusion (IVIF). CNN- and autoencoder-based methods learn effective local representations for extracting and integrating complementary information (Zhang & Ma (2021); Cheng et al. (2021); Wang et al. (2022); Park et al. (2024)), while GAN- and Transformer-based methods further improve visual quality by introducing adversarial learning and long-range dependency modeling (Ma et al. (2019; 2020); Wang et al. (2025); Ma et al. (2022); Zhao et al. (2024); Tian et al. (2023)). Other studies exploit background-detail decomposition or downstream task guidance to enhance structural and semantic information (Li et al. (2023); Li & Wu (2024); Liu et al. (2022); Bai et al. (2025)). Although these methods have achieved impressive fusion performance, they are predominantly full-precision and require substantial computational and storage resources, limiting their deployment on

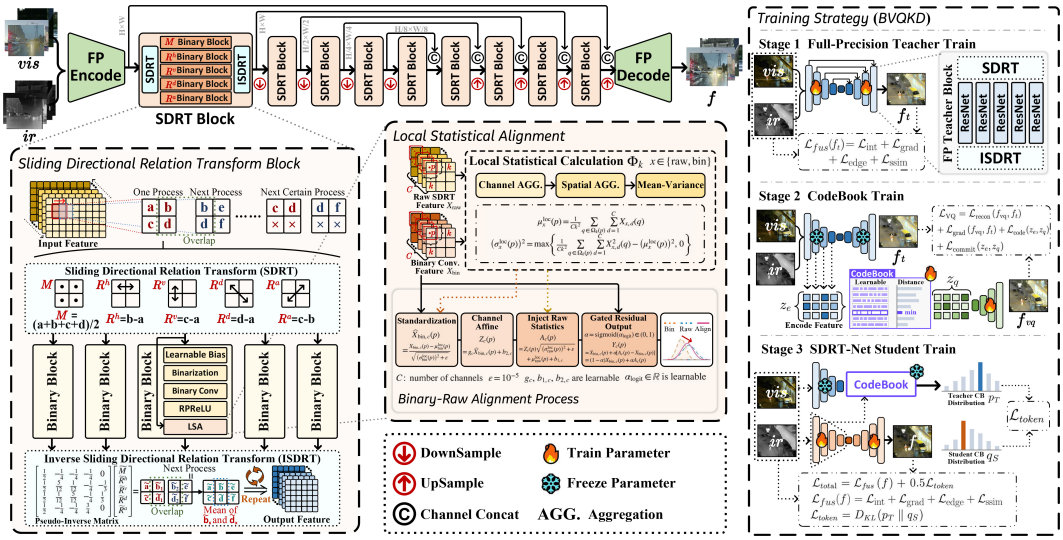


Figure 2: Overview of our proposed method SDRTNet.

resource-constrained devices. In this work, we explore binary IVIF and propose SDRTNet, which encodes local structures through directional relations to preserve richer complementary information under efficient binary computation.

Binary Neural Network. Binary Neural Networks (BNNs) quantize weights and activations into $\{-1, +1\}$, reducing computational and storage costs (Courbariaux et al. (2016)). XNOR-Net introduces scaling factors for better numerical approximation, while Bi-Real Net, IR-Net, and ReActNet improve information retention or activation adaptation (Rastegari et al. (2016); Liu et al. (2020a); Qin et al. (2020); Liu et al. (2020b)). For low-level vision, BTM removes batch normalization through distribution-based training (Jiang et al. (2021)); E2FIF preserves an end-to-end full-precision information flow; BBCU redesigns binary convolution units; and FABNet and BNNPan exploit frequency-aware learning or statistical compensation (Song et al. (2023); Xia et al. (2023); Jiang et al. (2023); Hou et al. (2025)). Despite these advances, binary IVIF remains sensitive to the loss of local structures and magnitude information. The limited-capacity network must also learn useful fusion patterns from the information it retains. These difficulties motivate our study of binary representation, feature propagation, and capacity-compatible training in SDRTNet.

3 Method

3.1 Overview

Given an infrared image ir and a visible image vis, SDRTNet aims to generate a fused image f. As illustrated in Fig. 2, SDRTNet adopts a hybrid-precision encoder–decoder architecture. The images are first concatenated and processed by a full-precision encoder. The intermediate features pass through a multi-scale binary backbone composed of SDRT Blocks, while a full-precision decoder reconstructs the fused image. Within each block, SDRT produces a base component and four directional relation components, which are processed by independent binary branches. LSA then uses local statistics of the raw SDRT features to calibrate the binary-convolution features before ISDRT reconstructs the spatial features. This supplies local magnitude information alongside the learned relation features. Skip connections and channel concatenation pass multi-scale information between the encoder and decoder. During training, BVQKD uses a full-precision teacher to provide additional fusion-pattern guidance through codeword-assignment distributions.

3.2 Design of SDRTNet

Encoding local relations helps distinguish structures, but their binary signs still lose magnitude information needed for reconstruction. Retaining these cues also leaves the network with the task of learning effective fusion patterns under limited capacity. SDRTNet addresses these connected dif-

ficulties through SDRT/ISDRT for local relation representation and reconstruction, LSA for feature propagation with local statistical guidance, and BVQKD for capacity-compatible training.

Binary Representation via SDRT. Point-wise binarization can assign the same signs to different local patterns, making the relations among neighboring features difficult to distinguish. SDRT makes these relations explicit by operating on sliding overlapping 2×2 feature patches. For the window whose upper-left position is (i, j), we denote one patch of local feature $X \in \mathbb{R}^{m \times n}$ as

$$X_{i,j} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \mathbb{R}^{2\times 2}. \tag{1}$$

SDRT extracts one base component and four directional relation components:

$$\begin{aligned} M_{i,j} &= \frac{a+b+c+d}{2}, \quad R^h_{i,j} = b-a, \quad R^v_{i,j} = c-a, \\ R^d_{i,j} &= d-a, \qquad\qquad R^a_{i,j} = c-b. \end{aligned} \tag{2}$$

Then SDRT calculates each patch. And $M \in \mathbb{R}^{(m-1)\times(n-1)}$ represents the local base feature, while same dimension R^h, R^v, R^d, and R^a denote the horizontal, vertical, main-diagonal, and anti-diagonal relations, respectively. The five components are processed by independent Binary Blocks. For $X^\ell_{\text{raw}} \in \{M, R^h, R^v, R^d, R^a\}$, each branch is formulated as $X^\ell_{\text{bin}} = \text{RPReLU}\left(\text{BConv}_\ell\left(\text{Sign}(X^\ell_{\text{raw}} + b_\ell)\right)\right)$, where b_ℓ is a learnable bias. Before spatial reconstruction, LSA calibrates X^ℓ_{bin} using the local statistics of X^ℓ_{raw}, producing aligned relation feature Y^ℓ.

Let $\widehat{\mathbf{y}} = [Y^M, Y^h, Y^v, Y^d, Y^a]^\mathsf{T}$ collect the five calibrated results. ISDRT then reconstructs the local spatial patch as

$$\text{vec}\left(\widetilde{\mathrm{X}}\right) = \mathbf{A}^+\widehat{\mathbf{y}}, \qquad \mathbf{A}^+ = \begin{bmatrix} \frac{1}{2} & -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & 0 \\ \frac{1}{2} & \frac{5}{12} & \frac{1}{12} & -\frac{1}{4} & -\frac{1}{3} \\ \frac{1}{2} & \frac{1}{12} & \frac{5}{12} & -\frac{1}{4} & \frac{1}{3} \\ \frac{1}{2} & -\frac{1}{4} & -\frac{1}{4} & \frac{3}{4} & 0 \end{bmatrix}. \tag{3}$$

The derivation of $\mathbf{A}^+$ is provided in the Appendix A. The reconstructed vectors $\text{vec}\left(\widetilde{\mathrm{X}}\right)$ are reshaped into 2×2 patches and averaged over overlapping windows to obtain the final spatial feature $\widetilde{\mathrm{X}}$.

Feature Propagation Consistency via LSA. Binary activations of the relation components encode thresholded directional differences but do not retain their exact magnitudes. To alleviate this problem, LSA takes the raw SDRT feature X^ℓ_{raw} and its corresponding binary-convolution feature X^ℓ_{bin} as inputs. For clarity, the branch index is omitted below. For each spatial position p, Local Statistical Calculation $\Phi(\cdot)$ performs channel aggregation on the features and their squared values across all C channels, followed by spatial aggregation that averages them over the local $k \times k$ neighborhood centered at p, yielding local mean and variance. Applied separately to X_{raw} and X_{bin}, it produces the raw statistics $\left(\mu^{\text{loc}}_{\text{raw}}(p), (\sigma^{\text{loc}}_{\text{raw}}(p))^2\right)$ and binary statistics $\left(\mu^{\text{loc}}_{\text{bin}}(p), (\sigma^{\text{loc}}_{\text{bin}}(p))^2\right)$. The binary statistics are used to standardize X_{bin}, while the raw statistics provide the reference mean and variance for its subsequent alignment.

At channel c and position p, LSA standardizes the feature value produced by binary convolution as $\widehat{X}_{\text{bin},c}(p) = \left(X_{\text{bin},c}(p) - \mu^{\text{loc}}_{\text{bin}}(p)\right)/\sqrt{(\sigma^{\text{loc}}_{\text{bin}}(p))^2 + \epsilon}$, where $\epsilon = 10^{-5}$ ensures numerical stability. A learnable channel-wise affine transformation produces $Z_c(p) = g_c\widehat{X}_{\text{bin},c}(p) + b_{2,c}$, where g_c and $b_{2,c}$ are the scale and bias for channel c. This feature value is scaled and shifted using raw SDRT statistics, giving $A_c(p) = Z_c(p)\sqrt{(\sigma^{\text{loc}}_{\text{raw}}(p))^2 + \epsilon} + \mu^{\text{loc}}_{\text{raw}}(p) + b_{1,c}$, where $b_{1,c}$ is a learnable channel bias. The calibrated feature value is obtained through a gated residual, $Y_c(p) = X_{\text{bin},c}(p) + \alpha\left[A_c(p) - X_{\text{bin},c}(p)\right]$, where $\alpha = \text{sigmoid}(\alpha_{\text{logit}})$ is controlled by the learnable logit α_{logit}. The affine parameters and gate control the statistical calibration without requiring exact equality of local feature statistics. Finally, the calibrated branch features $Y^\ell \in \{Y^M, Y^h, Y^v, Y^d, Y^a\}$ are collected as $\widehat{\mathbf{y}}$ and passed to ISDRT for local spatial reconstruction.

Capacity-Compatible Training via BVQKD. With local structures and magnitude information available for reconstruction, the binary network still needs to learn fusion patterns. A full-precision

teacher offers guidance, but its continuous features may be difficult for the student to match. BVQKD organizes teacher knowledge in a fusion-aware codebook and guides the student through codeword-assignment distributions.

In **Stage 1**, a full-precision teacher generates the fused image f_t from ir and vis. It follows the same SDRT–ISDRT architecture as SDRTNet, except that each binary block is replaced with a full-precision ResNet block. The teacher is optimized with intensity, gradient, edge, and structural-similarity losses:

$$\mathcal{L}_{fus}(f_t) = 1.5\,\mathcal{L}_{\text{int}} + 1.5\,\mathcal{L}_{\text{grad}} + 0.5\,\mathcal{L}_{\text{edge}} + 2.0\,\mathcal{L}_{\text{ssim}}. \tag{4}$$

These coefficients are selected to balance the relative magnitudes of the four loss terms. After training, the teacher parameters are frozen.

In **Stage 2**, features from the frozen teacher encoder are projected into latent space as z_e. Each spatial embedding in z_e is quantized by selecting its nearest codeword from the learnable codebook $\{e_\rho\}_{\rho=1}^{K}$:

$$\rho^* = \arg\min_{\rho} \|z_e - e_\rho\|_2^2, \qquad \widetilde{z}_q = e_{\rho^*}. \tag{5}$$

Thus, $\widetilde{z}_q$ represents this embedding using the selected codeword e_{ρ^*}. A lightweight decoder reconstructs an image from the quantized feature, producing $f_{vq} = D_q(\widetilde{z}_q)$. The tokenizer and codebook are optimized with reconstruction, gradient, codebook, and commitment losses:

$$\mathcal{L}_{\text{VQ}}(f_{vq}) = \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{grad}} + \mathcal{L}_{\text{code}} + 0.25\,\mathcal{L}_{\text{commit}}. \tag{6}$$

In this stage, $\mathcal{L}_{\text{recon}} = \|f_{vq} - f_t\|_1$ and $\mathcal{L}_{\text{grad}} = \|G(f_{vq}) - G(f_t)\|_1$ encourage reconstruction of the teacher's fused-image content and spatial details, respectively. Meanwhile, $\mathcal{L}_{\text{code}} = \|\widetilde{z}_q - \text{sg}(z_e)\|_2^2$ updates the selected codeword toward the projected embedding, whereas $\mathcal{L}_{\text{commit}} = \|z_e - \text{sg}(\widetilde{z}_q)\|_2^2$ encourages z_e to approach its assigned codeword. Here, $G(\cdot)$ denotes the Sobel-gradient magnitude and $\text{sg}(\cdot)$ denotes stop-gradient. The two latent-space losses thus bring projected embeddings and assigned codewords closer. After training, the codebook is frozen.

In **Stage 3**, the frozen teacher encoder provides features F_t, which are used with the fusion-aware codebook to obtain a teacher codeword-assignment distribution p_T. The SDRTNet student extracts its own features and predicts the corresponding distribution q_S over the same codewords. This supervision describes how teacher features relate to the codewords, allowing the student to learn their assignment patterns without directly matching continuous feature values:

$$\mathcal{L}_{token} = D_{\text{KL}}(p_T \| q_S), \tag{7}$$

where KL divergence is averaged over all spatial locations. The student is optimized with $\mathcal{L}_{total}$:

$$\mathcal{L}_{total}(f) = \mathcal{L}_{fus}(f) + 0.5\,\mathcal{L}_{token}, \tag{8}$$

where f denotes the student fusion output. The fusion loss remains the primary training objective, while the token loss supplies additional guidance for learning the teacher's fusion patterns. BVQKD is used only during training and introduces no additional inference cost.

4 Experiment

4.1 Infrared–Visible Image Fusion

Experimental Setup. All experiments are conducted on a single NVIDIA GeForce RTX 3090 GPU. We randomly select 2000 pairs of infrared-visible image from the LLVIP(Jia et al. (2021)) for training. During training, the two modalities undergo synchronized random resized cropping to 448×448 and horizontal flipping. The three-stage training strategy(BVQKD) uses a batch size of 4 and the AdamW optimizer(Ilya & Frank (2019)) with an initial learning rate of 2×10^{-4}. The full-precision teacher, codebook, and SDRTNet student are trained for 50, 30, and 50 epochs, respectively, with cosine annealing learning-rate decay. The trained model is evaluated on the LLVIP(Jia et al. (2021)), TNO(Toet & Hogervorst (2012)), M3FD(Liu et al. (2022)), and MSRS(Tang et al. (2022)) datasets to assess its generalization across different scenes and imaging conditions.

SOTA Methods and Metrics. We compare SDRTNet with ten representative methods, including five full-precision IVIF methods, TCMoA(Zhu et al. (2024)), ControlFusion(Tang et al. (2025)),

VDMUFusion(Shi et al. (2025)), DRFusion(Shi et al. (2026)), and STAFusion (Zhai et al. (2026)), as well as five binary methods, E2FIF(Song et al. (2023)), BBCU(Xia et al. (2023)), FABNet(Jiang et al. (2023)), S2BNet(Jiang et al. (2025)), and BNNPan(Hou et al. (2025)). Six widely used metrics are employed for quantitative evaluation. EN(Roberts et al. (2008)) and MI(Qu et al. (2002)) assess information richness and transfer. SF(Eskicioglu & Fisher (1995)) and AG(Cui et al. (2015)) quantify spatial detail. EI(Balakrishnan & Priya (2018)) measures edge strength, and CC(Teo & Lau (2012)) measures structural correlation with the source images. Higher values indicate better performance for all metrics.

Qualitative Experiment. Fig. 3 compares fusion results on three representative infrared-visible scenes. In the first scene, the red-box crop shows distant building structures, while the green-box crop highlights a pedestrian in the mist. Several methods retain thermal target cues but show blurred building patterns or incomplete pedestrian boundaries. SDRTNet retains clearer background structures and a more distinct pedestrian outline, with better contrast around the target. In the second scene, the red-box crop contains a small target partly hidden by haze, and the green-box crop shows mountain and cloud textures around it. Some competing results make the target more prominent but lose background details. SDRTNet preserves a more distinct target together with more visible local texture changes. In the third scene, the red-box crop highlights a pedestrian in an open area, while the green-box crop covers nearby vegetation and ground. SDRTNet produces a sharper pedestrian outline and retains fine background structures in these regions. Across the displayed examples, the results suggest that SDRTNet can combine infrared target cues with visible details, including in the low-contrast and haze-affected scenes shown here.

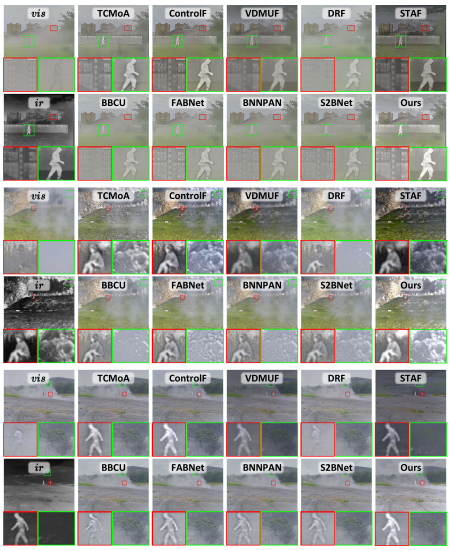


Figure 3: Qualitative comparison of SDRTNet with existing full-precision and binary methods.

Quantitative Experiment. Tab. 1 reports quantitative comparisons on the LLVIP, TNO, M3FD, and MSRS datasets, where higher values indicate better fusion quality. Despite using binary operations and a compact low-parameter architecture, SDRTNet consistently achieves the strongest or near-strongest performance among binary methods and remains competitive with full-precision approaches. On LLVIP, it ranks first among BNN methods on all six metrics, with an AG score of 3.98 that exceeds all full-precision methods and a CC score of 0.931 that is comparable to the best full-precision result. On TNO, SDRTNet achieves the best binary results on EN, MI, SF, EI, and AG, while remaining close to the full-precision leaders. On M3FD, it even surpasses all full-precision methods in EN and MI, reaching 6.79 and 13.57, respectively. On MSRS, it obtains the best or tied-best binary performance on five metrics and remains competitive with the full-precision methods. These results demonstrate that SDRTNet can preserve high-quality fusion performance under substantially reduced parameter and computational costs, offering a favorable efficiency–accuracy trade-off for resource-constrained applications.

4.2 Ablation

Binary Representation. To examine whether SDRT preserves fusion-relevant spatial information during binarization, we replace SDRT with other binary transformations in the feature reconstruction experiment, in which binarization is performed in the method's transformation domain, after which the processed features are inversely transformed back to the spatial domain and compared with raw spatial features. We compare direct point-wise Plain Binary, DWT Binary(Alessio (2016)), Non-overlap SDRT(NSDRT), and the complete overlapping SDRT. The reconstruction quality is evaluated using MSE and MAE for numerical errors, PSNR for reconstruction fidelity, Correlation and SSIM for structural consistency, and NMI(Estevez et al. (2009)) for retained information. As shown in Tab. 2, Plain Binary produces substantial feature distortion, with an MSE of 0.1527 and a PSNR of only 15.0766. DWT Binary and Non-overlap SDRT alleviate this degradation, while SDRT achieves the lowest MSE and MAE, together with the highest PSNR, Correlation, SSIM, and NMI.

IVIF on LLVIP Dataset							IVIF on TNO Dataset						
Method	**EN↑**	**MI↑**	**SF↑**	**EI↑**	**AG↑**	**CC↑**	**Method**	**EN↑**	**MI↑**	**SF↑**	**EI↑**	**AG↑**	**CC↑**
Full-Precision Methods							*Full-Precision Methods*						
TCMoA	7.58	15.15	14.34	40.92	3.92	0.928	TCMoA	6.80	13.60	11.29	44.00	4.27	0.684
ControlF	7.36	14.72	15.78	41.15	3.94	0.895	ControlF	6.79	13.58	10.73	42.21	4.16	0.674
VDMUF	7.27	14.55	8.36	25.03	2.47	0.959	VDMUF	6.14	12.28	6.24	23.01	2.40	0.668
DRF	7.53	15.06	10.26	23.00	2.18	0.920	DRF	6.56	13.12	10.44	27.82	2.87	0.673
STAF	6.99	13.98	4.36	11.66	1.12	0.937	STAF	5.78	11.57	5.55	24.29	2.35	0.611
Binary Methods							*Binary Methods*						
E2FIF	7.51	15.02	9.14	28.18	3.56	0.918	E2FIF	6.54	13.08	7.75	33.37	3.26	0.682
BBCU	7.51	15.02	10.01	30.19	3.77	0.918	BBCU	6.55	13.10	7.65	31.62	3.08	0.673
FABNet	7.52	15.03	10.02	29.86	3.85	0.919	FABNet	6.58	13.16	8.89	36.70	3.66	0.678
BNNPan	7.47	14.94	9.69	28.93	3.68	0.922	BNNPan	6.51	13.01	9.35	35.35	3.63	0.680
S2BNet	7.49	14.97	9.36	28.96	3.67	0.924	S2BNet	6.55	13.10	8.40	34.12	3.40	0.680
Ours	$\mathbf{7.54}^{2}_{1}$	$\mathbf{15.07}^{2}_{1}$	$\mathbf{10.62}^{3}_{1}$	$\mathbf{30.57}^{3}_{1}$	$\mathbf{3.98}^{1}_{1}$	$\mathbf{0.931}^{3}_{1}$	**Ours**	$\mathbf{6.69}^{3}_{1}$	$\mathbf{13.38}^{3}_{1}$	$\mathbf{9.94}^{4}_{1}$	$\mathbf{39.85}^{3}_{1}$	$\mathbf{3.97}^{3}_{1}$	$\mathbf{0.681}^{3}_{2}$
IVIF on M3FD Dataset							**IVIF on MSRS Dataset**						
Method	**EN↑**	**MI↑**	**SF↑**	**EI↑**	**AG↑**	**CC↑**	**Method**	**EN↑**	**MI↑**	**SF↑**	**EI↑**	**AG↑**	**CC↑**
Full-Precision Methods							*Full-Precision Methods*						
TCMoA	6.72	13.45	14.74	52.98	5.12	0.585	TCMoA	6.95	13.90	12.21	43.36	4.08	0.550
ControlF	6.77	13.53	14.90	52.06	5.01	0.585	ControlF	7.18	14.36	12.54	44.04	4.12	0.549
VDMUF	6.45	12.90	8.39	31.23	3.09	0.461	VDMUF	6.21	12.43	7.08	25.69	2.52	0.555
DRF	6.63	13.27	11.49	39.48	3.82	0.589	DRF	6.94	13.88	11.49	40.45	3.78	0.549
STAF	6.26	12.52	5.01	18.39	1.73	0.389	STAF	5.81	11.62	3.96	15.46	1.46	0.524
Binary Methods							*Binary Methods*						
E2FIF	6.74	13.48	13.34	49.42	4.69	0.487	E2FIF	6.96	13.92	10.90	39.54	4.09	0.551
BBCU	6.70	13.40	13.40	50.27	4.06	0.487	BBCU	6.97	13.94	10.07	39.85	3.70	0.552
FABNet	6.73	13.46	13.63	50.01	4.40	0.498	FABNet	6.96	13.91	10.17	39.62	3.55	0.551
BNNPan	6.62	13.24	13.31	50.67	4.36	0.504	BNNPan	6.95	13.91	10.04	39.25	3.69	0.551
S2BNet	6.70	13.41	13.13	49.62	4.85	0.486	S2BNet	6.90	13.80	10.93	39.90	3.49	0.553
Ours	$\mathbf{6.79}^{1}_{1}$	$\mathbf{13.57}^{1}_{1}$	$\mathbf{13.67}^{3}_{1}$	$\mathbf{50.35}^{4}_{2}$	$\mathbf{4.88}^{3}_{1}$	$\mathbf{0.515}^{4}_{1}$	**Ours**	$\mathbf{6.97}^{2}_{1}$	$\mathbf{13.95}^{2}_{1}$	$\mathbf{10.95}^{4}_{1}$	$\mathbf{40.12}^{4}_{1}$	$\mathbf{3.79}^{4}_{2}$	$\mathbf{0.553}^{2}_{1}$

Table 1: Quantitative comparison on the LLVIP, TNO, M3FD, and MSRS datasets. The darker the background color, the higher the rank of the method. For SDRTNet, the superscript and subscript denote its rank among all methods and among BNN methods, respectively.

	Feature Reconstruction after Binary Processing							*Fusion Performance on TNO*					
Method	MSE↓	MAE↓	PSNR↑	Corr↑	SSIM↑	NMI↑	Method	EN↑	MI↑	SF↑	EI↑	AG↑	CC↑
Plain B.	0.1527	0.2535	15.0766	0.8029	0.7380	0.2113	Plain B.	4.01	8.01	5.95	16.63	1.66	0.419
DWT B.	0.0281	0.0975	22.0282	0.9186	0.9162	0.3606	DWT B.	6.04	12.09	7.53	30.24	3.03	0.673
NSDRT B.	0.0277	0.0991	22.1009	0.9200	0.9176	0.3549	NSDRT B.	5.86	11.71	8.12	34.21	3.46	0.672
Ours	**0.0187**	**0.0693**	**23.7540**	**0.9470**	**0.9410**	**0.4384**	**Ours**	**6.69**	**13.38**	**9.94**	**39.85**	**3.97**	**0.681**

Table 2: Evaluation of binary representation and its effect on IVIF. **Bold** and underlined values indicate the best and second-best results, respectively.

These results suggest that performing binarization in the SDRT's transformation domain, preserves raw spatial features more faithfully, especially their local structures .

Fig. 4 visualizes the spatial features and fused images. "B. Fea." denotes Binary Feature, while "Fus." denotes Fusion. Plain B. Fea. shows blurred structures and weak local contrast, consistent with its larger reconstruction errors. DWT B. Fea. and Non-overlap SDRT B. Fea. retain more visible structures, although fine details remain less distinct. SDRT B. Fea. shows clearer local changes, and SDRT B. Fus. produces sharper details in the highlighted regions. These observations agree with the quantitative comparison and suggest that the proposed representation helps retain local structures through binary processing.

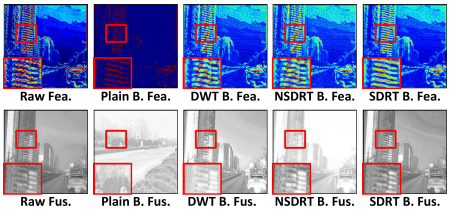


Figure 4: Visualization of different binary transformations.

The benefit of retaining these local structures is also reflected in the fusion results in Tab. 2. SDRT achieves the highest values on all six metrics among the compared representations. Relative to Plain Binary, it improves information content, local detail, and correlation with the source images. Its advantage over DWT Binary and Non-overlap SDRT further indicates that the proposed directional relations and overlapping neighborhoods provide useful information for the final fusion result.

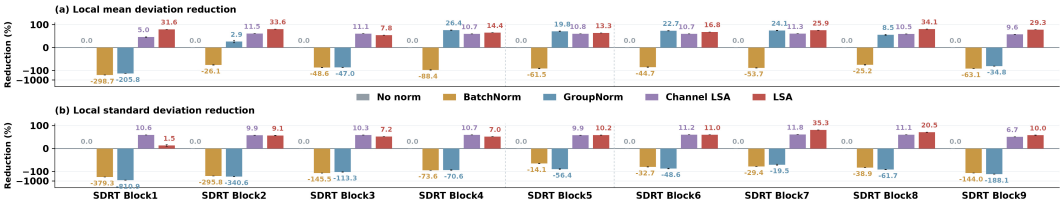


Figure 5: Reduction in local mean and standard deviation differences across SDRT blocks.

Feature Propagation Consistency. To examine the role of local statistical guidance during feature propagation, we compare LSA with No norm, BatchNorm, GroupNorm, and Channel LSA, keeping the remaining architecture unchanged. No norm removes statistical calibration, while Channel LSA retains channel aggregation but excludes spatial aggregation. Fig. 5 shows that LSA reduces the local mean and standard deviation discrepancies relative to raw SDRT features in all nine blocks. Channel LSA also reduces these discrepancies, while No norm gives zero reduction. BatchNorm and GroupNorm often increase them. Tab. 3 shows that LSA achieves the best fusion results among these variants. Its advantage over Channel LSA supports using spatial neighborhoods when estimating local statistics. Together, these results support using raw SDRT statistics as a reference for processing relation features before spatial reconstruction.

Method	EN↑	MI↑	SF↑	EI↑	AG↑	CC↑
No norm	4.75	9.49	6.08	19.87	1.73	0.526
BatchNorm	3.27	6.55	6.53	17.57	1.78	0.482
GroupNorm	6.01	12.03	5.77	14.64	1.91	0.676
Channel LSA	<u>6.08</u>	<u>12.15</u>	<u>7.41</u>	<u>29.02</u>	<u>2.84</u>	<u>0.679</u>
Ours	**6.69**	**13.38**	**9.94**	**39.85**	**3.97**	**0.681**

Table 3: Comparison of normalization and statistical calibration methods.

Capacity-Compatible Training. To assess whether BVQKD provides capacity-compatible supervision for BNNs, we compare five training objectives under the same student architecture in Stage 3. Fusion Only trains the student solely with the fusion loss. Building on this baseline, Output KD matches the student's fused output to the teacher's. Feature L1 KD forces the binary student feature at each channel and spatial position to match the corresponding continuous teacher feature with L1 loss. Hard-token KD assigns each teacher feature to a single codeword and uses this one-hot assignment to supervise the student, discarding its relations to the other codewords. As shown in Tab. 4, output- and feature-level matching provide limited or inconsistent improvements, while Hard-token KD achieves stronger overall performance, indicating that discrete supervision is more compatible with binary representations. BVQKD further achieves the best results on all metrics by preserving the relative probabilities among codewords, providing richer discrete supervision within the capacity of the binary student.

Method	**EN↑**	**MI↑**	**SF↑**	**EI↑**	**AG↑**	**CC↑**
Fusion Only	5.45	10.89	6.22	18.40	2.19	0.612
Output KD	5.90	11.81	7.11	18.09	2.87	0.646
Feature L1 KD	5.71	11.42	<u>8.32</u>	20.74	2.78	0.619
Hard-token KD	<u>6.36</u>	<u>12.72</u>	7.12	<u>29.03</u>	<u>2.97</u>	<u>0.675</u>
BVQKD	**6.69**	**13.38**	**9.94**	**39.85**	**3.97**	**0.681**

Table 4: Comparison of different student-training objectives. **Bold** and <u>underlined</u> values indicate the best and second-best results, respectively.

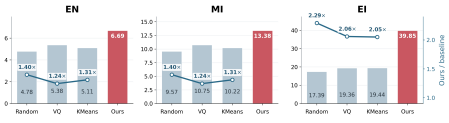


Figure 6: Codebook ablation where lines show the improving ratios of BVQKD.

We further examine how the codebook $\mathcal{E}$ for discrete supervision should be constructed in Stage 2. Frozen Random Codebook keeps randomly initialized codewords fixed, without codebook learning. VQ Only learns $\mathcal{E}$ only using $\mathcal{L}_{\text{code}} + 0.25\mathcal{L}_{\text{commit}}$. Offline KMeans constructs and freezes $\mathcal{E}$ using cluster centers obtained from pre-extracted teacher embeddings z_e. In contrast, BVQKD jointly learns the tokenizer and codebook using the complete $\mathcal{L}_{\text{VQ}}$. Fig. 6 shows that VQ Only and Offline KMeans improve upon the random codebook but remain clearly inferior to BVQKD, particularly in EI. This indicates that latent-space quantization or offline clustering alone cannot ensure that the codewords encode fusion-relevant structures. By incorporating $\mathcal{L}_{\text{recon}}$ and $\mathcal{L}_{\text{grad}}$, BVQKD produces discrete supervision that better preserves both information content and edge details.

4.3 Downstream Tasks

Object Detection. To evaluate whether the fused images produced by SDRTNet retain information useful for high-level visual understanding, we conduct downstream object detection experiments

Method	Multi-Modal Detection							Multi-Modal Semantic Segmentation								
	Peo	Car	Bus	Mot	Lam	Tru	mAP@0.5	Unl	Car	Per	Bik	Cur	CS	GD	CC	Bum	mIoU
vis	74.0	91.4	93.6	64.9	80.7	82.7	81.2	96.2	68.9	42.9	52.3	37.0	46.7	24.5	49.7	55.3	52.6
ir	84.6	88.3	89.7	62.9	56.5	84.2	77.7	96.3	66.8	54.7	47.9	32.6	40.2	29.1	40.0	51.5	51.0
TCMoA	82.9	91.9	92.5	68.6	80.7	84.2	83.5	96.6	71.3	51.9	52.2	46.3	46.5	25.3	52.3	53.2	55.1
ControlF	83.4	91.1	92.6	68.7	80.1	85.3	83.5	96.3	67.7	51.0	50.4	36.4	46.8	23.5	52.2	55.7	53.3
VDMUF	83.1	91.7	93.6	66.2	78.7	83.4	82.8	96.5	70.1	54.1	51.7	39.4	46.1	24.4	50.4	56.4	54.3
DRF	80.5	90.4	93.6	73.7	77.3	86.1	83.6	96.3	67.2	51.8	50.5	36.2	44.2	24.2	46.5	52.1	52.1
STAF	84.2	89.8	92.4	64.8	70.2	85.5	81.2	96.4	68.9	54.9	49.5	34.4	43.7	26.7	47.7	52.5	52.7
E2FIF	84.0	92.0	92.4	68.2	79.4	84.0	83.3	96.3	67.6	52.5	51.3	37.9	45.8	24.7	49.3	48.6	52.7
BBCU	83.6	91.5	92.6	67.0	79.2	84.5	83.1	96.4	69.2	53.4	42.0	38.3	46.8	24.1	49.8	53.9	52.7
FABNet	83.8	91.7	93.1	66.3	80.7	84.5	83.4	96.4	68.9	52.8	50.8	37.4	46.2	24.9	49.8	52.2	53.3
BNNPan	84.0	91.6	92.2	65.6	78.9	83.3	82.6	96.4	68.8	53.5	51.3	39.2	44.9	24.5	51.4	51.4	53.5
S2BNet	83.5	91.9	92.4	66.6	80.6	85.0	83.3	96.3	67.8	53.1	51.4	38.0	44.6	24.1	50.4	53.5	53.2
Ours	**84.2**$^{2}_{1}$	**92.2**$^{1}_{1}$	**91.4**$^{12}_{6}$	**68.7**$^{2}_{1}$	**81.3**$^{1}_{1}$	**85.3**$^{3}_{1}$	**83.8**$^{1}_{1}$	**96.7**$^{1}_{1}$	**71.1**$^{2}_{1}$	**54.5**$^{3}_{1}$	**52.1**$^{3}_{1}$	**43.2**$^{2}_{1}$	**47.3**$^{1}_{1}$	**26.2**$^{3}_{1}$	**52.1**$^{3}_{1}$	**55.9**$^{2}_{1}$	**55.5**$^{1}_{1}$

Table 5: Downstream results on Multi-Modal Detection and Multi-Modal Semantic Segmentation.

on M3FD-Detection (Liu et al. (2022)). The dataset contains 4200 paired infrared-visible images covering six categories: People, Car, Bus, Motorcycle, Lamp, and Truck. We divide the dataset into training, validation, and test sets using a 7:1:2 ratio. The fused images generated by each method are used as detector inputs, while the detector architecture and training configuration remain unchanged for a fair comparison. Specifically, we adopt YOLOv11 (ultralytics (2024); Khanam & Hussain (2024)) and train it using SGD for 200 epochs with a batch size of 16 and an initial learning rate of 0.01. Detection performance is evaluated using category-wise AP and the overall mAP@0.5.

As shown in Tab. 5, SDRTNet achieves the highest mAP@0.5 of 83.8%, exceeding DRFusion, the strongest competing fusion method, and outperforming the visible-only and infrared-only inputs by 2.6 and 6.1 percentage points, respectively. SDRTNet obtains the best results for Car and Lamp while remaining competitive for People, Motorcycle, and Truck. The improvement over both single-modality inputs confirms that its fused images preserve complementary visible textures and infrared thermal cues rather than favoring only one modality. Moreover, its advantage over existing fusion methods indicates that the information retained by SDRTNet remains discriminative for downstream recognition despite its binary architecture.

Semantic Segmentation. We further evaluate the pixel-level semantic information preserved by SDRTNet on MSRS-Segmentation (Tang et al. (2022)). The dataset consists of 1444 paired infrared-visible images spanning nine semantic categories: Unlabeled, Car, Person, Bike, Curve, Car Stop, Guardrail, Color Cone, and Bump. Following the official data split, the fused images generated by different methods are separately used to train and evaluate the same segmentation model. We employ DeepLabV3+ (Chen et al. (2018)) as the segmentation network and optimize it using SGD and cross-entropy loss for 1000 epochs. The batch size is set to 16, the initial learning rate is 0.01, and cosine annealing is used to adjust the learning rate during training. Performance is measured using category-wise IoU and the overall mIoU.

As reported in Tab. 5, SDRTNet achieves the highest mIoU of 55.5%, surpassing the strongest competing fusion method by 1.1 percentage points and the visible-only and infrared-only inputs by 2.9 and 4.5 points, respectively. It ranks first for Unlabeled, Car, Curve, and Car Stop and second for Bike and Bump. Its improvements on both object categories and structurally sensitive categories indicate that SDRTNet preserves semantic content and boundary information required for dense prediction. Together with the detection results, these findings demonstrate that SDRTNet retains fusion information beneficial to both object-level recognition and pixel-level semantic understanding.

5 Conclusion

This work presents SDRTNet, a binary IVIF framework that connects binary representation, feature propagation, and capacity-compatible training. SDRT encodes local directional relations so that binary features can distinguish more local structures, and ISDRT reconstructs spatial features from overlapping neighborhoods. LSA supplies local magnitude information through the statistics of raw SDRT features, providing a reference for calibrating binary-convolution features. BVQKD helps the network learn fusion patterns by expressing teacher guidance through codeword-assignment distributions, without additional inference cost. Experiments on IVIF benchmarks and downstream tasks show that these components support competitive fusion quality with reduced computational and storage costs. These results support considering the information represented, its use for reconstruction, and the learning of fusion patterns together in binary IVIF.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript organization, language editing, LaTeX drafting, and brainstorming the organization of experiments and analyses. All AI-assisted material was reviewed and independently verified by the authors, including experimental protocols, code, numerical results, citations, and technical claims. The authors take full responsibility for the final content of the manuscript.

REPRODUCIBILITY STATEMENT

The architecture and principal equations are described in Section 3, and the training and evaluation protocols are summarized in Section 4. The appendix provides additional implementation details, dataset construction and partitioning procedures, hyperparameters, ablations, mechanism analyses, and complete evaluation results.

REFERENCES

Silvia Maria Alessio. *Discrete Wavelet Transform (DWT)*, pp. 645–714. Springer International Publishing, Cham, 2016. ISBN 978-3-319-25468-5. doi: 10.1007/978-3-319-25468-5_14.

Haowen Bai, Jiangshe Zhang, Zixiang Zhao, Yichen Wu, Lilun Deng, Yukun Cui, Tao Feng, and Shuang Xu. Task-driven image fusion with learnable fusion loss. In *Proceedings of 38th the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7457–7468, 2025. doi: 10.1109/CVPR52734.2025.00699.

Rajalingam Balakrishnan and Rainy Priya. International journal of engineering science invention. pp. 52–60, 08 2018.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (eds.), *Proceedings of the 15th European Conference on Computer Vision*, pp. 833–851, Cham, 2018. Springer International Publishing. ISBN 978-3-030-01234-2.

Chunyang Cheng, Xiao-Jun Wu, Tianyang Xu, and Guoyang Chen. Unifusion: A lightweight unified image fusion network. *IEEE Transactions on Instrumentation and Measurement*, 70:1–14, 2021. doi: 10.1109/TIM.2021.3109379.

Matthieu Courbariaux, Itay Hubara, Daniel Soudry, and Ran El-Yaniv. Binarized neural networks: Training deep neural networks with weights and activations constrained to +1 or -1. 02 2016. doi: 10.48550/arXiv.1602.02830.

Guangmang Cui, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Detail preserved fusion of visible and infrared images using regional saliency extraction and multi-scale image decomposition. *Optics Communications*, 341:199–209, 2015. ISSN 0030-4018.

A.M. Eskicioglu and P.S. Fisher. Image quality measures and their performance. *IEEE Transactions on Communications*, 43(12):2959–2965, 1995. doi: 10.1109/26.477498.

Pablo A. Estevez, Michel Tesmer, Claudio A. Perez, and Jacek M. Zurada. Normalized mutual information feature selection. *IEEE Transactions on Neural Networks*, 20(2):189–201, 2009. doi: 10.1109/TNN.2008.2005601.

Junming Hou, Xiaoyu Chen, Ran Ran, Xiaofeng Cong, Xinyang Liu, Jian Wei You, and Liang-Jian Deng. Binarized neural network for multi-spectral image fusion. In *Proceedings of the 38th IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2236–2245, 2025. doi: 10.1109/CVPR52734.2025.00214.

Loshchilov Ilya and Hutter Frank. Decoupled weight decay regularization. In *Proceedings of the 7th International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=Bkg6RiCqY7`.

Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. Llvip: A visible-infrared paired dataset for low-light vision. In *Proceedings of the 34th IEEE/CVF International Conference on Computer Vision*, pp. 3496–3504, Montreal, BC, Canada, 2021. Institute of Electrical and Electronics Engineers. doi: 10.1109/ICCVW54120.2021.00389.

Xinrui Jiang, Nannan Wang, Jingwei Xin, Keyu Li, Xi Yang, and Xinbo Gao. Training binary neural network without batch normalization for image super-resolution. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, volume 35, pp. 1700–1707, 2021. doi: 10.1609/aaai.v35i2.16263.

Xinrui Jiang, Nannan Wang, Jingwei Xin, Keyu Li, Xi Yang, Jie Li, Xiaoyu Wang, and Xinbo Gao. Fabnet: Frequency-aware binarized network for single image super-resolution. *IEEE Transactions on Image Processing*, 32:6234–6247, 2023. doi: 10.1109/TIP.2023.3328565.

Yizhen Jiang, Meng-Yao Ma, Anqi Zhu, Xiaowen Ma, Jiaxin Li, and Wei Zhang. Spatial-spectral binarized neural network for panchromatic and multi-spectral images fusion. *arXiv*, abs/2509.23321, 2025.

Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements, 2024. URL https://api.semanticscholar.org/CorpusID:273532028.

Hui Li and Xiao-Jun Wu. Crossfuse: A novel cross attention mechanism based infrared and visible image fusion approach. *Information Fusion*, 103:102147, 2024. ISSN 1566-2535. doi: 10.1016/j.inffus.2023.102147.

Hui Li, Tianyang Xu, Xiao-Jun Wu, Jiwen Lu, and Josef 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. doi: 10.1109/TPAMI.2023.3268209.

Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *Proceedings of the 35th IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5802–5811, New Orleans, LA, USA, 2022. Institute of Electrical and Electronics Engineers. doi: 10.1109/CVPR52688.2022.00571.

Zechun Liu, Wenhan Luo, Baoyuan Wu, Xin Yang, Wei Liu, and Kwang-Ting Cheng. Bi-real net: Binarizing deep network towards real-network performance. *International Journal of Computer Vision*, 128(1):202–219, 2020a. doi: 10.1007/s11263-019-01227-8.

Zechun Liu, Zhiqiang Shen, Marios Savvides, and Kwang-Ting Cheng. Reactnet: Towards precise binary neural network with generalized activation functions. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Proceedings of the 16th European Conference on Computer Vision*, pp. 143–159, Cham, 2020b. Springer International Publishing. ISBN 978-3-030-58568-6.

Jiayi Ma, Wei Yu, Pengwei Liang, Chang Li, and Junjun Jiang. Fusiongan: A generative adversarial network for infrared and visible image fusion. *Information Fusion*, 48:11–26, 2019. ISSN 1566-2535. doi: 10.1016/j.inffus.2018.09.004.

Jiayi Ma, Han Xu, Junjun Jiang, Xiaoguang Mei, and Xiao-Ping Zhang. Ddcgan: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion. *IEEE Transactions on Image Processing*, 29:4980–4995, 2020. doi: 10.1109/TIP.2020.2977573.

Jiayi Ma, Linfeng Tang, Fan Fan, Jun Huang, Xiaoguang Mei, and Yong Ma. Swinfusion: Cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA Journal of Automatica Sinica*, 9(7):1200–1217, 2022. doi: 10.1109/JAS.2022.105686.

Seonghyun Park, An Gia Vien, and Chul Lee. Cross-modal transformers for infrared and visible image fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(2):770–785, 2024. doi: 10.1109/TCSVT.2023.3289170.

Haotong Qin, Ruihao Gong, Xianglong Liu, Mingzhu Shen, Ziran Wei, Fengwei Yu, and Jingkuan Song. Forward and backward information retention for accurate binary neural networks. In *Proceedings of the 33rd IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2247–2256, 2020. doi: 10.1109/CVPR42600.2020.00232.

Guihong Qu, Dali Zhang, and Pingfan Yan. Information measure for performance of image fusion. *Electronics Letters*, 38(7):313, January 2002. doi: 10.1049/el:20020212.

Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, and Ali Farhadi. Xnor-net: Imagenet classification using binary convolutional neural networks. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (eds.), *Proceedings of the 14th European Conference on Computer Vision*, pp. 525–542, Cham, 2016. ISBN 978-3-319-46493-0.

Wesley Roberts, Jan van Aardt, and Fethi Ahmed. Assessment of image fusion procedures using entropy, image quality, and multispectral classification. *Journal of Applied Remote Sensing*, 2: 1–28, 05 2008. doi: 10.1117/1.2945910.

Yu Shi, Yu Liu, Juan Cheng, Z. Jane Wang, and Xun Chen. Vdmufusion: A versatile diffusion model-based unsupervised framework for image fusion. *IEEE Transactions on Image Processing*, 34:441–454, 2025. doi: 10.1109/TIP.2024.3512365.

Yu Shi, Yu Liu, Zhong-Cheng Wu, Juan Cheng, Huafeng Li, and Xun Chen. Degradation-robust fusion: An efficient degradation-aware diffusion framework for multimodal image fusion in arbitrary degradation scenarios. In *Proceedings of the 39th IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33848–33858, June 2026.

Chongxing Song, Zhiqiang Lang, Wei Wei, and Lei Zhang. E2fif: Push the limit of binarized deep imagery super-resolution using end-to-end full-precision information flow. *IEEE Transactions on Image Processing*, 32:5379–5393, 2023. doi: 10.1109/TIP.2023.3315540.

Linfeng Tang, Jiteng Yuan, Hao Zhang, Xingyu Jiang, and Jiayi Ma. Piafusion: A progressive infrared and visible image fusion network based on illumination aware. *Information Fusion*, 83-84:79–92, 2022. ISSN 1566-2535. doi: 10.1016/j.inffus.2022.03.007.

Linfeng Tang, Hao Zhang, Han Xu, and Jiayi Ma. Deep learning-based image fusion: A survey. *Journal of Image and Graphics*, 28(1):3–36, 2023. doi: 10.11834/jig.220422.

Linfeng Tang, Yeda Wang, Zhanchuan Cai, Junjun Jiang, and Jiayi Ma. Controlfusion: A controllable image fusion network with language-vision degradation prompts. In *Proceedings of the 39th Advances in Neural Information Processing Systems*, 2025.

Tee-Ann Teo and Chi-Chung Lau. Pyramid-based image empirical mode decomposition for the fusion of multispectral and panchromatic images. *EURASIP Journal on Advances in Signal Processing*, 2012(1):4, 2012. ISSN 1687-6180. doi: 10.1186/1687-6180-2012-4.

Rui Tian, Zuxuan Wu, Qi Dai, Han Hu, Yu Qiao, and Yu-Gang Jiang. Resformer: Scaling vits with multi-resolution training. In *Proceedings of the 36th IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22721–22731, 2023. doi: 10.1109/CVPR52729.2023.02176.

Alexander Toet and Maarten Hogervorst. Progress in color night vision. *Optical Engineering*, 51, 02 2012. doi: 10.1117/1.OE.51.1.010901.

yolov11 ultralytics. `https://github.com/ultralytics/ultralytics`, 2024. Accessed: 2024-10-09.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6309–6318, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Xue Wang, Zheng Guan, Shishuang Yu, Jinde Cao, and Ya Li. Infrared and visible image fusion via decoupling network. *IEEE Transactions on Instrumentation and Measurement*, 71:1–13, 2022. doi: 10.1109/TIM.2022.3216413.

Zhishe Wang, Zhuoqun Zhang, Wuqiang Qi, Fengbao Yang, and Jiawei Xu. Freqgan: Infrared and visible image fusion via unified frequency adversarial learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(1):728–740, 2025. doi: 10.1109/TCSVT.2024.3460172.

Bin Xia, Yulun Zhang, Yitong Wang, Yapeng Tian, Wenming Yang, Timofte Radu, and Gool Luc Van. Basic binary convolution unit for binarized image restoration network. In *Proceedings of the 11th International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=h8T5dZWTZ-Z`.

Junfeng Yang, Xueyang Fu, Yuwen Hu, Yue Huang, Xinghao Ding, and John Paisley. Pannet: A deep network architecture for pan-sharpening. In *Proceedings of the 16th IEEE/CVF International Conference on Computer Vision*, pp. 1753–1761, 2017. doi: 10.1109/ICCV.2017.193.

Yi Zhai, Baoping Song, Jinyong Cheng, Aimei Dong, and Guohua Lv. Stafusion: An adversarial learning network for infrared and visible image fusion via swin transformer. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 10(1):191–205, 2026. doi: 10.1109/TETCI.2025.3572144.

Hao Zhang and Jiayi Ma. Sdnet: A versatile squeeze-and-decomposition network for real-time image fusion. *International Journal of Computer Vision*, 129:2761 – 2785, 2021. doi: 10.1007/s11263-021-01501-8.

Xingchen Zhang. Deep learning-based multi-focus image fusion: A survey and a comparative study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4819–4838, 2022. doi: 10.1109/TPAMI.2021.3078906.

Xuchong Zhang, Han Zhai, Jiaxing Liu, Zhiping Wang, and Hongbin Sun. Real-time infrared and visible image fusion network using adaptive pixel weighting strategy. *Information Fusion*, 99:101863, 2023. ISSN 1566-2535. doi: 10.1016/j.inffus.2023.101863.

Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Kai Zhang, Shuang Xu, Dongdong Chen, Radu Timofte, and Luc Van Gool. Equivariant multi-modality image fusion. In *Proceedings of the 37th IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25912–25921, Seattle, WA, USA, 2024. Institute of Electrical and Electronics Engineers. doi: 10.1109/CVPR52733.2024.02448.

Pengfei Zhu, Yang Sun, Bing Cao, and Qinghua Hu. Task-customized mixture of adapters for general image fusion. In *Proceedings of the 37th IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7099–7108, 2024. doi: 10.1109/CVPR52733.2024.00678.

A Derivation of the ISDRT Matrix

This section derives the fixed linear synthesis matrix used by ISDRT to map five processed SDRT components to a 2×2 spatial patch. The synthesis operator is defined by an unweighted Euclidean least-squares criterion.

A.1 Matrix Representation and Rank of SDRT

For a 2×2 feature patch, we use row-wise vectorization:

$$X_{i,j} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \qquad \mathbf{x} = \mathrm{vec}_{\mathrm{r}}(X_{i,j}) = [a \quad b \quad c \quad d]^{\mathsf{T}}. \tag{9}$$

SDRT computes a base component, defined as a normalized sum, and four directional difference components:

$$\begin{aligned} M &= \frac{a+b+c+d}{2}, \quad R^h = b-a, \quad R^v = c-a, \\ R^d &= d-a, \qquad\qquad R^a = c-b. \end{aligned} \tag{10}$$

Collecting these components into a vector gives

$$\mathbf{y} = \begin{bmatrix} M & R^h & R^v & R^d & R^a \end{bmatrix}^\mathsf{T} = \mathbf{A}\mathbf{x}, \tag{11}$$

where

$$\mathbf{A} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \\ 0 & -1 & 1 & 0 \end{bmatrix} \in \mathbb{R}^{5\times 4}. \tag{12}$$

The five-component representation is redundant because

$$R^a = R^v - R^h. \tag{13}$$

Nevertheless, $\mathbf{A}$ has full column rank. Indeed, $\mathbf{A}\mathbf{x} = \mathbf{0}$ implies $a = b = c = d$ from the directional differences, and the base component then implies that their common value is zero. Hence, $\mathrm{rank}(\mathbf{A}) = 4$ and $\mathbf{A}^\mathsf{T}\mathbf{A}$ is positive definite.

As a rectangular matrix, $\mathbf{A}$ has no two-sided inverse, but it admits left inverses. We adopt its Moore–Penrose pseudoinverse as the synthesis operator, corresponding to the least-squares criterion defined below.

A.2 Least-Squares Formulation

For a fixed channel and spatial position, we collect the processed components supplied to ISDRT as

$$\widehat{\mathbf{y}} = \begin{bmatrix} Y^M & Y^h & Y^v & Y^d & Y^a \end{bmatrix}^\mathsf{T} \in \mathbb{R}^5. \tag{14}$$

Here, Y^M, Y^h, Y^v, Y^d, and Y^a denote the processed base, horizontal, vertical, main-diagonal, and anti-diagonal components, respectively. The channel and spatial indices are omitted for clarity. These processed components need not satisfy the consistency relation in Eq. 13.

For a candidate patch vector $\mathbf{u} = [u_a, u_b, u_c, u_d]^\mathsf{T}$, define

$$\mathcal{J}(\mathbf{u}) = \|\mathbf{A}\mathbf{u} - \widehat{\mathbf{y}}\|_2^2 . \tag{15}$$

This criterion assigns equal weights to the five squared residuals under the component normalization in Eq. 10. ISDRT defines the reconstructed patch vector as

$$\widetilde{\mathbf{x}} = \operatorname*{arg\,min}_{\mathbf{u}\in\mathbb{R}^4} \mathcal{J}(\mathbf{u}). \tag{16}$$

Since $\mathbf{A}$ has full column rank, the least-squares problem has a unique solution. Setting $\nabla_\mathbf{u}\mathcal{J} = 2\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{u} - \widehat{\mathbf{y}})$ to zero yields the normal equations

$$\mathbf{A}^\mathsf{T}\mathbf{A}\widetilde{\mathbf{x}} = \mathbf{A}^\mathsf{T}\widehat{\mathbf{y}}. \tag{17}$$

The unique least-squares solution is therefore

$$\widetilde{\mathbf{x}} = (\mathbf{A}^\mathsf{T}\mathbf{A})^{-1}\mathbf{A}^\mathsf{T}\widehat{\mathbf{y}}. \tag{18}$$

A.3 Synthesis Matrix and Reconstruction Properties

Direct multiplication gives

$$\mathbf{A}^\mathsf{T}\mathbf{A} = \frac{1}{4}\begin{bmatrix} 13 & -3 & -3 & -3 \\ -3 & 9 & -3 & 1 \\ -3 & -3 & 9 & 1 \\ -3 & 1 & 1 & 5 \end{bmatrix}. \tag{19}$$

Its determinant is

$$\det(\mathbf{A}^\mathsf{T}\mathbf{A}) = 12, \tag{20}$$

and its inverse is

$$(\mathbf{A}^\mathsf{T}\mathbf{A})^{-1} = \frac{1}{48}\begin{bmatrix} 21 & 9 & 9 & 9 \\ 9 & 29 & 13 & -3 \\ 9 & 13 & 29 & -3 \\ 9 & -3 & -3 & 45 \end{bmatrix}. \tag{21}$$

The synthesis operator is the Moore–Penrose pseudoinverse of $\mathbf{A}$:

$$\mathbf{A}^+ = (\mathbf{A}^\mathsf{T}\mathbf{A})^{-1}\mathbf{A}^\mathsf{T} \in \mathbb{R}^{4\times 5}. \tag{22}$$

Substituting Eqs. 12 and 21 gives the ISDRT synthesis matrix

$$\begin{aligned} \mathbf{A}^+ &= \frac{1}{48}\begin{bmatrix} 21 & 9 & 9 & 9 \\ 9 & 29 & 13 & -3 \\ 9 & 13 & 29 & -3 \\ 9 & -3 & -3 & 45 \end{bmatrix}\begin{bmatrix} \frac{1}{2} & -1 & -1 & -1 & 0 \\ \frac{1}{2} & 1 & 0 & 0 & -1 \\ \frac{1}{2} & 0 & 1 & 0 & 1 \\ \frac{1}{2} & 0 & 0 & 1 & 0 \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{2} & -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & 0 \\ \frac{1}{2} & \frac{5}{12} & \frac{1}{12} & -\frac{1}{4} & -\frac{1}{3} \\ \frac{1}{2} & \frac{1}{12} & \frac{5}{12} & -\frac{1}{4} & \frac{1}{3} \\ \frac{1}{2} & -\frac{1}{4} & -\frac{1}{4} & \frac{3}{4} & 0 \end{bmatrix}. \end{aligned} \tag{23}$$

The reconstructed vector is

$$\widetilde{\mathbf{x}} = \mathbf{A}^+\widehat{\mathbf{y}} = \begin{bmatrix} \widetilde{a} & \widetilde{b} & \widetilde{c} & \widetilde{d} \end{bmatrix}^\mathsf{T}, \tag{24}$$

which is reshaped into the corresponding spatial patch:

$$\widetilde{X}_{i,j} = \begin{bmatrix} \widetilde{a} & \widetilde{b} \\ \widetilde{c} & \widetilde{d} \end{bmatrix}. \tag{25}$$

The derived matrix satisfies

$$\mathbf{A}^+\mathbf{A} = \mathbf{I}_4. \tag{26}$$

The identity $\mathbf{A}^+\mathbf{A} = \mathbf{I}_4$ confirms that $\mathbf{A}^+$ is a left inverse of the SDRT matrix. For the processed components $\widehat{\mathbf{y}}$, ISDRT constructs the spatial patch using the unique least-squares solution defined above.

B Ablation on SDRT

To examine the role of learnable processing in the base and directional branches, we compare two SDRT variants while retaining the complete LSA module and the remaining network architecture. ***M-pass*** replaces the binary block in the base branch of every SDRT block with an identity mapping, $Y^M = M$, while keeping the four learnable directional branches. ***R-pass*** replaces the binary blocks in all four directional branches with identity mappings, $Y^r = R^r$, while keeping the learnable base branch. Both variants retain the corresponding components and the original reconstruction operation. Each model is trained from scratch for 50 epochs on the same 2000 LLVIP training pairs. In Tab. 6, both variants give lower scores than Ours on all reported metrics. The lower scores in both variants support a role for learnable processing of the base and directional components within the complete framework.

Method	EN↑	MI↑	SF↑	EI↑	AG↑
M-pass	6.42	12.84	8.93	30.45	3.15
R-pass	6.54	13.08	7.40	30.47	3.02
Ours	**6.69**	**13.38**	**9.94**	**39.85**	**3.97**

Table 6: Ablation of directional relation encoding and learnable component processing in SDRT.