T^{3}FUSION: TEST-TIME TRAINING FOR TASK-GENERAL IMAGE FUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Image fusion combines complementary observations into a more informative representation for subsequent perception and analysis. However, existing deep fusion models often generalize poorly to unseen datasets and tasks because fusion rules learned offline may not capture the relationships between sources in new inputs. To address this limitation, we propose **T^{3}Fusion**, a test-time training framework that adapts fusion by learning cross-source relations from the current input pair during inference. Specifically, a fusion network extracts multiscale features from the two source images. We fit compact bidirectional mappings between these features through closed-form ridge regression so that each source predicts the other's features. A learned fusion readout combines these predictions and their residuals with the original features to refine the representations passed to the decoder for image reconstruction. The network weights remain fixed while the mappings are fitted to each test pair without fused ground truth or test-time backpropagation. Extensive experiments across five fusion tasks show that a single MSRS-trained model generalizes to unseen datasets and tasks without task-specific fine-tuning, achieving better fusion quality and downstream task performance.

1 INTRODUCTION

Image fusion combines complementary observations of the same scene into a single image with more complete and reliable information than any individual source (Zhang & Demiris, 2023). It is an important component of computational imaging and multimodal perception because a single observation cannot capture all relevant scene content under every sensing condition (Xu et al., 2022). By consolidating complementary evidence, fusion supports subsequent visual interpretation and decision-making (Liu et al., 2022a; Zhao et al., 2023a). Its core challenge is to preserve useful information from each source while suppressing source-specific degradation.

Recent deep fusion methods have improved the integration of complementary source information (Zhao et al., 2023c; Tang et al., 2025; Jiang et al., 2026; Xia et al., 2026), while unified frameworks further extend this capability across fusion tasks (Xu et al., 2022; Liu et al., 2026). When target-task data are limited and retraining is costly, transferring a model trained on an available task is valuable. However, fusion rules learned offline may generalize poorly to unseen datasets and tasks with different source contributions. The local degradation observed in the DPOFusion (Su et al., 2026) results in Fig. 1(a) illustrates this limitation and motivates learning from the current source pair to adapt fusion to its specific characteristics.

The aligned sources provide a natural learning signal because each can supervise prediction of the other's features without a fused target. Test-time training (TTT) offers a framework for exploiting such supervision at inference by learning an adapted state to guide prediction (Sun et al., 2020; Bartler et al., 2022; Sun et al., 2025), as depicted in Fig. 1(b). For image fusion, the learned state must ultimately guide the construction of the fused image. Cross-source prediction alone does not determine which information should be preserved because useful source-specific details may also be difficult to predict. The central challenge is therefore to connect learning from the current pair with a fusion process that can interpret and preserve complementary source information.

In this paper, we propose **T^{3}Fusion** as a test-time training framework for task-general image fusion. Specifically, a fusion network extracts source features that combine multiscale context with

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.

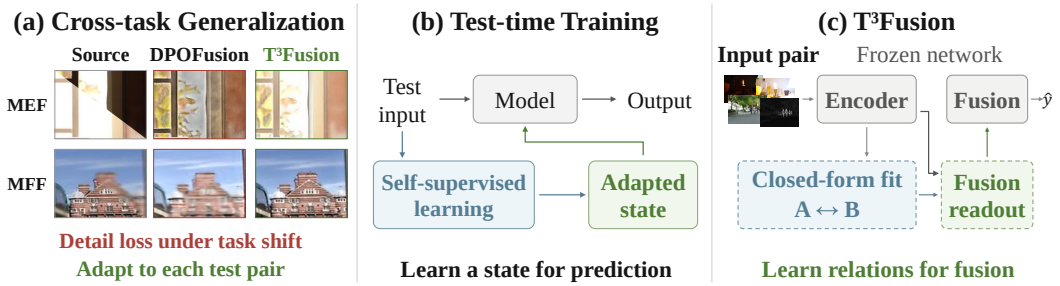

Figure 1: Cross-task fusion and test-time learning. (a) Qualitative comparisons between DPOFusion (Su et al., 2026) and T^{3}Fusion on multi-exposure fusion (MEF) and multi-focus fusion (MFF). (b) Test-time training learns an adapted state from test observations through self-supervision to guide prediction. (c) T^{3}Fusion connects closed-form cross-source fitting to a learned fusion readout.

full-resolution detail. We fit compact bidirectional mappings between these features through closed-form ridge regression so that each source predicts the other's features at held-out spatial locations. A learned fusion readout combines these predictions and their residuals with the original features to refine the representations used for image reconstruction. The fitted relations condition a fusion process learned offline, while the original features remain available to preserve complementary source information. The network weights are learned offline and remain fixed at test time, when only the mapping matrices are fitted without fused ground truth or test-time backpropagation. Our contributions are summarized as follows:

- We formulate test-time learning for image fusion through current-pair cross-source prediction, providing supervision for adapting fusion representations without fused ground truth.
- We develop T^{3}Fusion to couple cross-source prediction with a learned fusion readout, enabling pair-specific feature updates while retaining access to the original source information.
- Extensive experiments show that T^{3}Fusion generalizes to unseen infrared-visible datasets and four additional fusion tasks, achieving better fusion quality and downstream task performance.

2 Related Work

Image fusion and cross-task generalization. Deep image fusion combines complementary observations through learned representations and fusion rules (Li & Wu, 2019; Zhao et al., 2023b;c). Unified frameworks extend fusion across tasks through shared learning, task interaction, and adaptive modules (Xu et al., 2022; Zhu et al., 2024; Cheng et al., 2025; Hu et al., 2025). Recent methods further explore robust source-aware representations (Li et al., 2026) and signal-level self-supervision (Wang et al., 2026b). DECC (Liu et al., 2026) demonstrates zero-shot transfer to other fusion tasks after infrared and visible image fusion training. These advances establish the potential of generalizable offline representations. T^{3}Fusion complements this direction by learning from individual source pairs at test time.

Test-time training and closed-form solvers. Test-time training adapts models through self-supervision on unlabeled observations (Sun et al., 2020). Subsequent work explores masked reconstruction (Gandelsman et al., 2022) and meta-learned initialization (Bartler et al., 2022). TTT layers (Sun et al., 2025) incorporate this learning process into the architecture by representing the hidden state as a learner, and ViT^3 (Han et al., 2026) extends this principle to vision. Differentiable closed-form solvers enable learning through ridge regression (Bertinetto et al., 2019), while test-time regression provides a unifying view of associative memory (Wang et al., 2026a). T^{3}Fusion applies these learning principles to the construction of a fused image from paired observations.

Test-time adaptation for image fusion. Zheng et al. (Zheng et al., 2024a) adapt a fusion model through quality-aware self-supervision with a fusion adapter. TTTFusion (Xie & Tang, 2025) conditions normalization and fusion on source-feature statistics, while TTD (Cao et al., 2024) derives dynamic fusion weights from relative dominability. These approaches use the test observations to adjust model parameters or determine how source information is combined. T^{3}Fusion differs by formulating adaptation as explicit cross-source prediction within each input pair.

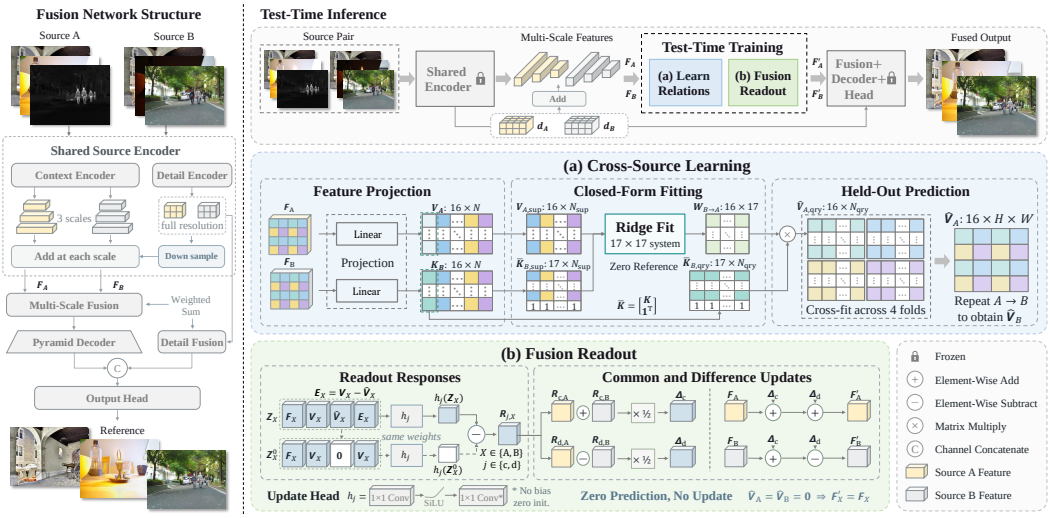

Figure 2: Overview of T^{3}Fusion. Cross-source learning fits bidirectional relations to the current pair, and fusion readout converts their predictions into feature updates for image reconstruction. Persistent network parameters remain fixed at test time.

3 METHOD

3.1 FUSION FRAMEWORK

Problem formulation. Given aligned luminance images $\boldsymbol{I}_A, \boldsymbol{I}_B \in [0,1]^{H\times W}$, image fusion aims to produce an output $\boldsymbol{Y}$ that preserves useful information from both sources while suppressing source-specific degradation. The contribution of each source depends on the observed pair, but no fused target is available to directly supervise their combination.

Spatial alignment nevertheless provides feature correspondences for bidirectional prediction. Each source can supervise prediction of the other's features, providing a learning signal for estimating cross-source relations from the current pair. We introduce a temporary relation state $\mathcal{W}$ and formulate fusion as

$$\mathcal{W}^\star = \arg\min_{\mathcal{W}} \mathcal{L}_{\text{cs}}(\mathcal{W}; \boldsymbol{I}_A, \boldsymbol{I}_B, \boldsymbol{\theta}), \quad \boldsymbol{Y} = \mathcal{F}_{\boldsymbol{\theta}}(\boldsymbol{I}_A, \boldsymbol{I}_B; \mathcal{W}^\star), \tag{1}$$

where $\mathcal{F}_{\boldsymbol{\theta}}$ denotes the complete fusion model and $\mathcal{L}_{\text{cs}}$ is a regularized cross-source prediction objective. T^{3}Fusion represents $\mathcal{W}$ through compact bidirectional mappings fitted by closed-form ridge regression. The persistent parameters $\boldsymbol{\theta}$ are learned offline and remain fixed during inference.

The fitted relations describe cross-source predictability, while useful source-specific details may remain difficult to predict. We therefore provide predictions and residuals alongside the original features to a learned fusion readout. Offline fusion supervision trains the readout to interpret these quantities in their feature context. At test time, fitting updates the relation state within this learned fusion process. This separates estimating how the current sources are related from learning how their relations should influence fusion.

Network architecture. Fig. 2 presents the framework. A shared encoder combines a Swin context pyramid (Liu et al., 2022b) with a full-resolution CNN detail branch. For each source $X \in \{A, B\}$, it produces context features $\mathbf{F}^s_X$ at three scales with widths $(32, 64, 128)$ and a detail tensor $\mathbf{d}_X$. Pooled detail features are projected and added to the context features before channel normalization.

At each scale, a shared scorer receives the source feature, pair mean, and absolute source difference. A softmax across sources produces channel- and position-dependent fusion weights

$$\mathbf{F}^s_{\text{f}} = \mathbf{W}^s_A \odot \mathbf{F}^s_A + \mathbf{W}^s_B \odot \mathbf{F}^s_B, \qquad \mathbf{W}^s_A + \mathbf{W}^s_B = \mathbf{1}. \tag{2}$$

The same selection rule fuses the detail tensors. A top-down decoder with lateral connections (Lin et al., 2017) combines the context and detail paths to produce the fused image. Shared source processing and symmetric fusion preserve the output under source swapping. The cross-source learning and readout described next refine the context features before fusion. The output before these updates is denoted by $\boldsymbol{Y}_0$.

3.2 CROSS-SOURCE LEARNING

We estimate the relation state by predicting features of one source from the other within each input pair. Corresponding spatial locations provide supervision without a fused target. Fig. 2(a) shows how feature projection, relation fitting, and held-out prediction turn this supervision into cross-source predictions for the fusion readout.

Feature projection. At scale s, projections $\boldsymbol{P}_k^s, \boldsymbol{P}_v^s \in \mathbb{R}^{r \times C_s}$ produce compact keys and values with $r = 16$. Keys serve as regression inputs and values as prediction targets. Keys are derived from the fusion network features and values from a frozen teacher copy. Projection rows are normalized, and both representations are standardized using a shared mean $\boldsymbol{\mu}_s$ and diagonal standard deviation $\boldsymbol{\Sigma}_s$ calibrated once on training data. Teacher features are detached, while both projections are learned offline. A constant one is appended to each key $\boldsymbol{k}_X^s(p)$ to obtain $\bar{\boldsymbol{k}}_X^s(p)$ for fitting an intercept.

Closed-form fitting. To predict tokens excluded from fitting, we partition the spatial grid into four interleaved folds using $g(i, j) = 2(i \bmod 2) + (j \bmod 2)$. For each query fold, a support set is drawn from the other three folds. Support sets contain up to 1,024 deterministically sampled valid tokens, excluding padding. The target values of the query fold do not participate in fitting.

For $B \to A$, support keys from source B are paired with values from source A at corresponding positions. Suppressing scale, fold, and direction indices, let $\bar{\boldsymbol{K}} \in \mathbb{R}^{(r+1)\times n}$ and $\boldsymbol{V} \in \mathbb{R}^{r \times n}$ collect these representations. The ridge objective and its solution are

$$\begin{aligned} \boldsymbol{W}^* &= \arg\min_{\boldsymbol{W}} \left[\frac{1}{n} \|\boldsymbol{W}\bar{\boldsymbol{K}} - \boldsymbol{V}\|_{\mathrm{F}}^2 + \rho \|\boldsymbol{W}\|_{\mathrm{F}}^2 \right], \\ &= \boldsymbol{V}\bar{\boldsymbol{K}}^\top \left(\bar{\boldsymbol{K}}\bar{\boldsymbol{K}}^\top + n\rho \boldsymbol{I}_{r+1} \right)^{-1}. \end{aligned} \tag{3}$$

The positive coefficient ρ is learned offline and regularizes all coefficients, including the intercept. The solution depends on the key and cross-source second-order statistics of the support set, allowing a fixed prediction space to capture relations specific to each pair. We evaluate Eq. (3) using a linear solve without iterative gradient updates. The reverse direction is fitted after exchanging the sources.

Held-out prediction. Each token is predicted by the mapping fitted without its fold

$$\widehat{\boldsymbol{v}}_A^s(p) = \boldsymbol{W}_{B\to A, g(p)}^{s,*} \bar{\boldsymbol{k}}_B^s(p), \qquad \widehat{\boldsymbol{v}}_B^s(p) = \boldsymbol{W}_{A\to B, g(p)}^{s,*} \bar{\boldsymbol{k}}_A^s(p). \tag{4}$$

Restoring spatial positions assembles the prediction maps $\widehat{\mathbf{V}}_A^s$ and $\widehat{\mathbf{V}}_B^s$. Support and query tokens are disjoint, although their encoded features may share spatial context. Three scales, four folds, and two prediction directions require 24 linear solves per pair.

3.3 FUSION READOUT

Cross-source prediction captures relations within the current pair, but useful source-specific details may also be difficult to predict. The readout in Fig. 2(b) therefore combines predictions, residuals, and original features to translate the learned relations into feature updates for fusion.

Readout responses. Let $\mathbf{V}_X^s$ denote the spatial value tensor and $\mathbf{E}_X^s = \mathbf{V}_X^s - \widehat{\mathbf{V}}_X^s$ its prediction residual. The head input $\mathbf{Z}_X^s$ concatenates the fusion network feature, target value, prediction, and residual. To center the response at zero prediction, we form a reference input $\mathbf{Z}_X^{s,0}$ by replacing the prediction with zero and the residual with the target value. For $j \in \{\mathrm{c}, \mathrm{d}\}$, the readout response is

$$\mathbf{R}_{j,X}^s = h_j^s(\mathbf{Z}_X^s) - h_j^s(\mathbf{Z}_X^{s,0}). \tag{5}$$

Each head uses a 1×1 convolution, SiLU, and a bias-free 1×1 output convolution. Heads share parameters across sources, and their output convolutions are zero-initialized. The residual explicitly parameterizes the discrepancy between values and predictions without adding an independent observation. Through the nonlinear head, the original features and values condition the response to cross-source predictions. Reference subtraction centers this response at zero prediction while retaining that feature context.

Common and difference updates. Common and difference refer to the source-exchange symmetry of the updates. Separate heads learn a shared correction applied equally to both sources and a

rebalancing correction applied with opposite signs

$$\begin{aligned} \mathbf{\Delta}^s_{\mathrm{c}} &= \tfrac{1}{2}(\mathbf{R}^s_{\mathrm{c},A} + \mathbf{R}^s_{\mathrm{c},B}), \quad \mathbf{\Delta}^s_{\mathrm{d}} = \tfrac{1}{2}(\mathbf{R}^s_{\mathrm{d},A} - \mathbf{R}^s_{\mathrm{d},B}), \\ \mathbf{F}'^{,s}_A &= \mathbf{F}^s_A + \mathbf{\Delta}^s_{\mathrm{c}} + \mathbf{\Delta}^s_{\mathrm{d}}, \quad \mathbf{F}'^{,s}_B = \mathbf{F}^s_B + \mathbf{\Delta}^s_{\mathrm{c}} - \mathbf{\Delta}^s_{\mathrm{d}}. \end{aligned} \tag{6}$$

The fusion module recomputes its selection weights from the refined features, and the decoder produces $\boldsymbol{Y}$ using the original detail tensors. Source swapping preserves the common update, reverses the difference update, and exchanges the refined features. The fused output therefore remains invariant to source order. If both directional predictions vanish at every scale, the actual and reference head inputs coincide, all updates vanish, and $\boldsymbol{Y} = \boldsymbol{Y}_0$.

3.4 Training Objectives

Offline training first learns a fusion prior and then learns how cross-source predictions should refine it. We train the fusion network before learning the projections and readout with that network fixed.

Fusion network training. We first pretrain the encoder and decoder with the reconstruction loss $\mathcal{L}_{\mathrm{rec}} = \frac{1}{2}\sum_{X\in\{A,B\}} \langle |R_{\boldsymbol{\theta}}(\boldsymbol{I}_X) - \boldsymbol{I}_X| \rangle$, where $R_{\boldsymbol{\theta}}$ denotes source reconstruction and $\langle\cdot\rangle$ denotes the spatial mean. To preserve the stronger source gradient at each position, we define

$$\mathcal{L}_{\mathrm{grad}}(\boldsymbol{Y}; \boldsymbol{I}_A, \boldsymbol{I}_B) = \sum_{d\in\{x,y\}} \left\langle \left| |\boldsymbol{S}_d * \boldsymbol{Y}| - \max(|\boldsymbol{S}_d * \boldsymbol{I}_A|, |\boldsymbol{S}_d * \boldsymbol{I}_B|) \right| \right\rangle, \tag{7}$$

where $\boldsymbol{S}_x, \boldsymbol{S}_y$ are fixed Sobel kernels and $*$ denotes convolution. The fusion network is then trained with $\mathcal{L}_{\mathrm{fusion}} = \mathcal{L}_{\mathrm{grad}}(\boldsymbol{Y}_0; \boldsymbol{I}_A, \boldsymbol{I}_B) + \lambda_{\mathrm{rec}}\mathcal{L}_{\mathrm{rec}}$.

Projection and readout training. With the fusion network fixed, we learn the projections, readout heads, and ridge coefficient by backpropagating through the linear solve and fused output. Cross-source prediction is supervised on held-out tokens through

$$\mathcal{L}_{\mathrm{q}} = \frac{1}{6}\sum_{s=1}^{3} \sum_{X\in\{A,B\}} \frac{1}{r|\mathcal{U}_s|} \sum_{p\in\mathcal{U}_s} \|\widehat{\boldsymbol{v}}^s_X(p) - \boldsymbol{v}^s_X(p)\|_2^2, \tag{8}$$

where $\mathcal{U}_s$ contains the valid query tokens at scale s. The loss averages over channels and tokens, with equal weight across scales and directions.

We reuse the gradient loss at coarse resolutions to preserve larger structures. Let $B_0(\boldsymbol{I}) = \boldsymbol{I}$ and $B_{\ell+1}(\boldsymbol{I}) = \mathrm{AvgPool}_2(\boldsymbol{G} * B_\ell(\boldsymbol{I}))$, where $\boldsymbol{G}$ is a 5×5 Gaussian kernel with unit standard deviation. The coarse loss is $\mathcal{L}_{\mathrm{c}}(\boldsymbol{Y}) = \frac{1}{3}\sum_{\ell=1}^{3} \mathcal{L}^{\mathrm{valid}}_{\mathrm{grad}}(B_\ell(\boldsymbol{Y}); B_\ell(\boldsymbol{I}_A), B_\ell(\boldsymbol{I}_B))$, where the superscript valid excludes the one-pixel Sobel boundary from the spatial mean. The final objective is

$$\mathcal{L}_{\mathrm{TTT}} = \mathcal{L}_{\mathrm{grad}}(\boldsymbol{Y}; \boldsymbol{I}_A, \boldsymbol{I}_B) + \lambda_{\mathrm{q}}\mathcal{L}_{\mathrm{q}} + \lambda_{\mathrm{c}}\mathcal{L}_{\mathrm{c}}(\boldsymbol{Y}) + s_{\mathrm{q}} + s_{\mathrm{c}}. \tag{9}$$

Here $\lambda_t = \lambda_t^0 \exp(-s_t)$ for $t \in \{\mathrm{q}, \mathrm{c}\}$. The learned log-scales adjust the positive auxiliary weights, while their additive terms regularize this adjustment.

4 Experiments

4.1 Settings and Details

Datasets and metrics. We train T^{3}Fusion on MSRS (Tang et al., 2022) and evaluate infrared-visible image fusion (IVIF) on its test split, FMB (Liu et al., 2023), M3FD (Liu et al., 2022a), and RoadScene (Xu et al., 2022). To assess cross-task generalization, we further evaluate the same model on MEFB (Zhang, 2021) for multi-exposure fusion (MEF), MFI-WHU (Zhang et al., 2021) for multi-focus fusion (MFF), PET–MRI pairs from the Harvard Whole Brain Atlas (Johnson & Becker, 1995) for medical image fusion (MIF), and Polarization (Morimatsu et al., 2020) for multi-polarization fusion (MPF). We report VIF, Q_{abf}, and Q_{cb} for fusion quality, FMB mIoU for semantic segmentation, and M3FD mAP_{50} and $\mathrm{mAP}_{50:95}$ for object detection.

Implementation details. We implement T^{3}Fusion in PyTorch and train it with AdamW using paired 256×256 crops and a weight decay of 10^{-4}. The fusion network undergoes 50 epochs of reconstruction pretraining and 150 epochs of fusion training, both with batch size 16 and an initial learning rate

Table 1: IVIF comparison. Higher is better; **bold** and underlined values denote the best and second-best results throughout the tables.

Method	MSRS			FMB			M3FD			RoadScene		
	VIF↑	Q_{abf} ↑	Q_{cb} ↑	VIF↑	Q_{abf} ↑	Q_{cb} ↑	VIF↑	Q_{abf} ↑	Q_{cb} ↑	VIF↑	Q_{abf} ↑	Q_{cb} ↑
TarDAL CVPR'22	0.716	0.416	0.274	0.807	0.462	0.403	0.708	0.427	0.414	0.627	0.434	0.427
CDDFuse CVPR'23	1.084	0.676	0.574	1.061	0.672	0.465	0.917	0.610	0.462	0.713	0.482	0.490
SHIP CVPR'24	0.968	0.658	0.537	1.040	0.697	0.491	0.915	0.643	0.491	0.729	0.541	0.471
CAF IJCAI'24	0.566	0.419	0.513	0.824	0.552	0.446	0.747	0.512	0.462	0.690	0.509	0.478
A2RNet AAAI'25	0.775	0.453	0.510	0.718	0.398	0.416	0.613	0.336	0.421	0.625	0.359	0.467
SAGE CVPR'25	0.762	0.537	0.495	0.969	0.637	0.448	0.856	0.583	0.455	0.641	0.383	0.442
HCLFuse NeurIPS'25	0.880	0.435	0.480	0.931	0.496	0.442	0.802	0.465	0.452	0.593	0.391	0.475
DPOFusion CVPR'26	0.716	0.296	0.471	0.695	0.365	0.422	0.611	0.338	0.436	0.557	0.330	0.517
ReCoFuse CVPR'26	0.879	0.442	0.452	0.815	0.460	0.436	0.727	0.452	0.448	0.706	0.444	0.513
T^{3}Fusion Ours	**1.135**	**0.726**	**0.606**	**1.106**	**0.718**	**0.506**	**0.994**	**0.681**	**0.519**	**0.820**	**0.606**	**0.523**

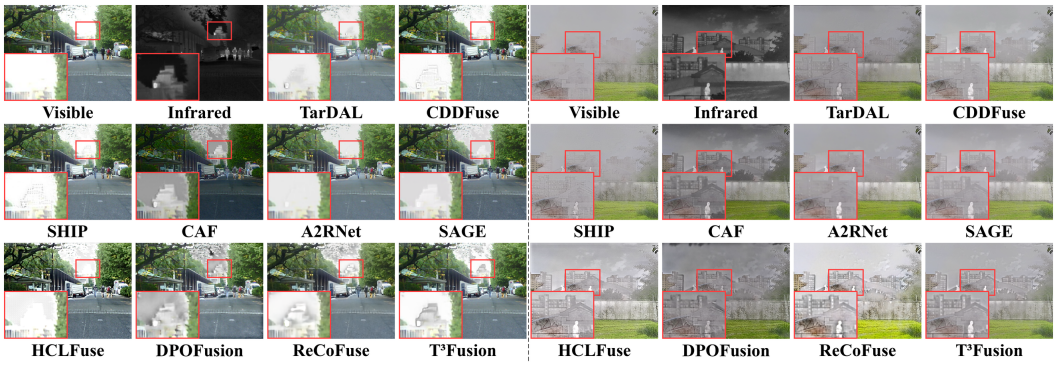

Figure 3: Visual comparisons on MSRS (left) and FMB (right). Insets highlight local details.

of 10^{-4}. We then freeze the fusion network and train the adaptation parameters for 50 epochs with batch size 1 and an initial learning rate of 5×10^{-5}. We use cosine decay during pretraining and adaptation, and a constant rate during fusion training. All main evaluations use the same checkpoint, with one round of closed-form fitting per test pair and no test-time gradient updates.

Compared methods. We compare against TarDAL (Liu et al., 2022a), CDDFuse (Zhao et al., 2023b), SHIP (Zheng et al., 2024b), CAF (Liu et al., 2024), A2RNet (Li et al., 2025), SAGE (Wu et al., 2025), HCLFuse (Guo et al., 2025), DPOFusion (Su et al., 2026), and ReCoFuse (Zhang et al., 2026). We use released implementations and checkpoints.

4.2 Infrared-Visible Fusion

In Tab. 1, T^{3}Fusion achieves the highest VIF, Q_{abf}, and Q_{cb} on MSRS and the unseen FMB, M3FD, and RoadScene datasets without target-dataset fine-tuning. On RoadScene, its VIF and Q_{abf} reach 0.820 and 0.606, compared with the next-best scores of 0.729 and 0.541, respectively. These results support effective preservation of source information and structural details across different imaging conditions and scene distributions. Fig. 3 shows that T^{3}Fusion preserves infrared building contours alongside visible foliage and street textures in MSRS, and pedestrian outlines together with surrounding window and wall structures in FMB. The enlarged regions illustrate the integration of salient infrared structures and complementary visible details while retaining surrounding details.

4.3 Cross-Task Generalization

We evaluate the same MSRS-trained model on MEF, MFF, MIF, and MPF without target-task fine-tuning, using identical cross-source learning and fusion readout across tasks. In Tab. 2, T^{3}Fusion achieves the highest Q_{abf} on all four datasets and the highest VIF on MEFB, MFI-WHU, and PET-MRI. Its Q_{abf} scores on MEFB and MFI-WHU are 0.733 and 0.682, exceeding the next-best results of 0.669 and 0.643, respectively. These results support structural preservation across unseen fusion tasks. Fig. 4 shows preserved window boundaries and calendar details across exposure dif-

Table 2: Zero-shot transfer to MEF, MFF, MIF, and MPF without target-task fine-tuning.

Method	MEFB (MEF)			MFI-WHU (MFF)			PET-MRI (MIF)			Polarization (MPF)		
	VIF↑	Q_{abf} ↑	Q_{cb} ↑	VIF↑	Q_{abf} ↑	Q_{cb} ↑	VIF↑	Q_{abf} ↑	Q_{cb} ↑	VIF↑	Q_{abf} ↑	Q_{cb} ↑
TarDAL CVPR'22	0.754	0.276	0.323	0.987	0.479	0.611	0.307	0.283	0.356	0.444	0.208	0.273
CDDFuse CVPR'23	1.347	0.637	0.418	1.127	0.620	0.642	0.624	0.698	0.522	**1.222**	0.599	0.349
SHIP CVPR'24	1.304	0.669	0.380	1.063	0.643	0.676	0.608	0.698	0.329	1.023	0.620	0.318
CAF IJCAI'24	1.132	0.566	0.338	0.988	0.577	0.630	0.524	0.629	**0.550**	0.576	0.343	0.385
A2RNet AAAI'25	0.924	0.366	0.312	0.981	0.418	0.613	0.483	0.406	0.335	0.891	0.438	0.307
SAGE CVPR'25	0.865	0.392	0.262	1.038	0.562	0.616	0.481	0.464	0.289	0.874	0.518	0.347
HCLFuse NeurIPS'25	1.198	0.425	0.403	0.943	0.393	0.452	0.478	0.539	0.545	0.980	0.322	**0.394**
DPOFusion CVPR'26	0.903	0.389	**0.455**	0.810	0.369	0.519	0.416	0.420	0.282	0.850	0.267	0.332
ReCoFuse CVPR'26	0.993	0.471	0.381	0.909	0.488	0.570	0.592	0.653	0.346	0.979	0.392	0.292
T^{3}Fusion Ours	**1.368**	**0.733**	0.413	**1.187**	**0.682**	**0.686**	**0.798**	**0.764**	0.439	1.140	**0.682**	0.357

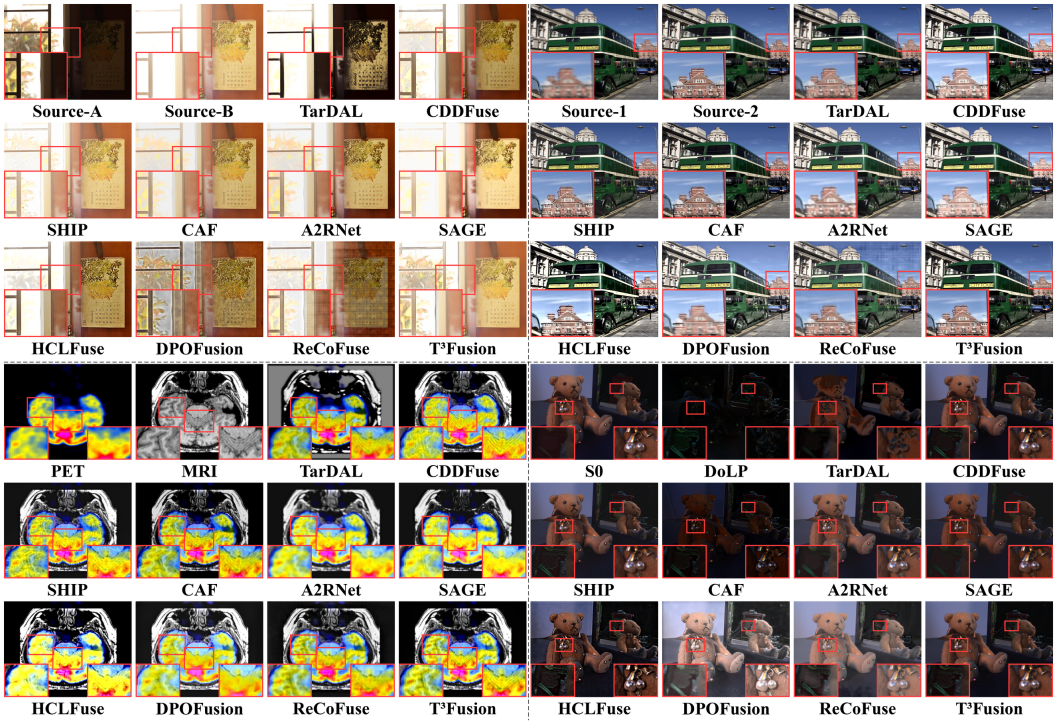

Figure 4: Zero-shot visual comparisons on MEF, MFF, MIF, and MPF using the same MSRS-trained model. Insets highlight local details.

ferences in MEF, and focused foreground vehicles together with distant building structures in MFF. PET–MRI results retain PET color distributions and MRI anatomical boundaries, while MPF results preserve the bear's appearance and polarization-dependent edges around the reflective pendant and background objects.

4.4 Downstream Perception

We evaluate FMB semantic segmentation with SegFormer-B1 (Xie et al., 2021) and M3FD object detection with YOLOv8s (Jocher et al., 2023), averaging results over fixed IR- and VI-trained evaluators. In Tab. 3, T^{3}Fusion leads the compared methods with 50.05% mIoU, 59.02% mAP_{50}, and 37.57% $mAP_{50:95}$. It also achieves the highest traffic-light and truck IoU and car, motorcycle, and lamp AP_{50}, although performance varies across categories. Fig. 5 complements these scores with preserved vehicle boundaries and small traffic-signal regions in FMB segmentation, alongside detections of the foreground truck, roadside pedestrian, and distant cars in M3FD. Together, these results show that the fused images retain useful information for dense semantic prediction and object localization in road scenes.

Table 3: Downstream comparison (%), averaged over fixed IR- and VI-trained evaluators.

Method	FMB Segmentation							M3FD Detection							
	Side.	Light	Sign	Car	Truck	Bus	mIoU ↑	Per.	Car	Bus	Motor.	Lamp	Truck	mAP_{50} ↑	$mAP_{50:95}$ ↑
TarDAL CVPR'22	32.87	17.88	42.45	**73.71**	11.23	14.08	48.09	**65.28**	78.08	67.29	44.03	38.99	46.50	56.69	35.67
CDDFuse CVPR'23	**36.32**	20.72	41.25	72.47	11.86	14.90	49.36	60.03	80.02	64.55	45.37	44.05	54.46	58.08	36.68
SHIP CVPR'24	34.94	20.54	40.52	72.42	9.38	9.83	47.65	57.46	76.15	54.82	41.06	45.02	46.96	53.58	33.29
CAF IJCAI'24	34.00	16.72	42.42	72.80	14.40	14.43	48.47	60.43	80.66	68.73	43.76	39.09	**55.69**	58.06	36.78
A2RNet AAAI'25	29.88	17.92	41.01	69.98	10.30	14.33	47.09	63.95	77.07	**72.85**	46.68	37.89	51.51	58.32	36.85
SAGE CVPR'25	32.91	20.17	**42.51**	72.38	14.82	19.18	49.51	59.85	79.24	70.94	44.37	44.08	54.59	58.84	37.34
HCLFuse NeurIPS'25	34.94	20.78	39.74	72.16	14.03	12.48	47.65	56.75	76.12	48.42	39.71	40.09	46.17	51.21	32.45
DPOFusion CVPR'26	32.77	17.80	38.08	67.00	10.37	13.52	43.82	52.20	74.14	53.71	41.53	40.85	46.42	51.47	31.84
ReCoFuse CVPR'26	35.24	19.57	37.89	68.66	10.62	**28.74**	47.65	57.97	77.16	57.48	42.40	41.65	51.19	54.64	34.41
T^3Fusion Ours	35.27	**22.94**	41.67	72.52	**16.22**	15.14	**50.05**	62.58	**80.90**	61.19	**47.24**	**47.36**	54.82	**59.02**	**37.57**

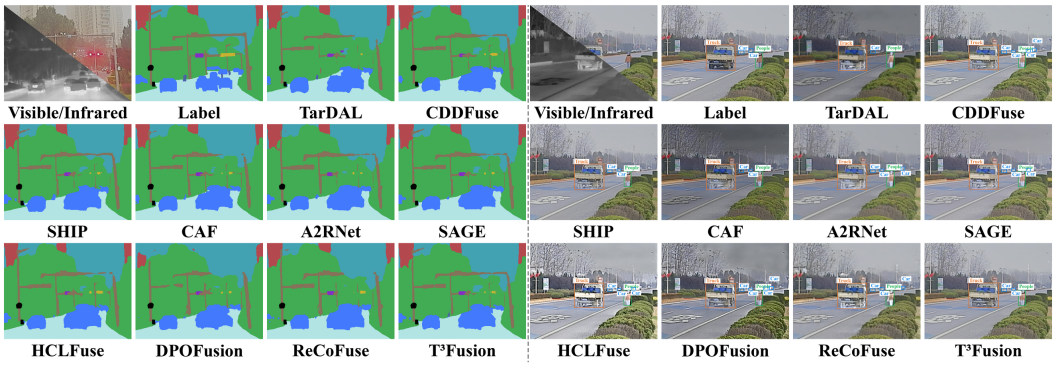

Figure 5: Visual comparisons of FMB semantic segmentation and M3FD object detection.

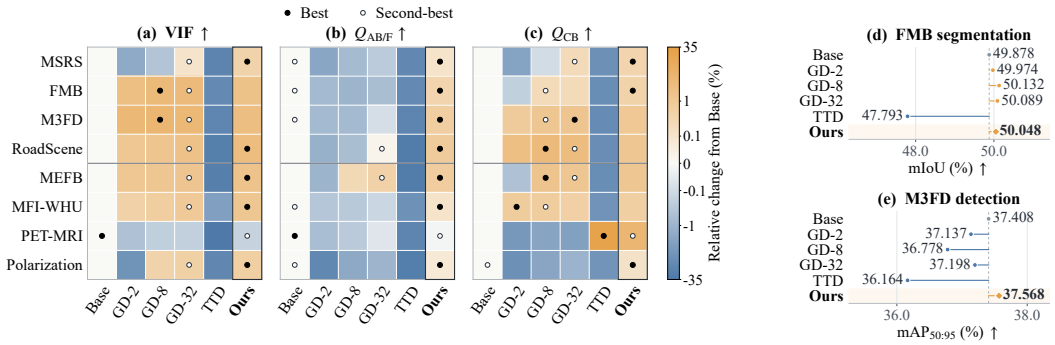

Figure 6: Comparison of gradient descent (GD), TTD (Cao et al., 2024), and our closed-form fitting.

4.5 Ablation Study

Effects of the TTT strategy. Fig. 6 compares closed-form fitting with 2, 8, and 32 gradient-descent steps on the same ridge objective. All variants retain the checkpoint trained with closed-form fitting and its fixed projections and readout. Closed-form TTT achieves higher Q_{abf} on every dataset and 37.57% detection $mAP_{50:95}$ versus 36.78–37.20% for GD, while segmentation scores remain close. These results favor direct relation fitting within the trained framework without selecting an iteration budget. They measure the downstream effect of solver choice, with fusion quality jointly determined by the relations and learned readout. The official complete TTD method (Cao et al., 2024) serves as a separate external reference.

Effects of current-pair relations. Table 4 replaces all 24 mappings with those from another pair in the same dataset, keeping the checkpoint, images, projections, and thus prediction targets fixed. Compared with the mean of three shuffle trials, matched relations reduce held-out prediction MSE by 5.64–54.74% and improve all three fusion metrics across eight datasets. In this fixed prediction space, lower MSE indicates more relevant relations, while fusion gains show that the readout uses them beyond auxiliary prediction. These consistent gains support pair-specific refinement with the

Table 4: Prediction-error reduction and fusion gains from current-pair versus shuffled relations. Fusion gains are in 10^{-3}.

Dataset	MSE red.(%)↑	ΔVIF ↑	ΔQ_{abf} ↑	ΔQ_{cb} ↑
MSRS	24.69	+1.882	+1.742	+0.529
FMB	33.77	+4.561	+0.571	+0.774
M3FD	39.96	+2.728	+1.184	+2.102
RoadScene	33.13	+3.847	+1.678	+0.635
MEFB	54.74	+1.350	+1.100	+0.230
MFI-WHU	31.51	+0.732	+0.696	+0.656
PET-MRI	5.64	+0.141	+0.092	+0.096
Polarization	32.00	+0.878	+1.012	+1.863

Table 5: Ablation study of relation estimation and fusion readout. Bold marks the best unrounded score, even when displayed values are identical.

Method	M3FD			MEFB		
	VIF↑	Q_{abf} ↑	Q_{cb} ↑	VIF↑	Q_{abf} ↑	Q_{cb} ↑
Static relations	0.989	0.680	0.517	**1.369**	0.731	0.411
Generator	0.990	0.680	0.517	1.368	0.730	0.409
Shared reader	0.986	0.677	0.516	1.358	0.732	0.413
T^{3}Fusion	**0.994**	**0.681**	**0.519**	1.368	**0.733**	**0.413**
Method	PET-MRI			MFI-WHU		
	VIF↑	Q_{abf} ↑	Q_{cb} ↑	VIF↑	Q_{abf} ↑	Q_{cb} ↑
Static relations	**0.798**	0.763	0.402	1.184	0.681	0.686
Generator	0.797	0.764	0.409	**1.188**	**0.682**	0.685
Shared reader	0.797	0.759	0.405	1.185	0.679	0.678
T^{3}Fusion	0.798	**0.764**	**0.439**	1.187	0.682	**0.686**

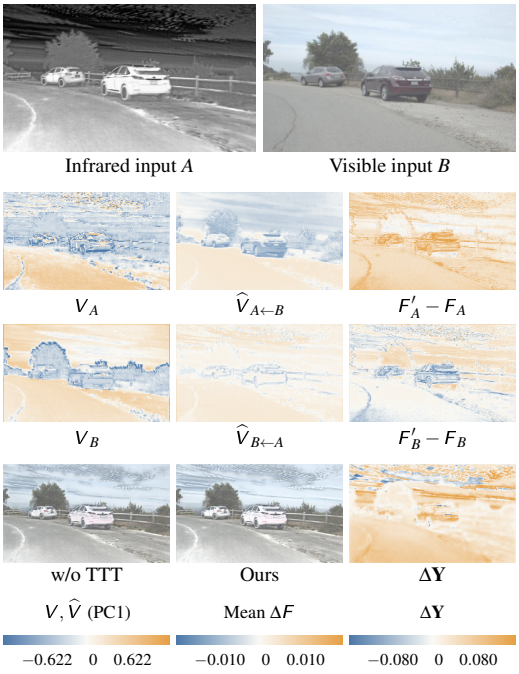

Figure 7: Visualization of cross-source predictions, feature updates, and output differences.

same readout across unseen tasks. Both conditions retain the original features and fusion prior, so the comparison measures the additional benefit of pair matching.

Effects of relation estimation. Table 5 compares alternatives trained on MSRS with the same frozen fusion network and 50-epoch adaptation protocol. Static relations learn fixed mappings offline. Generator predicts mappings from the support-set key and cross-source statistics used by ridge regression, with a narrower readout to match the overall adaptation budget. These controls test offline relation learning and direct input conditioning. T^{3}Fusion exceeds Static relations in Q_{abf} and Generator in Q_{cb} on all four datasets. On PET–MRI, the latter rises from 0.409 to 0.439. These results support explicit regression within the complete design, with benefits varying by dataset and metric.

Effects of fusion readout. Shared reader retains current-pair fitting but updates each source directly through one source-shared head, widened to match the two-head parameter count. Separate heads independently learn shared and opposite corrections while preserving source-order symmetry. For the reported checkpoints, T^{3}Fusion achieves higher Q_{abf} on all four datasets in Table 5, supporting this update structure at matched readout capacity. Fig. 7 complements the comparison by showing predictions, feature updates, and local output changes around vehicles and guardrail edges, illustrating where relation learning affects fusion.

5 Conclusion

In this paper, we present T^{3}Fusion, a test-time training framework for task-general image fusion. It fits compact bidirectional relations through closed-form ridge regression and uses a learned readout to interpret their predictions alongside the original source features. This connects pair-specific learning to fusion refinement while keeping persistent parameters fixed, without fused ground truth or test-time backpropagation. A single MSRS-trained model generalizes to unseen IVIF datasets and four additional fusion tasks. Controlled comparisons support the usefulness of pair-matched relations and the structured readout within this transferable framework. The computational cost of relation fitting and feature refinement makes deployment sensitive to input resolution and available resources. Future work will explore optimized solvers and temporal reuse of relations for efficient and temporally consistent video fusion.

AI USE STATEMENT

In this work, we used generative AI tools to assist with experimental scripting and workflow organization, technical discussions, result summarization, figure and table preparation, and manuscript editing. The authors reviewed the AI-assisted code and manuscript content, checked reported results against experimental outputs, and verified implementations through tests and experimental runs. Generative AI was not used to produce synthetic training or evaluation datasets. The authors take responsibility for the final methods, experiments, claims, and conclusions.

REFERENCES

Alexander Bartler, Andre Bühler, Felix Wiewel, Mario Döbler, and Bin Yang. MT3: Meta test-time training for self-supervised test-time adaption. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 151, pp. 3080–3090, 2022.

Luca Bertinetto, Joao F. Henriques, Philip H. S. Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. In *Proceedings of the International Conference on Learning Representations*, 2019.

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

Bing Cao, Yinan Xia, Yi Ding, Changqing Zhang, and Qinghua Hu. Test-time dynamic image fusion. In *Advances in Neural Information Processing Systems*, volume 37, pp. 2080–2105, 2024.

Chunyang Cheng, Tianyang Xu, Zhenhua Feng, Xiaojun Wu, Zhangyong Tang, Hui Li, Zeyang Zhang, Sara Atito, Muhammad Awais, and Josef Kittler. One model for ALL: Low-level task interaction is a key to task-agnostic image fusion. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 28102–28112, 2025.

Yossi Gandelsman, Yu Sun, Xinlei Chen, and Alexei A. Efros. Test-time training with masked autoencoders. In *Advances in Neural Information Processing Systems*, volume 35, pp. 29374–29385, 2022.

Lin Guo, Xiaoqing Luo, Wei Xie, Zhancheng Zhang, Hui Li, Rui Wang, Zhenhua Feng, and Xiaoning Song. Revisiting generative infrared and visible image fusion based on human cognitive laws. In *Advances in Neural Information Processing Systems*, volume 38, pp. 96322–96352, 2025.

Dongchen Han, Yining Li, Tianyu Li, Zixuan Cao, Ziming Wang, Jun Song, Yu Cheng, Bo Zheng, and Gao Huang. ViT^3: Unlocking test-time training in vision. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 51–61, 2026.

Xingyu Hu, Junjun Jiang, Chenyang Wang, Kui Jiang, Xianming Liu, and Jiayi Ma. Balancing task-invariant interaction and task-specific adaptation for unified image fusion. In *Proceedings of the International Conference on Computer Vision*, pp. 11262–11272, 2025.

Haiyang Jiang, Huiqin Zhang, Yanduo Zhang, Jiayi Ma, Junjun Jiang, and Huabing Zhou. PhyFusion: Physics-aware infrared and visible image fusion via modality-specific physical priors. In *Proceedings of the Conference on Computer Vision and Pattern Recognition Findings*, pp. 4993–5002, 2026.

Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLOv8. Ultralytics software, 2023.

Keith A. Johnson and J. Alex Becker. The whole brain atlas. Harvard Medical School, 1995.

Hui Li and Xiao-Jun Wu. DenseFuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing*, 28(5):2614–2623, 2019.

Jiawei Li, Hongwei Yu, Jiansheng Chen, Xinlong Ding, Jinlong Wang, Jinyuan Liu, Bochao Zou, and Huimin Ma. A^{2}RNet: Adversarial attack resilient network for robust infrared and visible image fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 4770–4778, 2025.

Xingyuan Li, Songcheng Du, Yang Zou, Haoyuan Xu, Zhiying Jiang, and Jinyuan Liu. UniFusion: A unified image fusion framework with robust representation and source-aware preservation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 33869–33880, 2026.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 2117–2125, 2017.

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 Conference on Computer Vision and Pattern Recognition*, pp. 5802–5811, 2022a.

Jinyuan Liu, Zhu Liu, Guanyao Wu, Long Ma, Risheng Liu, Wei Zhong, Zhongxuan Luo, and Xin Fan. Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In *Proceedings of the International Conference on Computer Vision*, pp. 8115–8124, 2023.

Jinyuan Liu, Guanyao Wu, Zhu Liu, Long Ma, Risheng Liu, and Xin Fan. Where elegance meets precision: Towards a compact, automatic, and flexible framework for multi-modality image fusion and applications. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 1110–1118, 2024.

Xiaowen Liu, Jing Li, Hongtao Huo, Haozhe Cao, Renhua Wang, and Xu Dong. More than meets the eye: A unified image fusion framework via semantic-pixel entropy trade-off for zero-shot generalization. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 41510–41520, 2026.

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin Transformer V2: Scaling up capacity and resolution. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 12009–12019, 2022b.

Miki Morimatsu, Yusuke Monno, Masayuki Tanaka, and Masatoshi Okutomi. Monochrome and color polarization demosaicking using edge-aware residual interpolation. In *Proceedings of the International Conference on Image Processing*, pp. 2571–2575, 2020.

Mansour Nejati, Shadrokh Samavi, and Shahram Shirani. Multi-focus image fusion using dictionary-based sparse representation. *Elsevier Information Fusion*, 25:72–84, 2015.

Weijian Su, Songqian Zhang, Yuqi Han, Jian Zhuang, Yongdong Huang, and Qiang Zhang. Fusion in your way: Aligning image fusion with heterogeneous demands via direct preference optimization. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 41499–41509, 2026.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A. Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *Proceedings of the International Conference on Machine Learning*, volume 119, pp. 9229–9248, 2020.

Yu Sun, Xinhao Li, Karan Dalal, Jiarui Xu, Arjun Vikram, Genghan Zhang, Yann Dubois, Xinlei Chen, Xiaolong Wang, Sanmi Koyejo, Tatsunori Hashimoto, and Carlos Guestrin. Learning to (learn at test time): RNNs with expressive hidden states. In *Proceedings of the International Conference on Machine Learning*, volume 267, pp. 57503–57522, 2025.

Linfeng Tang, Jiteng Yuan, Hao Zhang, Xingyu Jiang, and Jiayi Ma. PIAFusion: A progressive infrared and visible image fusion network based on illumination aware. *Elsevier Information Fusion*, 83–84:79–92, 2022.

Linfeng Tang, Yeda Wang, Zhanchuan Cai, Junjun Jiang, and Jiayi Ma. ControlFusion: A controllable image fusion network with language-vision degradation prompts. In *Advances in Neural Information Processing Systems*, volume 38, pp. 118304–118327, 2025.

Ke Alexander Wang, Jiaxin Shi, and Emily B. Fox. Test-time regression: A unifying framework for designing sequence models with associative memory. *Journal of Machine Learning Research*, 27 (169):1–41, 2026a.

Zeyu Wang, Jiawei Feng, Jiayu Wang, Pengjie Wang, and Haiyu Song. SigFusion: Unified signal-level self-supervised learning paradigm for image fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 10385–10393, 2026b.

Guanyao Wu, Haoyu Liu, Hongming Fu, Yichuan Peng, Jinyuan Liu, Xin Fan, and Risheng Liu. Every SAM drop counts: Embracing semantic priors for multi-modality image fusion and beyond. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 17882–17891, 2025.

Jianghan Xia, Hong Song, Jinfu Li, Yucong Lin, Shihan Ma, Jingfan Fan, Danni Ai, Tianyu Fu, Deqiang Xiao, and Jian Yang. RegionFuse: Region-adaptive pixel distribution learning for infrared and visible image fusion. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 19539–19548, 2026.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. SegFormer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems*, volume 34, pp. 12077–12090, 2021.

Qinhua Xie and Hao Tang. TTTFusion: A test-time training-based strategy for multimodal medical image fusion in surgical robots, 2025.

Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. U2Fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1): 502–518, 2022.

Hao Zhang, Zhuliang Le, Zhenfeng Shao, Han Xu, and Jiayi Ma. MFF-GAN: An unsupervised generative adversarial network with adaptive and gradient joint constraints for multi-focus image fusion. *Elsevier Information Fusion*, 66:40–53, 2021.

Hao Zhang, Shuhan Yang, Linfeng Tang, Xunpeng Yi, and Jiayi Ma. ReCoFuse: Ultra-robust image fusion via restorative multi-modal diffusion reciprocal coupling. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 33838–33847, 2026.

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

Xingchen Zhang and Yiannis Demiris. Visible and infrared image fusion using deep learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10535–10554, 2023.

Wenda Zhao, Shigeng Xie, Fan Zhao, You He, and Huchuan Lu. MetaFusion: Infrared and visible image fusion via meta-feature embedding from object detection. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 13955–13965, 2023a.

Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Shuang Xu, Zudi Lin, Radu Timofte, and Luc Van Gool. CDDFuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 5906–5916, 2023b.

Zixiang Zhao, Haowen Bai, Yuanzhi Zhu, Jiangshe Zhang, Shuang Xu, Yulun Zhang, Kai Zhang, Deyu Meng, Radu Timofte, and Luc Van Gool. DDFM: Denoising diffusion model for multi-modality image fusion. In *Proceedings of the International Conference on Computer Vision*, pp. 8082–8093, 2023c.

Guoqing Zheng, Zhenqi Fu, Xiaopeng Lin, Xueye Chu, Yue Huang, and Xinghao Ding. Infrared and visible image fusion via test-time training. In *Proceedings of the Chinese Conference on Pattern Recognition and Computer Vision*, volume 14434, pp. 77–88, 2024a.

Naishan Zheng, Man Zhou, Jie Huang, Junming Hou, Haoying Li, Yuan Xu, and Feng Zhao. Probing synergistic high-order interaction in infrared and visible image fusion. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 26384–26395, 2024b.

Pengfei Zhu, Yang Sun, Bing Cao, and Qinghua Hu. Task-customized mixture of adapters for general image fusion. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 7099–7108, 2024.

Appendix

A Derivations

A.1 Closed-Form Solution

Problem and dimensions. Fix one scale, direction, and query fold, and suppress these indices. For $n > 0$ support tokens, set $d = r + 1$ and write

$$\bar{\boldsymbol{K}} = \begin{bmatrix} \boldsymbol{K} \\ \mathbf{1}_n^\top \end{bmatrix} \in \mathbb{R}^{d\times n}, \qquad \boldsymbol{V} \in \mathbb{R}^{r\times n}, \qquad \boldsymbol{W} = [\boldsymbol{A}\ \boldsymbol{b}] \in \mathbb{R}^{r\times d}. \tag{S1}$$

Thus $\boldsymbol{W}\bar{\boldsymbol{K}} = \boldsymbol{A}\boldsymbol{K} + \boldsymbol{b}\mathbf{1}_n^\top$ is an affine prediction. Tensor symbols follow the main paper; w_{ac}, $\bar{k}_{ci}$, and v_{ai} below denote their scalar entries. With $\rho > 0$,

$$\begin{aligned} J(\boldsymbol{W}) &= \frac{1}{n}\|\boldsymbol{W}\bar{\boldsymbol{K}} - \boldsymbol{V}\|_{\mathrm{F}}^2 + \rho\|\boldsymbol{W}\|_{\mathrm{F}}^2 \\ &= \frac{1}{n}\sum_{a=1}^{r}\sum_{i=1}^{n}\left(\sum_{c=1}^{d} w_{ac}\bar{k}_{ci} - v_{ai}\right)^2 + \rho\sum_{a=1}^{r}\sum_{c=1}^{d} w_{ac}^2. \end{aligned} \tag{S2}$$

The last column of $\boldsymbol{W}$ is the intercept, so the penalty includes $\rho\|\boldsymbol{b}\|_2^2$ as well as $\rho\|\boldsymbol{A}\|_{\mathrm{F}}^2$.

Derivative and normal equations. For any output index a and coefficient index b, the chain rule gives

$$\begin{aligned} \frac{\partial J}{\partial w_{ab}} &= \frac{2}{n}\sum_{i=1}^{n}\left(\sum_{c=1}^{d} w_{ac}\bar{k}_{ci} - v_{ai}\right)\bar{k}_{bi} + 2\rho w_{ab}, \\ \nabla_{\boldsymbol{W}} J &= \frac{2}{n}(\boldsymbol{W}\bar{\boldsymbol{K}} - \boldsymbol{V})\bar{\boldsymbol{K}}^\top + 2\rho\boldsymbol{W}. \end{aligned} \tag{S3}$$

Setting this gradient to zero, dividing by two, and collecting $\boldsymbol{W}$ yields

$$\begin{aligned} &\frac{1}{n}\boldsymbol{W}\bar{\boldsymbol{K}}\bar{\boldsymbol{K}}^\top - \frac{1}{n}\boldsymbol{V}\bar{\boldsymbol{K}}^\top + \rho\boldsymbol{W} = 0, \\ &\boldsymbol{W}\underbrace{\left(\frac{1}{n}\bar{\boldsymbol{K}}\bar{\boldsymbol{K}}^\top + \rho\boldsymbol{I}_d\right)}_{\boldsymbol{M}} = \underbrace{\frac{1}{n}\boldsymbol{V}\bar{\boldsymbol{K}}^\top}_{\boldsymbol{B}}. \end{aligned} \tag{S4}$$

Here $\boldsymbol{M} \in \mathbb{R}^{d\times d}$ is symmetric and $\boldsymbol{B} \in \mathbb{R}^{r\times d}$. For any nonzero $\boldsymbol{x} \in \mathbb{R}^d$,

$$\boldsymbol{x}^\top\boldsymbol{M}\boldsymbol{x} = \frac{1}{n}(\bar{\boldsymbol{K}}^\top\boldsymbol{x})^\top(\bar{\boldsymbol{K}}^\top\boldsymbol{x}) + \rho\boldsymbol{x}^\top\boldsymbol{x} = \frac{1}{n}\|\bar{\boldsymbol{K}}^\top\boldsymbol{x}\|_2^2 + \rho\|\boldsymbol{x}\|_2^2 > 0. \tag{S5}$$

Consequently $\boldsymbol{M}$ is invertible even when the Keys are rank-deficient. Right-multiplying Eq. (S4) by $\boldsymbol{M}^{-1}$, and using $\boldsymbol{M} = n^{-1}(\bar{\boldsymbol{K}}\bar{\boldsymbol{K}}^\top + n\rho\boldsymbol{I}_d)$, gives

$$\begin{aligned} \boldsymbol{W}^* = \boldsymbol{B}\boldsymbol{M}^{-1} &= \frac{1}{n}\boldsymbol{V}\bar{\boldsymbol{K}}^\top\left[n(\bar{\boldsymbol{K}}\bar{\boldsymbol{K}}^\top + n\rho\boldsymbol{I}_d)^{-1}\right] \\ &= \boldsymbol{V}\bar{\boldsymbol{K}}^\top(\bar{\boldsymbol{K}}\bar{\boldsymbol{K}}^\top + n\rho\boldsymbol{I}_d)^{-1}. \end{aligned} \tag{S6}$$

Uniqueness. Let $\boldsymbol{E} = \boldsymbol{W}^*\bar{\boldsymbol{K}} - \boldsymbol{V}$ and let $\boldsymbol{D}$ be any perturbation of $\boldsymbol{W}^*$. Use the Frobenius inner product $\langle \boldsymbol{U}, \boldsymbol{T}\rangle_{\mathrm{F}} = \mathrm{tr}(\boldsymbol{U}^\top\boldsymbol{T})$. Expanding both squares in Eq. (S2) gives

$$\begin{aligned} J(\boldsymbol{W}^* + \boldsymbol{D}) - J(\boldsymbol{W}^*) &= \frac{2}{n}\langle \boldsymbol{E}, \boldsymbol{D}\bar{\boldsymbol{K}}\rangle_{\mathrm{F}} + 2\rho\langle \boldsymbol{W}^*, \boldsymbol{D}\rangle_{\mathrm{F}} + \frac{1}{n}\|\boldsymbol{D}\bar{\boldsymbol{K}}\|_{\mathrm{F}}^2 + \rho\|\boldsymbol{D}\|_{\mathrm{F}}^2 \\ &= 2\left\langle \frac{1}{n}\boldsymbol{E}\bar{\boldsymbol{K}}^\top + \rho\boldsymbol{W}^*, \boldsymbol{D}\right\rangle_{\mathrm{F}} + \frac{1}{n}\|\boldsymbol{D}\bar{\boldsymbol{K}}\|_{\mathrm{F}}^2 + \rho\|\boldsymbol{D}\|_{\mathrm{F}}^2 \\ &= \frac{1}{n}\|\boldsymbol{D}\bar{\boldsymbol{K}}\|_{\mathrm{F}}^2 + \rho\|\boldsymbol{D}\|_{\mathrm{F}}^2. \end{aligned} \tag{S7}$$

The inner product vanishes by Eq. (S3). The final expression is strictly positive for every $\boldsymbol{D} \neq 0$, proving uniqueness. In code, $\boldsymbol{M}^\top = \boldsymbol{M}$ permits the equivalent system $\boldsymbol{M}(\boldsymbol{W}^*)^\top = \boldsymbol{B}^\top$: a linear solver returns $(\boldsymbol{W}^*)^\top$ without explicitly forming an inverse. This guarantees a unique regression solution, not an improvement in fusion quality.

A.2 Differentiation Through the Solve

During offline learning, support membership and n are fixed for a given forward pass, whereas Keys, Values, and ρ depend on trainable parameters. Differentiating $\boldsymbol{W}^*\boldsymbol{M} = \boldsymbol{B}$ by the product rule gives

$$\begin{aligned}(\mathrm{d}\boldsymbol{W}^*)\boldsymbol{M} + \boldsymbol{W}^*(\mathrm{d}\boldsymbol{M}) &= \mathrm{d}\boldsymbol{B},\\ \mathrm{d}\boldsymbol{W}^* &= (\mathrm{d}\boldsymbol{B} - \boldsymbol{W}^*\mathrm{d}\boldsymbol{M})\boldsymbol{M}^{-1},\\ \mathrm{d}\boldsymbol{M} &= \frac{1}{n}\left[(\mathrm{d}\bar{\boldsymbol{K}})\bar{\boldsymbol{K}}^\top + \bar{\boldsymbol{K}}(\mathrm{d}\bar{\boldsymbol{K}})^\top\right] + (\mathrm{d}\rho)\boldsymbol{I}_d,\\ \mathrm{d}\boldsymbol{B} &= \frac{1}{n}\left[(\mathrm{d}\boldsymbol{V})\bar{\boldsymbol{K}}^\top + \boldsymbol{V}(\mathrm{d}\bar{\boldsymbol{K}})^\top\right].\end{aligned} \tag{S8}$$

Substituting the last two lines into the second and collecting terms gives

$$\mathrm{d}\boldsymbol{W}^* = \left\{\frac{1}{n}\left[(\mathrm{d}\boldsymbol{V})\bar{\boldsymbol{K}}^\top + (\boldsymbol{V} - \boldsymbol{W}^*\bar{\boldsymbol{K}})(\mathrm{d}\bar{\boldsymbol{K}})^\top - \boldsymbol{W}^*(\mathrm{d}\bar{\boldsymbol{K}})\bar{\boldsymbol{K}}^\top\right] - \boldsymbol{W}^*\mathrm{d}\rho\right\}\boldsymbol{M}^{-1}. \tag{S9}$$

To make the backward computation explicit, let L be a scalar outer loss, $\boldsymbol{G} = \partial L/\partial \boldsymbol{W}^*$ its incoming gradient, and $\boldsymbol{U} = \boldsymbol{G}\boldsymbol{M}^{-\top} = \boldsymbol{G}\boldsymbol{M}^{-1}$. Cyclic permutation inside a trace gives $\langle \boldsymbol{G}, \boldsymbol{T}\boldsymbol{M}^{-1}\rangle_{\mathrm{F}} = \langle \boldsymbol{U}, \boldsymbol{T}\rangle_{\mathrm{F}}$. Applying this identity to each term,

$$\begin{aligned}\mathrm{d}L\big|_{\mathrm{solve}} &= \langle \boldsymbol{G}, \mathrm{d}\boldsymbol{W}^*\rangle_{\mathrm{F}}\\ &= \left\langle \frac{1}{n}\boldsymbol{U}\bar{\boldsymbol{K}}, \mathrm{d}\boldsymbol{V}\right\rangle_{\mathrm{F}} + \left\langle \frac{1}{n}\left[\boldsymbol{U}^\top(\boldsymbol{V} - \boldsymbol{W}^*\bar{\boldsymbol{K}}) - (\boldsymbol{W}^*)^\top\boldsymbol{U}\bar{\boldsymbol{K}}\right], \mathrm{d}\bar{\boldsymbol{K}}\right\rangle_{\mathrm{F}} - \langle \boldsymbol{U}, \boldsymbol{W}^*\rangle_{\mathrm{F}}\,\mathrm{d}\rho.\end{aligned} \tag{S10}$$

The coefficients of the three differentials are therefore

$$\begin{aligned}\left.\frac{\partial L}{\partial \boldsymbol{V}}\right|_{\mathrm{solve}} &= \frac{1}{n}\boldsymbol{U}\bar{\boldsymbol{K}}, \qquad \left.\frac{\partial L}{\partial \rho}\right|_{\mathrm{solve}} = -\langle \boldsymbol{U}, \boldsymbol{W}^*\rangle_{\mathrm{F}},\\ \left.\frac{\partial L}{\partial \bar{\boldsymbol{K}}}\right|_{\mathrm{solve}} &= \frac{1}{n}\left[\boldsymbol{U}^\top(\boldsymbol{V} - \boldsymbol{W}^*\bar{\boldsymbol{K}}) - (\boldsymbol{W}^*)^\top\boldsymbol{U}\bar{\boldsymbol{K}}\right].\end{aligned} \tag{S11}$$

Their dimensions are $r \times n$, scalar, and $d \times n$, respectively. The appended row of ones is constant, so its differential is zero. For $\rho = 0.1\exp(a)$, the chain rule gives $\partial L/\partial a = \rho\,\partial L/\partial \rho$; contributions are summed over the regression problems sharing a. These are the *solve-path* contributions: Values also receive gradients through the head inputs and query targets, and Keys through query predictions. The frozen teacher features remain detached while their Value projection is trainable. No backward pass is performed at inference.

A.3 Cross-Fitting and Query Loss

Let $\mathcal{U}_s$ be the valid token grid of size $\lceil H/2^s\rceil \times \lceil W/2^s\rceil$, $s \in \{1,2,3\}$, after bottom/right padding of the input to a multiple of eight. Define

$$\begin{aligned}g(i,j) &= 2(i \bmod 2) + (j \bmod 2), \qquad \mathcal{Q}_{s,g} = \{p \in \mathcal{U}_s : g(p) = g\},\\ \mathcal{S}_{s,g} &\subseteq \mathcal{U}_s \setminus \mathcal{Q}_{s,g}, \qquad |\mathcal{S}_{s,g}| = \min(1024, |\mathcal{U}_s \setminus \mathcal{Q}_{s,g}|).\end{aligned} \tag{S12}$$

When the candidate count N exceeds 1,024, the selected offsets in the flattened ordered list are $\mathrm{round}(j(N-1)/1023)$, $j = 0, \ldots, 1023$; otherwise all candidates are used. Both directions use the same indices. For $B \to A$, the columns are $\bar{\boldsymbol{k}}^s_B(p)$ and $\boldsymbol{v}^s_A(p)$ at matching support positions. The reverse fit exchanges A and B. Each valid query is then predicted once:

$$\widehat{\boldsymbol{v}}^s_A(p) = \boldsymbol{W}^{s,*}_{B\to A,g(p)}\bar{\boldsymbol{k}}^s_B(p), \qquad \widehat{\boldsymbol{v}}^s_B(p) = \boldsymbol{W}^{s,*}_{A\to B,g(p)}\bar{\boldsymbol{k}}^s_A(p). \tag{S13}$$

Because the four query sets are disjoint and their union is $\mathcal{U}_s$, the channel/token mean for each direction and scale, followed by the mean over six direction–scale combinations, is

$$\begin{aligned}\mathcal{L}_{\mathrm{q}} &= \frac{1}{6}\sum_{s=1}^{3}\sum_{X\in\{A,B\}}\frac{1}{r|\mathcal{U}_s|}\sum_{g=0}^{3}\sum_{p\in\mathcal{Q}_{s,g}}\sum_{c=1}^{r}(\widehat{v}^s_{X,c}(p) - v^s_{X,c}(p))^2\\ &= \frac{1}{6}\sum_{s=1}^{3}\sum_{X\in\{A,B\}}\frac{1}{r|\mathcal{U}_s|}\sum_{p\in\mathcal{U}_s}\|\widehat{\boldsymbol{v}}^s_X(p) - \boldsymbol{v}^s_X(p)\|_2^2.\end{aligned} \tag{S14}$$

Thus folds contribute in proportion to their valid query counts; they are not averaged equally when their sizes differ. Padding-only tokens are excluded from fitting and this loss. If a tiny input gives an empty support or query set, the implementation skips that solve and retains the zero reference mapping; Eq. (S2) is used only for nonempty support. Boundary tokens can contain padding context and nearby features share receptive fields. Holding out query targets therefore does not establish statistical independence of support and query features.

A.4 Source-Exchange Symmetry

A tilde denotes the computation after swapping the two luminance inputs. Shared encoders, projections, and calibration imply, at every scale, $\widetilde{\mathbf{F}}^s_A = \mathbf{F}^s_B$, $\widetilde{\bar{K}}^s_A = \bar{K}^s_B$, and $\widetilde{V}^s_A = V^s_B$, with the converse identities for source B. The support sets depend only on spatial indices and validity, so they are unchanged. Substituting the swapped Keys and Values into the closed form, or using uniqueness of the swapped regression problems, gives

$$\begin{aligned} \widetilde{W}^{s,*}_{B\to A,g} &= W^{s,*}_{A\to B,g}, \qquad \widetilde{W}^{s,*}_{A\to B,g} = W^{s,*}_{B\to A,g}, \\ \widetilde{\hat{v}}^s_A(p) &= \widetilde{W}^{s,*}_{B\to A,g(p)} \widetilde{\bar{k}}^s_B(p) = W^{s,*}_{A\to B,g(p)} \bar{k}^s_A(p) = \hat{v}^s_B(p). \end{aligned} \tag{S15}$$

The other prediction exchanges identically. Suppress s below. Define channel concatenation by $[\,\cdot\,]$ and write the head inputs explicitly:

$$\mathbf{Z}_X = [\mathbf{F}_X, \mathbf{V}_X, \widehat{\mathbf{V}}_X, \mathbf{V}_X - \widehat{\mathbf{V}}_X], \qquad \mathbf{Z}^0_X = [\mathbf{F}_X, \mathbf{V}_X, 0, \mathbf{V}_X]. \tag{S16}$$

Both inputs exchange under swapping. Since each head h_j is shared,

$$\widetilde{\mathbf{R}}_{j,A} = h_j(\widetilde{\mathbf{Z}}_A) - h_j(\widetilde{\mathbf{Z}}^0_A) = h_j(\mathbf{Z}_B) - h_j(\mathbf{Z}^0_B) = \mathbf{R}_{j,B}, \qquad \widetilde{\mathbf{R}}_{j,B} = \mathbf{R}_{j,A}. \tag{S17}$$

Substitution into the common and difference updates then yields

$$\begin{aligned} \widetilde{\boldsymbol{\Delta}}_{\mathrm{c}} &= \tfrac{1}{2}(\mathbf{R}_{\mathrm{c},B} + \mathbf{R}_{\mathrm{c},A}) = \boldsymbol{\Delta}_{\mathrm{c}}, \qquad \widetilde{\boldsymbol{\Delta}}_{\mathrm{d}} = \tfrac{1}{2}(\mathbf{R}_{\mathrm{d},B} - \mathbf{R}_{\mathrm{d},A}) = -\boldsymbol{\Delta}_{\mathrm{d}}, \\ \widetilde{\mathbf{F}}'_A &= \mathbf{F}_B + \boldsymbol{\Delta}_{\mathrm{c}} - \boldsymbol{\Delta}_{\mathrm{d}} = \mathbf{F}'_B, \qquad \widetilde{\mathbf{F}}'_B = \mathbf{F}_A + \boldsymbol{\Delta}_{\mathrm{c}} + \boldsymbol{\Delta}_{\mathrm{d}} = \mathbf{F}'_A. \end{aligned} \tag{S18}$$

For completeness, the feature mean and half-difference change as

$$\begin{aligned} \tfrac{1}{2}(\mathbf{F}'_A + \mathbf{F}'_B) &= \tfrac{1}{2}(\mathbf{F}_A + \mathbf{F}_B + 2\boldsymbol{\Delta}_{\mathrm{c}}) = \tfrac{1}{2}(\mathbf{F}_A + \mathbf{F}_B) + \boldsymbol{\Delta}_{\mathrm{c}}, \\ \tfrac{1}{2}(\mathbf{F}'_A - \mathbf{F}'_B) &= \tfrac{1}{2}(\mathbf{F}_A - \mathbf{F}_B + 2\boldsymbol{\Delta}_{\mathrm{d}}) = \tfrac{1}{2}(\mathbf{F}_A - \mathbf{F}_B) + \boldsymbol{\Delta}_{\mathrm{d}}. \end{aligned} \tag{S19}$$

To verify that fusion preserves this symmetry, consider any channel and spatial position. Let $\mathbf{M}_F = (\mathbf{F}'_A + \mathbf{F}'_B)/2$ and $\mathbf{D}_F = |\mathbf{F}'_A - \mathbf{F}'_B|$. Both are unchanged by swapping. The shared scorer q produces logits $\ell_X = q(\mathbf{F}'_X, \mathbf{M}_F, \mathbf{D}_F)$ at that position, so $\widetilde{\ell}_A = \ell_B$ and $\widetilde{\ell}_B = \ell_A$. Therefore

$$\begin{aligned} \widetilde{\alpha}_A &= \frac{e^{\widetilde{\ell}_A}}{e^{\widetilde{\ell}_A} + e^{\widetilde{\ell}_B}} = \frac{e^{\ell_B}}{e^{\ell_B} + e^{\ell_A}} = \alpha_B, \qquad \widetilde{\alpha}_B = \alpha_A, \\ \widetilde{F}_{\mathrm{fused}} &= \widetilde{\alpha}_A \widetilde{F}'_A + \widetilde{\alpha}_B \widetilde{F}'_B = \alpha_B F'_B + \alpha_A F'_A = F_{\mathrm{fused}}. \end{aligned} \tag{S20}$$

The same two-source selection rule fuses the original detail tensors, so its output is also unchanged. All decoder inputs, the sigmoid, and output crop are consequently identical, proving $\widetilde{\boldsymbol{Y}} = \boldsymbol{Y}$ in exact arithmetic. This assumes shared source processing and source-independent sampling. Modality-specific color recovery must retain the designated chrominance source when the inputs are reordered.

A.5 Zero-Prediction Reference

If $\widehat{\mathbf{V}}^s_A = \widehat{\mathbf{V}}^s_B = 0$ at every scale, Eq. (S16) gives, separately for each source,

$$\begin{aligned} \mathbf{V}^s_X - \widehat{\mathbf{V}}^s_X &= \mathbf{V}^s_X, \qquad \mathbf{Z}^s_X = [\mathbf{F}^s_X, \mathbf{V}^s_X, 0, \mathbf{V}^s_X] = \mathbf{Z}^{s,0}_X, \\ \mathbf{R}^s_{j,X} &= h^s_j(\mathbf{Z}^{s,0}_X) - h^s_j(\mathbf{Z}^{s,0}_X) = 0, \\ \boldsymbol{\Delta}^s_{\mathrm{c}} &= \tfrac{1}{2}(0+0) = 0, \qquad \boldsymbol{\Delta}^s_{\mathrm{d}} = \tfrac{1}{2}(0-0) = 0, \qquad \mathbf{F}'^{,s}_X = \mathbf{F}^s_X. \end{aligned} \tag{S21}$$

The frozen scorers thus receive exactly the Base features and return the Base weights. Detail tensors are unchanged as well, so the decoder receives the same inputs and $\boldsymbol{Y} = \boldsymbol{Y}_0$. This holds for any learned head parameters. At initialization there is a second sufficient condition: writing a head as $h_j(\mathbf{Z}) = \mathbf{C}_{j,2}\,\mathrm{SiLU}(\mathbf{C}_{j,1}\mathbf{Z} + \boldsymbol{b}_{j,1})$, its bias-free output convolution has $\mathbf{C}_{j,2} = 0$, hence $h_j(\mathbf{Z}) = h_j(\mathbf{Z}^0) = 0$ for arbitrary predictions. All updates again vanish. Neither condition is a quality-based acceptance gate; inference never chooses between Base and TTT using a quality oracle.

B Additional Implementation Details

Tables S1 and S2 summarize the architecture and offline training. The context branch follows Swin V2 (Liu et al., 2022b); encoders, source scorers, and refinement heads share parameters between sources.

Table S1: Core architecture. Spatial sizes refer to padded inputs.

Component	Configuration
Context stages	Widths $(32, 64, 128)$; depths $(4, 2, 2)$; heads $(2, 4, 8)$; patch size 2; window size 8; MLP expansion 4; no dropout
Context resolution	$H/2 \times W/2$, $H/4 \times W/4$, $H/8 \times W/8$
Detail branch	3×3 input convolution, width 32; two local residual blocks
Local block	Depthwise 3×3, channel LayerNorm, 1×1 expansion by 2, GELU, 1×1 contraction; residual scale initialized to 0.1
Source scorer	$3C \to \max(C/2, 8) \to C$ pointwise convolutions; LayerNorm and GELU between them; bias-free output
Decoder	Width 96; two residual blocks per context level; bilinear upsampling; full-resolution detail concatenation; two output residual blocks; 3×3 luminance convolution and sigmoid
Projections	$r = 16$ per scale; row-normalized projection weights
Refinement heads	Input $(80, 112, 176)$, hidden $(32, 48, 64)$, output $(32, 64, 128)$; 1×1, SiLU, bias-free 1×1; separate common/difference heads

We use 975 MSRS training pairs. The adaptation stage uses the fixed 64-pair validation subset recorded in the released split. We use paired 256×256 random crops and horizontal flips during training. Fusion training updates the scorers for five epochs, then jointly updates the encoder, decoder, and scorers for 145 epochs. The final offline stage freezes the fusion network and learns the projections, refinement heads, and ridge coefficient. Feature normalization statistics are computed from the training pairs and fixed. The main results use the selected fusion-network epoch 147 and final adaptation epoch 50 across all datasets.

Table S2: Offline stages. AdamW uses default weight decay 10^{-4}, gradient-norm clipping at 1.0, and seed 3407.

	Reconstruction	Fusion	Adaptation
Epochs / batch size	50 / 16	150 / 16	50 / 1
Initial learning rate	10^{-4}	10^{-4}	5×10^{-5}
Schedule	Cosine	Constant	Cosine
Final learning rate	10^{-6}	10^{-4}	2.5×10^{-6}
Precision	AMP	AMP	FP32
Fusion network	Encoder/decoder updated	Staged update	Frozen

C Evaluation Settings

Datasets and metrics. The evaluation contains 1,691 pairs: MSRS (361) (Tang et al., 2022), FMB (280) (Liu et al., 2023), M3FD (300) (Liu et al., 2022a), and RoadScene (221) (Xu et al., 2022) for IVIF; MEFB (100) (Zhang, 2021) for MEF; MFI-WHU (120) (Zhang et al., 2021) for MFF; PET–MRI (269) (Johnson & Becker, 1995) for MIF; and Polarization (40 DoLP/S_0 pairs) (Morimatsu

et al., 2020) for MPF. Dataset means weight pairs equally. VIF sums the two source-to-fused scores; source-based SSIM averages the two scores. MFI-WHU all-focus references are used only for the separate reference-based evaluation.

Image preprocessing. Our network fuses $Y = 0.299R + 0.587G + 0.114B$ in $[0, 1]$ at native resolution; padding is removed from the output. IVIF retains released RGB outputs and uses visible chrominance for luminance-output methods; PET–MRI uses PET chrominance. MEF/MFF/MPF use the symmetric composition

$$q_X = -\frac{|Y_X - Y|}{0.08} - \frac{1}{2}\left(\frac{Y_X - 0.5}{0.20}\right)^2, \quad \alpha_X = \frac{e^{q_X}}{e^{q_A} + e^{q_B}}, \quad C_b = \alpha_A C_{b,A} + \alpha_B C_{b,B}, \tag{S22}$$

and analogously for C_r. RGB is clipped to $[0, 1]$; fusion metrics use the grayscale conversion of saved 8-bit RGB.

Downstream evaluation. For each task, two evaluators are trained separately on IR and VI: SegFormer-B1 (Xie et al., 2021) on FMB and YOLOv8s (Jocher et al., 2023) on M3FD. Table S3 gives their training settings. Their validation-selected checkpoints remain fixed across fusion methods. Reported scores average their metrics. FMB mIoU ignores class 0 and averages the 13 ground-truth-present classes $\{1, \ldots, 12, 14\}$. YOLO evaluation uses image size 640, confidence 0.001, and NMS IoU 0.7.

Table S3: Training of the fixed downstream evaluators. Both use AdamW, 300 epochs, and seed 3407.

	FMB / SegFormer-B1	M3FD / YOLOv8s
Train / validation / test	976 / 244 / 280	3,120 / 780 / 300
Batch size	8 (4, accumulated twice)	8
Crop / image size	512×512	640
Initial learning rate	6×10^{-5}	10^{-3}
Head LR factor	10	Native configuration
Weight decay	0.01	0.0005
Warm-up / schedule	1,500 steps / polynomial	5 epochs / cosine

D Further Evidence for Discussion

D.1 Correspondence Between Inputs and Relations

The intervention in Table 4 uses the fixed checkpoint. For each recipient pair, all 24 mapping matrices are either fitted to that pair or transferred together from a donor pair in the same dataset. The recipient features, projections, query targets, and readout remain fixed. We average three shuffle trials before computing dataset-level differences. These are repetitions of the mapping assignment, not independent training runs.

Held-out prediction MSE is evaluated on the query locations excluded from the corresponding fit. The reported reduction is

$$100\frac{\overline{E}_{\text{shuffle}} - E_{\text{matched}}}{\overline{E}_{\text{shuffle}}}. \tag{S23}$$

Fusion changes are matched minus shuffled dataset means, multiplied by 10^3 in Table 4. Prediction reductions are percentages. All three fusion differences are positive on all eight datasets, supporting input–relation correspondence within the fixed model. The accompanying data retain unrounded values.

Table S4: Complete GD sweep for VIF↑. All variants use the same trained checkpoint. Closed form is the original solver.

Dataset	GD-1	GD-2	GD-4	GD-8	GD-16	GD-32	GD-64	Closed form
MSRS	1.098706	1.110098	1.123433	1.125905	1.128310	1.131326	1.133792	**1.134789**
FMB	1.091778	1.111908	1.125095	**1.125775**	1.124409	1.119318	1.112385	1.105793
M3FD	0.986086	1.008526	1.015377	**1.015950**	1.014272	1.008585	1.003458	0.993597
RoadScene	0.806287	0.807542	0.808111	0.809328	0.811392	0.814036	0.816315	**0.820294**
MEFB	1.360859	1.362581	1.362623	1.363021	1.363634	1.364586	1.366967	**1.368432**
MFI-WHU	1.177595	1.178324	1.179524	1.180614	1.182060	1.183713	1.185099	**1.187267**
PET-MRI	0.795056	0.795254	0.796464	0.797109	0.797403	0.797250	0.796915	**0.797502**
Polarization	1.008313	1.054734	**1.149777**	1.138040	1.140412	1.139542	1.139975	1.140133

Table S5: Complete GD sweep for Q_{abf} ↑. All variants use the same trained checkpoint. Closed form is the original solver.

Dataset	GD-1	GD-2	GD-4	GD-8	GD-16	GD-32	GD-64	Closed form
MSRS	0.702896	0.708594	0.717964	0.720132	0.721527	0.723226	0.724707	**0.725828**
FMB	0.712052	0.710245	0.708103	0.708208	0.709453	0.711490	0.713540	**0.717863**
M3FD	0.673138	0.672436	0.671594	0.672353	0.674380	0.676604	0.678589	**0.681452**
RoadScene	0.592009	0.594054	0.596564	0.597876	0.599328	0.601287	0.603127	**0.606064**
MEFB	0.714245	0.720603	0.728088	0.729824	0.730695	0.731432	0.731964	**0.733232**
MFI-WHU	0.679661	0.678969	0.678982	0.679342	0.679891	0.680430	0.680968	**0.682103**
PET-MRI	0.759526	0.759430	0.760424	0.761697	0.762707	0.763182	0.763395	**0.763562**
Polarization	0.562138	0.590188	0.657539	0.669131	0.671646	0.674755	0.678230	**0.682084**

D.2 Closed-Form and Iterative Relation Fitting

The GD variants reuse the seed-3407 projections, normalization, ridge coefficient, and readout of T^{3}Fusion. Only the numerical solution of each relation problem changes. Let $\boldsymbol{A}$ denote the regularized support Gram matrix and $\boldsymbol{B}$ the corresponding cross-covariance term including the ridge reference. Starting from $\boldsymbol{W}_0 = 0$, the implementation applies

$$\boldsymbol{W}_{k+1} = \boldsymbol{W}_k + (\boldsymbol{B} - \boldsymbol{W}_k \boldsymbol{A})/L, \qquad L = \|\boldsymbol{A}\|_\infty. \tag{S24}$$

For the quadratic gradient $2(\boldsymbol{W}\boldsymbol{A} - \boldsymbol{B})$, this is a step size of $1/(2L)$. The bound is computed separately for each system. We evaluate 1, 2, 4, 8, 16, 32, and 64 steps without retraining. Figure 6 displays the retrospectively chosen subset of 2, 8, and 32 steps. Tables S4–S6 give the complete sweep to make the dependence on the update budget visible.

Table S6: Complete GD sweep for Q_{cb} ↑. All variants use the same trained checkpoint. Closed form is the original solver.

Dataset	GD-1	GD-2	GD-4	GD-8	GD-16	GD-32	GD-64	Closed form
MSRS	0.579890	0.587297	0.600186	0.603168	0.604100	0.605159	**0.606043**	0.605863
FMB	0.498512	0.502488	0.504602	0.504616	0.504556	0.504579	0.504849	**0.505700**
M3FD	0.510991	0.518079	0.520038	0.520269	**0.521150**	0.520869	0.520169	0.519284
RoadScene	0.518309	0.527994	0.533430	**0.533616**	0.532319	0.530180	0.527554	0.522732
MEFB	0.399382	0.406443	0.413534	**0.414168**	0.413839	0.413493	0.412913	0.413392
MFI-WHU	**0.687482**	0.687034	0.686723	0.686435	0.685810	0.685504	0.685242	0.685761
PET-MRI	0.401737	0.397626	0.401834	0.408561	0.403112	0.406021	0.416355	**0.438757**
Polarization	0.326150	0.327129	0.346472	0.347755	0.349871	0.352394	0.355494	**0.356525**

E Additional Multi-Task Training Study

We compare the MSRS-only T^{3}Fusion model used in the main paper with a version jointly trained on MSRS, SICE (Cai et al., 2018), and Lytro (Nejati et al., 2015). Both retain test-time relation fitting. We also compare with released implementations and checkpoints of U2Fusion (Xu et al.,

2022), TITA (Hu et al., 2025), UniFusion (Li et al., 2026), and SigFusion (Wang et al., 2026b), whose training data may differ from ours.

Table S7 shows that joint training improves VIF on all four datasets and all four metrics on MEFB relative to MSRS-only training. It also improves Q_{abf} and Q_{cb} on MFI-WHU, although the benefits vary across datasets and metrics. The jointly trained model achieves the highest VIF among the compared methods on all four datasets, while UniFusion obtains higher Q_{abf} and Q_{cb} on MFI-WHU.

Table S7: Comparison with additional training data. T^{3}Fusion uses either MSRS only or joint training on MSRS, SICE, and Lytro, with test-time adaptation in both settings. Bold and underlined values indicate the best and second-best scores.

Method	MSRS				RoadScene			
	AG↑	VIF↑	Q_{abf} ↑	Q_{cb} ↑	AG↑	VIF↑	Q_{abf} ↑	Q_{cb} ↑
U2Fusion TPAMI'22	2.289	0.561	0.340	0.498	4.759	0.682	0.505	**0.530**
TITA ICCV'25	3.447	1.038	0.667	0.571	4.644	0.769	0.529	0.466
UniFusion CVPR'26	3.597	1.054	0.650	0.545	5.379	0.791	0.557	0.498
SigFusion AAAI'26	3.613	1.109	0.701	0.585	5.040	0.736	0.485	0.513
T^{3}Fusion (MSRS only)	**3.765**	1.135	**0.726**	0.606	**5.832**	0.820	0.606	0.523
T^{3}Fusion (Joint)	3.755	**1.137**	0.721	**0.608**	5.827	**0.832**	**0.617**	0.522
Method	**MEFB**				**MFI-WHU**			
	AG↑	VIF↑	Q_{abf} ↑	Q_{cb} ↑	AG↑	VIF↑	Q_{abf} ↑	Q_{cb} ↑
U2Fusion TPAMI'22	4.124	1.239	0.568	**0.434**	6.624	1.004	0.591	0.571
TITA ICCV'25	5.366	1.383	0.695	0.380	7.501	1.201	0.692	0.737
UniFusion CVPR'26	4.270	1.455	0.707	0.413	8.005	1.221	**0.722**	**0.743**
SigFusion AAAI'26	5.452	1.179	0.333	0.401	6.341	0.761	0.299	0.395
T^{3}Fusion (MSRS only)	5.481	1.368	0.733	0.413	**8.494**	1.187	0.682	0.686
T^{3}Fusion (Joint)	**5.739**	**1.551**	**0.783**	0.432	8.320	**1.245**	0.705	0.715

F LIMITATIONS

Computational cost. Table S8 reports complete-method cost and throughput, including relation fitting and refinement for T^{3}Fusion.

Table S8: Efficiency comparison on an RTX 4090 using 256×256 input pairs, batch size 1, and FP32. One MAC counts as two FLOPs. FPS includes the complete fusion inference process.

Metric	TarDAL CVPR'22	CDDFuse CVPR'23	SHIP CVPR'24	CAF IJCAI'24	A2RNet AAAI'25	SAGE CVPR'25	HCLFuse NeurIPS'25	DPOFusion CVPR'26	ReCoFuse CVPR'26	U2Fusion TPAMI'22	TITA ICCV'25	UniFusion CVPR'26	SigFusion AAAI'26	T^{3}Fusion
Params (M)↓	0.30	1.19	0.55	0.15	10.61	**0.14**	27.33	778.40	155.47	0.66	1.39	27.04	22.13	1.07
GFLOPs↓	38.77	235.72	70.56	19.37	72.90	**8.53**	141.41	19069.68	1069.58	86.34	243.03	135.63	199.93	28.76
FPS↑	984.38	42.66	81.28	**1454.35**	178.57	1217.89	7.45	0.57	0.34	277.16	2.21	41.03	26.30	39.93

TTT solves 24 systems of size 17×17 per pair. Complete inference on an RTX 4090 at 256×256, batch one, FP32 takes 25.04 ms, compared with 15.90 ms without relational refinement, a 57.5% increase. Timing includes support statistics, prediction, and readout after warm-up and excludes weight loading, file I/O, external input transfer, and chrominance recovery.