# HYBRID FUSION: MINUTE-SCALE TRAINING FOR ZERO-SHOT CROSS-DOMAIN IMAGE FUSION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Image fusion integrates complementary information from multiple sources, yet most recent systems still learn to reconstruct every output pixel with patch training, long schedules, or large pretrained backbones. We propose Hybrid Fusion, a source-allocation formulation in which a U-Net predicts only a dense guidance map and a fixed Laplacian-pyramid kernel performs synthesis. The output therefore remains an explicit multi-scale mixture of the two sources, while the learnable task is reduced from image generation to deciding where each source should contribute. This design supports full-resolution optimization, reaches strong infrared-visible fusion quality after one to two epochs, improves downstream detection, and transfers without retraining to medical and video fusion. A unified single-GPU benchmark further shows that the practical advantage comes from rapid adaptation rather than a selectively cheap iteration. For accessibility, our Colab demo reproduces the one-epoch result in about five minutes on a free Google T4 GPU.

## 1 INTRODUCTION

Image fusion combines complementary measurements into a single representation for human inspection and downstream perception Albarqaan et al. (2024); Ma et al. (2019a). Infrared-visible fusion, for example, should retain thermal targets from infrared images without discarding the texture, color, and spatial context of visible images. Medical fusion must preserve functional responses and anatomical structure, while video fusion additionally faces changing content over time. These applications share a strict requirement: useful information should be transferred from the sources rather than invented by an unconstrained generator.

Classical multi-scale methods satisfy this faithfulness requirement through explicit decomposition and handcrafted coefficient rules Fu et al. (2020); Yao et al. (2023). They are efficient and interpretable, but a fixed rule cannot adapt its source preference to different objects, modalities, or scenes. Deep models learn this preference from data and have progressed from CNN and GAN formulations Zhang et al. (2020); Ma et al. (2019b); Xu et al. (2020a) to Transformer, diffusion, and large-prior systems Ma et al. (2022); Zhao et al. (2023a;b); Yi et al. (2024); Zhang et al. (2025). Their reported quality is strong, but the dominant formulation asks a network to synthesize the entire fused image. The model must simultaneously discover a fusion policy, reproduce low-frequency appearance, preserve high-frequency detail, and avoid artifacts. This broad output space encourages large backbones, cropped training, auxiliary priors, and long optimization schedules. It also creates a practical mismatch when a network trained on small patches is deployed on full-resolution images.

This motivates a simpler alternative. Fusion is fundamentally a decision about *which source should contribute at each location and scale*. We therefore retain learning only for this decision and delegate image construction to a fixed Laplacian-pyramid operator. A U-Net predicts a spatial allocation map, the map is resized across scales, and each fused coefficient is an explicit linear mixture of source coefficients. For a fixed allocation map, every pyramid coefficient is mixed from the two sources; this gives the network a constrained synthesis path before convergence. This constrained formulation makes full-resolution training practical and supplies a direct optimization path from fusion losses to source preference.

The resulting model converges in minutes rather than relying on a long schedule to first learn image reconstruction. After one pass over the 662-pair MSRS training set, it already obtains competitive



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.

VIF and $Q^{AB/F}$ and produces the strongest downstream YOLO detection result among the compared methods. MSRS-trained checkpoints also operate on PET-MRI, CT-MRI, SPECT-MRI, and video inputs without target-domain retraining. We evaluate these claims with a unified single-RTX-4090 reproduction benchmark, real wall-clock measurements, archived epoch-1/100 predictions, downstream detection, cross-domain tests, and extensive supplementary analyses. Registration is orthogonal to the present fusion operator; a concise misalignment study is included in the supplement rather than treated as a separate contribution.

Our contributions are:

- We formulate image fusion as *source allocation*: a learnable guidance network selects source contributions while a fixed Laplacian pyramid performs faithful pixel synthesis.
- The constrained task enables full-resolution training and useful performance after one epoch; the key result can be reproduced in about five minutes on a free Colab T4.
- We evaluate fusion quality and practical transfer through a unified IVF benchmark, downstream detection, medical and video tests, formulation and scale-map ablations, stability tests, and hardware/model-scaling studies.

## 2 Related Work

**Traditional and hybrid fusion.** Multi-scale decomposition separates images into frequency bands and combines them through deterministic coefficient rules Ma et al. (2019a); Fu et al. (2020). Such pipelines are inexpensive and make the source of each output component easier to trace, but a handcrafted saliency rule has limited ability to adapt across scenes. Hybrid methods such as HG-LPFN Yao et al. (2023) insert learned modules into a pyramid architecture, yet the network still participates in nonlinear reconstruction. RegionFuse also learns spatially adaptive fusion weights Xia et al. (2026). Our distinction from these methods is narrower and structural: the fixed pyramid is the complete coefficient-synthesis path, whereas learning is restricted to a single source-allocation policy.

**Reconstruction-based neural fusion.** CNN and GAN approaches learn image-level mappings or implicit fusion rules Ma et al. (2019b); Xu et al. (2020a). IFCNN Zhang et al. (2020) is attractive as a compact feed-forward framework, but its public release is primarily a pretrained inference package and does not expose a matching MSRS training recipe, which limits controlled retraining. Transformer designs such as SwinFusion and CDDFuse Ma et al. (2022); Zhao et al. (2023a) enlarge the receptive field and improve representation capacity, but their memory demand commonly leads to patch training and small inference batches. These methods optimize an image reconstruction space much larger than the source-selection space required by our formulation.

**Unified and large-prior models.** Recent systems pursue broader task coverage by importing increasingly strong priors. TC-MoA Zhu et al. (2024) routes task-specific adapters through a large MAE-ViT backbone. Its original recipe is designed for multi-task training, and the released configuration uses 448-pixel inputs; in our single-card adaptation the largest stable batch is two and 100 epochs require 213.3 minutes. DDFM Zhao et al. (2023b) replaces task-specific training with iterative diffusion sampling, so deployment cost is shifted from optimization to repeated denoising at inference. Text-IF and DTPF Yi et al. (2024); Zhang et al. (2025) add large frozen text/vision priors to obtain semantic guidance, but this introduces external checkpoints, additional memory, and longer adaptation schedules. These systems demonstrate broad capability, yet they are cumbersome when the practical requirement is simply a reliable fusion operator that can be retrained or deployed on modest hardware. We instead study the opposite design point: retain an explicit synthesis rule and learn only the source-allocation policy. Repository-derived settings, dependency notes, reproduction exceptions, and full efficiency measurements are given in the supplementary material.

## 3 Method

**Problem formulation.** Given a visible image $I_{vi} \in \mathbb{R}^{H\times W\times 3}$ and an infrared image $I_{ir} \in \mathbb{R}^{H\times W}$, we seek a fused color image that preserves visible appearance and infrared saliency. We

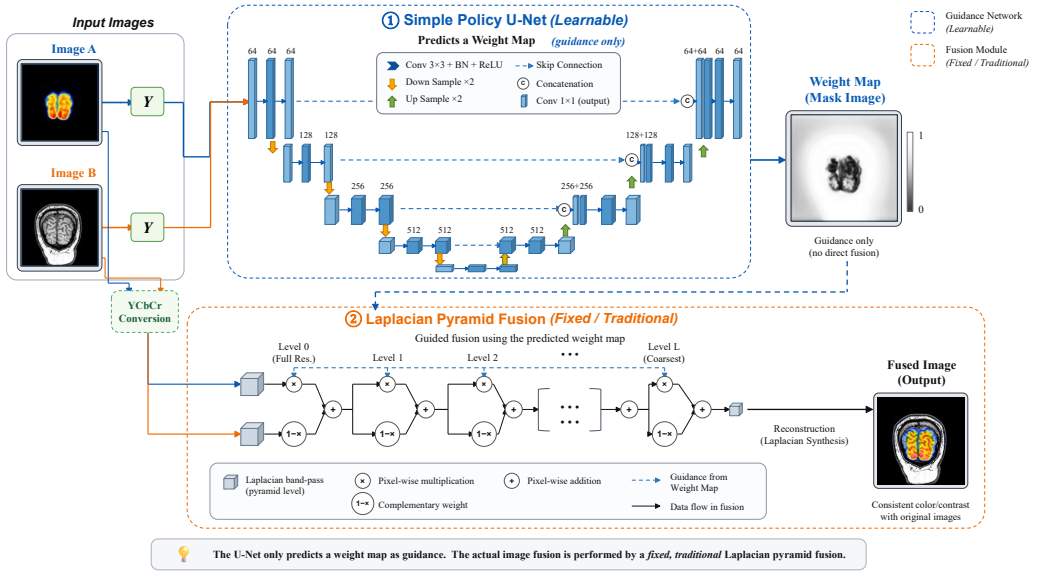


Figure 1: Hybrid Fusion separates guidance learning from synthesis. A U-Net predicts a dense source-allocation map from visible luminance and infrared inputs. A fixed Laplacian-pyramid kernel downsamples this map across scales, linearly mixes source coefficients, reconstructs fused luminance, and restores the visible chrominance.

convert $I_{vi}$ to YCbCr, fuse only its luminance $I_{vi}^{Y}$ with $I_{ir}$, and restore the untouched visible chrominance after fusion. This prevents the network from learning an unnecessary color-generation problem and guarantees that color remains anchored to the visible source.

**Guidance rather than reconstruction.** A four-stage U-Net receives the concatenated luminance pair and predicts $\mu \in [0,1]^{H\times W}$. We use a conventional encoder-decoder because the required output is a dense, spatially aligned policy rather than a semantic representation. Skip connections preserve local boundaries, while the bottleneck supplies enough context to distinguish thermal targets from visible texture. More elaborate attention backbones are possible, but they reintroduce the memory pressure that motivates patch-based training; model-scaling results in the supplement show that the formulation, rather than a particular backbone size, provides the early convergence.

**Fixed multi-scale synthesis.** The two luminance inputs are decomposed by the same Laplacian pyramid. The guidance map is resized to level $k$ and controls

$$L_f^k = (1-\mu^k) \odot L_{vi}^k + \mu^k \odot L_{ir}^k, \tag{1}$$

where $L^k$ denotes a pyramid coefficient map. Collapsing the fused pyramid yields $I_f^Y$. For a fixed $\mu^k$, Equation 1 linearly interpolates source coefficients; the complete input-to-output mapping remains nonlinear because the guidance map depends on the inputs. The pyramid supplies the reconstruction mechanism before learning begins, so optimization need only discover a useful source preference.

We predict one full-resolution map and resize it at all pyramid levels rather than predicting four independent maps. Separate scale-specific maps add parameters and permit inconsistent decisions across bands, but our controlled experiment finds essentially identical final VIF/$Q^{AB/F}$ for the one-map and four-map variants. We therefore retain the simpler shared policy; the final comparison is reported in the supplementary scale-map ablation. Likewise, four pyramid levels are sufficient to separate coarse appearance from high-frequency detail at the evaluated resolutions. Increasing decomposition complexity did not produce a measurable gain, whereas the four-level form keeps the synthesis rule and its interpretation compact.

**Unsupervised objective.** No fused ground truth is required. We optimize

$$\mathcal{L} = \lambda_{max}\mathcal{L}_{max} + \lambda_{grad}\mathcal{L}_{grad} + \lambda_{ssim}\mathcal{L}_{ssim} + \lambda_{con}\mathcal{L}_{con}, \tag{2}$$

Table 1: Reported downstream object detection on fused MSRS images. Fusion metrics can be sensitive to contrast and source copying; detection provides a complementary task-level view. All rows use the reported YOLO evaluation protocol.

| Metric | Visible | U2Fusion | IFCNN | SwinFusion | SuperFusion | CDDFuse | PSLPT | Text-IF | DTPF | Ours |
|---|---|---|---|---|---|---|---|---|---|---|
| mAP@50 | .7955 | .9280 | .9210 | .9110 | .9210 | .9220 | .9150 | .9236 | .9368 | **.9518** |
| mAP@50–95 | .5411 | .6570 | .6510 | .6410 | .6730 | .6620 | .5660 | .7052 | .7058 | **.7233** |

where $\mathcal{L}_{max}$ preserves the stronger source intensity, $\mathcal{L}_{grad}$ follows the strongest Sobel response, $\mathcal{L}_{ssim}$ maintains structure relative to both sources, and $\mathcal{L}_{con}$ regularizes intensity deviation. These terms supervise the allocation decision from complementary perspectives; none asks the network to reproduce a target fused image. A 240-point search reveals a broad high-performing region, and the supplement reports the complete loss grid, guidance-map evolution, shared/four-map comparison, and long-run trajectories. The stable region is important because the claimed rapid convergence should not depend on a single fragile weight tuple.

# 4 Experiments

### Setup and Reporting Protocol

We use MSRS Tang et al. (2022b), M3FD Liu et al. (2022), and RoadScene Xu et al. (2020b) for infrared-visible fusion; PET-MRI, CT-MRI, and SPECT-MRI for medical fusion; and M3SVD Tang et al. (2025) for zero-shot video evaluation. Metrics include EN Roberts et al. (2008), MI Ma et al. (2019a), VIF Han et al. (2013), $Q^{AB/F}$ Piella & Heijmans (2003), and SSIM Wang et al. (2004). Following common fusion evaluation, SSIM and VIF may sum preservation scores against both inputs and can therefore exceed one. Our U-Net is optimized with AdamW at $10^{-4}$ on full-resolution MSRS images.

We compare official implementations in a unified single-GPU benchmark. All trainable baselines use the same 662-pair MSRS training manifest, one RTX 4090, one fixed seed, and 100 epochs. We preserve the released model, loss, optimizer, scheduler, pretraining, and preprocessing, while selecting the largest stable single-GPU batch size. Reported times are real wall-clock measurements with at most one job per GPU. Repository configurations are treated as authoritative when paper and code settings differ.

Three methods require special reporting. IFCNN releases pretrained inference code but no defensible MSRS training loop; DDFM is a pretrained zero-shot diffusion method without an MSRS training epoch; and our SwinFusion migration completed but produced degenerate low-dynamic-range outputs. We therefore quote their published MSRS quality values in Table 2, marked by †, rather than presenting failed local outputs as method capability. SwinFusion's 100-epoch cost is retained from the completed single-GPU run; IFCNN and DDFM have no comparable 100-epoch cost. Full configurations, repository-derived schedules, failure evidence, and cross-dataset metrics are provided in the supplement.

### Infrared-Visible Fusion and Detection

Table 1 provides a downstream IVF comparison of whether fusion preserves cues useful to a downstream detector. DTPF and Text-IF are the strongest detector baselines, but both rely on large external semantic priors and require 171 and 97 minutes, respectively, for the unified 100-epoch adaptation. TC-MoA is still heavier: its MAE-L backbone, 448-pixel input, and batch size two lead to 213.3 minutes, while its first-epoch VIF/$Q^{AB/F}$ remains near the warm-up regime. At the other extreme, PSLPT finishes 100 epochs quickly because it starts from released stage-2 weights and updates only selected branches, yet its mAP@50–95 is .5660. SuperFusion and CDDFuse provide more conventional efficiency–quality compromises, but neither matches the detector accuracy of the strongest prior-based models.

Our model uses no external semantic encoder and reaches .9518 mAP@50 and .7233 mAP@50–95. The gain over DTPF is moderate rather than an order-of-magnitude separation, but it is obtained with a simpler dependency stack and a useful first-epoch solution. Table 2 supplies the complemen-

Table 2: MSRS fusion quality and optimization cost. Unmarked entries are obtained with our unified evaluator after local training. † denotes quality quoted from the official paper because a valid matching local training result was unavailable; ‡ denotes measured/estimated cost from the completed official-code adaptation despite unusable quality. Times are single-RTX-4090 wall-clock values.

| | Epoch 1 | | Epoch 100 / official | | | |
|---|---|---|---|---|---|---|
| Method | VIF↑ | $Q^{AB/F}$ ↑ | VIF↑ | $Q^{AB/F}$ ↑ | Batch | 100-epoch time↓ |
| U2Fusion Xu et al. (2020a) | .501 | .363 | .522 | .389 | 8 | 47 min |
| SuperFusion Tang et al. (2022a) | .841 | .605 | .986 | .615 | 64 | 77.7 min |
| CDDFuse Zhao et al. (2023a) | .765 | .552 | .811 | .623 | 4 | 92 min |
| PSLPT Wang et al. (2024) | .609 | .358 | 1.036 | .684 | 64 | **6.6 min** |
| TC-MoA Zhu et al. (2024) | .036 | .030 | .877 | .634 | 2 | 213.3 min |
| Text-IF Yi et al. (2024) | .746 | .523 | 1.019 | .692 | 4 | 97 min |
| DTPF Zhang et al. (2025) | .712 | .512 | 1.004 | .682 | 4 | 171 min |
| IFCNN Zhang et al. (2020) | – | – | .579† | .479† | – | – |
| SwinFusion Ma et al. (2022) | – | – | .825† | .558† | 8 | 32.3 min‡ |
| DDFM Zhao et al. (2023b) | – | – | .810† | .580† | – | zero-shot |
| Ours | **.996** | **.678** | **1.079** | **.721** | 4 | 58 min |




Figure 2: Qualitative infrared-visible fusion comparison.

tary VIF/$Q^{AB/F}$ view. These metrics are less decisive than detection because contrast amplification or source copying can change their ranking, so we use them to verify balanced source preservation rather than as the sole performance claim. Complete EN, MI, SSIM, cross-dataset values, and repository settings and full evaluation tables are provided in the supplement.

Figure 2 helps explain the detector ranking. Some outputs preserve strong thermal contrast but weaken visible boundaries, while others retain texture at the cost of target prominence. The allocation result keeps both cues without requiring a separate restoration or semantic-guidance stage. Three additional full-width comparisons in the supplement include difficult low-light, small-target, and cluttered scenes, rather than only the example shown here.

### Medical and Cross-Domain Generalization

The medical comparison exposes a different trade-off. DTPF is strong on PET and SPECT but depends on a large pretrained semantic prior and target-domain training; Text-IF has a similar dependency profile. By contrast, the MSRS model is evaluated without medical optimization and remains competitive on all three modality pairs. Its PET VIF reaches 1.025, while the 10-epoch MSRS

Table 3: Medical fusion comparison using VIF and $Q^{AB/F}$. Competitor values are quoted from the cited papers. "ZS" denotes our MSRS-trained model evaluated without medical training; "trained" uses corresponding medical data. Full SSIM, EN, and MI results are in the supplement.

| | PET-MRI | | CT-MRI | | SPECT-MRI | |
|---|---|---|---|---|---|---|
| Method | VIF | $Q^{AB/F}$ | VIF | $Q^{AB/F}$ | VIF | $Q^{AB/F}$ |
| PSLPT Wang et al. (2024) | .548 | .373 | .502 | .432 | .359 | .325 |
| EMFusion Xu & Ma (2021) | .685 | .783 | .552 | .475 | .665 | .692 |
| MSRPAN Fu et al. (2021) | .581 | .799 | .436 | .455 | .525 | .560 |
| SwinFusion Ma et al. (2022) | .703 | .683 | .522 | .545 | .744 | .720 |
| Zero Lahoud & Süsstrunk (2019) | .635 | .774 | .320 | .582 | .582 | .681 |
| CDDFuse Zhao et al. (2023a) | .650 | .765 | .526 | .530 | .786 | .719 |
| Text-IF Yi et al. (2024) | .640 | .690 | .542 | .561 | .747 | .715 |
| DTPF Zhang et al. (2025) | .909 | .782 | .641 | .657 | .888 | .746 |
| Ours (2e, ZS) | **1.025** | .772 | **.807** | .561 | .840 | .732 |
| Ours (10e, ZS) | .911 | .796 | .701 | .622 | **.965** | .776 |
| Ours (trained) | .937 | **.806** | .700 | **.664** | .956 | **.780** |

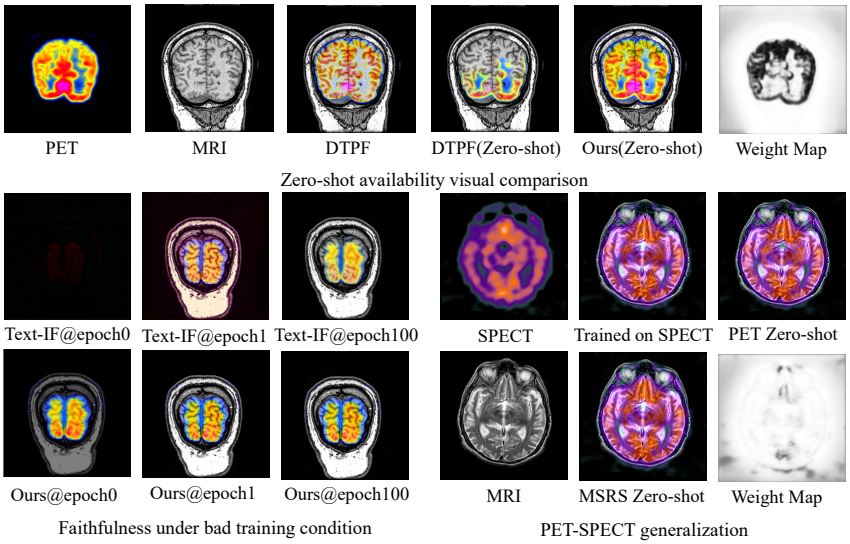


Figure 3: Zero-shot medical transfer. MSRS-trained checkpoints preserve source structure on unseen modalities; target-domain and cross-domain variants are shown for comparison.

model reaches .965 VIF on SPECT. The point is not that a natural-image model dominates every target-trained specialist, but that the fixed synthesis path remains usable under a severe domain shift. Full EN, MI, SSIM, additional qualitative examples, and cross-modality trajectories appear in the supplement.

Figure 3 visualizes the same property. The output is assembled from source pyramids, so domain shift can change allocation quality while the synthesis operator remains fixed. Table 5 further shows that the short schedule is not supported by one favorable seed: variation decreases by epoch 10 on most listed tasks. Registration remains a separate preprocessing problem; initialization behavior, misalignment examples, and the complete 100-run statistics are kept in the supplement so that this qualification does not dominate the main argument.

Table 4: Cross-domain medical generalization. The MSRS row uses the 10-epoch checkpoint on SPECT and 2-epoch checkpoint on PET; it does not represent one shared checkpoint.

| Training data | Eval. SPECT-MRI | | | Eval. PET-MRI | | |
|---|---|---|---|---|---|---|
| | SSIM | VIF | $Q^{AB/F}$ | SSIM | VIF | $Q^{AB/F}$ |
| SPECT-MRI | **1.270** | .956 | **.780** | **1.259** | .898 | .795 |
| PET-MRI | 1.233 | .861 | .754 | 1.253 | .937 | **.806** |
| MSRS (10e / 2e) | 1.214 | **.965** | .776 | 1.118 | **1.025** | .772 |

Table 5: Short-run stability over 100 independent trainings ($Q^{AB/F}$ mean$\pm$std). Full VIF and SSIM statistics are in the supplement.

| Infrared-visible | | | Medical | | |
|---|---|---|---|---|---|
| Dataset | Epoch 2 | Epoch 10 | Dataset | Epoch 2 | Epoch 10 |
| MSRS | .6665$\pm$.0072 | .6881$\pm$.0054 | PET-MRI | .7785$\pm$.0119 | .7950$\pm$.0022 |
| M3FD | .5642$\pm$.0445 | .6508$\pm$.0154 | CT-MRI | .5739$\pm$.0066 | .6100$\pm$.0075 |
| RoadScene | .6078$\pm$.0218 | .6406$\pm$.0099 | SPECT-MRI | .7593$\pm$.0168 | .7746$\pm$.0013 |

### Ablations and Convergence

The early-epoch rows in Table 2 separate per-iteration speed from optimization speed. PSLPT has the shortest measured 100-epoch wall-clock time, but it starts from a task-specific second-stage checkpoint. TC-MoA, Text-IF, and DTPF inherit large pretrained priors, yet their first-epoch scores remain below the allocation model and their complete schedules are longer. The comparison therefore does not claim that every iteration of our network is the cheapest; it shows that a simpler prediction target reaches a usable operating point with fewer passes over MSRS.

Table 6 probes the source of the optimization difference. The fixed pyramid alone is stable but cannot adapt its decision rule. The direct-reconstruction U-Net improves slowly and gives an imbalanced $Q^{AB/F}$ on MSRS. Its M3FD run exhausted memory under the archived implementation; because the exact memory allocation is unavailable, that row is not used as evidence of an inherent memory advantage. The allocation model reaches a stronger short-run operating point in this comparison. Additional direct-fusion outputs and initialization comparisons are shown in the supplement.

A separate map for each pyramid level produces nearly identical final performance (Supplementary Table S8). We therefore keep one interpretable map and resize it across scales. The result also argues against attributing performance to hidden scale-specific capacity.

The grid search in Supplementary Figure S7 shows a contiguous high-performing region rather than a single tuned point. The supplementary hardware table demonstrates that the same short schedule remains usable beyond the RTX 4090, including a free Colab T4. Model scaling reveals a conventional trade-off: the 17.26M model is slower than sub-million-parameter variants at inference, but it reaches the useful regime in fewer epochs. The supplement reports every scaling row, the 500-epoch trajectories, and the full stability statistics rather than presenting only the favorable operating point.

### Zero-Shot Video Fusion

To test whether transfer is limited to still images, we apply the MSRS model frame by frame to M3SVD Tang et al. (2025). The 10-epoch model obtains EN 7.343, MI 4.176, and SD 51.660, while the video-specific VideoFusion model reports 7.235/4.212/53.644. The results are mixed rather than uniformly superior: our entropy is higher, whereas VideoFusion retains slightly higher MI and SD. This is expected because the frame-wise model has no temporal module. The complete table, including DDFM, TC-MoA, and Text-IF, is provided in the supplement.

Table 6: Formulation ablation. Direct U-Net reconstruction uses the same learnable backbone but predicts fused pixels.

| | MSRS | | M3FD | | RoadScene | |
|---|---|---|---|---|---|---|
| Method | VIF | $Q$ | VIF | $Q$ | VIF | $Q$ |
| Laplacian | .862 | .624 | .745 | .580 | .703 | .527 |
| Heuristic map | .887 | .622 | .748 | .573 | .714 | .519 |
| Direct U-Net, 200e | .981 | .362 | OOM | OOM | – | – |
| Direct U-Net, 1000e | .868 | .570 | OOM | OOM | – | – |
| Hybrid, 2e | .997 | .674 | .827 | .605 | .787 | **.581** |
| Hybrid, 10e | **1.032** | **.694** | **.924** | **.664** | **.808** | .556 |

### Why the Optimization Converges Quickly

The empirical timing results are best understood together with the formulation. A reconstruction model must learn a mapping from two input images to an unconstrained output image. Even with an unsupervised objective, the network is responsible for low-frequency appearance, high-frequency texture, local contrast, and the interaction between modalities. The loss can often be reduced through many visually different outputs, so optimization must discover both a synthesis mechanism and a fusion policy. Patch training reduces memory use but also exposes the model to a different spatial context from full-resolution inference.

Hybrid Fusion removes most of this search space. The Laplacian kernel already specifies how a valid image is decomposed, combined, and reconstructed. The network only selects a mixing coefficient. Equation 1 also gives a direct gradient path from the fusion loss to the allocation map at every scale. The initial model therefore starts from a meaningful family of outputs and can improve rapidly by learning where infrared saliency should dominate and where visible texture should be retained. This interpretation is consistent with the early scores in Table 2, the direct-reconstruction comparison in Table 6, and a broad region of useful loss weights in Supplementary Figure S7. These observations support the design rationale but do not by themselves isolate every factor in the optimization speed.

The advantage should not be confused with claiming the cheapest individual iteration. PSLPT completes its restricted second-stage update especially quickly, and compact feed-forward baselines can execute inexpensive cropped epochs. Conversely, TC-MoA spends 213.3 minutes adapting MAE-L adapters on a single card, while Text-IF and DTPF retain external semantic encoders and take 97 and 171 minutes in the same 100-epoch accounting. DDFM avoids MSRS training but pays for iterative diffusion during every inference. These methods are reasonable when task unification, text interaction, or diffusion priors are the objective; they are unnecessarily elaborate when the requirement is a compact, retrainable fusion operator. Reporting epoch-1 quality, 100-epoch quality, initialization, and wall-clock together makes this distinction explicit.

### Faithfulness, Interpretability, and Transfer

The learned guidance map is directly interpretable as a source preference. Bright regions indicate increased infrared contribution and dark regions preserve visible luminance. Unlike a latent feature tensor, the map has the same spatial support as the input and participates in the explicit linear rule in Eq. 1. Inspecting the map therefore reveals whether the network is emphasizing a thermal target, suppressing an uninformative region, or retaining visible edges. Supplementary visualizations show that early training produces a soft allocation field, whereas extended in-domain optimization produces a sharper and more specialized map.

This constrained output space is also useful under domain shift. A medical image or video frame may differ greatly from MSRS in texture, intensity distribution, and semantics, yet the pyramid operation itself remains defined. The model can be imperfect without becoming undefined: every coefficient is still assembled from the two sources. The zero-shot medical results in Table 3 and the M3SVD experiment support this mechanism. In particular, the natural-scene model remains competitive with methods trained on the target medical domain, while cross-training between PET and

SPECT preserves strong VIF and $Q^{AB/F}$. The fixed synthesis operator contributes to this transfer; the present comparisons do not isolate how much additional benefit comes from the learned map.

The explicit allocation map also helps interpretation. When a score changes, the cause can be related to source allocation rather than an opaque generative prior. The source-copy baselines in the complete supplement illustrate why this matters: copying one modality may look favorable under a particular metric but fails to integrate complementary evidence. Our combination of multi-metric evaluation, downstream detection, qualitative comparisons, and cross-domain tests reduces reliance on any single measurement.

### Robustness and Practical Accessibility

Table 5 summarizes 100 independent short runs. The remaining spread is largest on M3FD at epoch 2 and becomes substantially smaller by epoch 10; medical $Q^{AB/F}$ is already stable after two epochs. The complete VIF/SSIM statistics and 500-epoch trajectories in the supplement show both the early saturation and the later in-domain specialization, preventing the short-run claim from resting on a single seed or a selectively chosen metric.

The hardware experiment targets reproducibility rather than peak throughput. An RTX 4060 laptop reaches VIF 1.001 and $Q^{AB/F}$ .681 after a short run, while a free-tier Colab T4 reaches 1.016/.691 in about five minutes for one epoch. This is the practical contrast with methods that require large external checkpoints, sliding-window inference, or multi-stage repository setup: the core result can be inspected without specialized infrastructure. The supplementary model-scaling table adds an important qualification: very small guidance networks are faster per pass but need more optimization and still trail the default model. Exact device settings and the full scaling rows are listed in the supplement.

## 5 Limitations

This work studies general-purpose fusion under the standard assumption of spatially aligned and complementary source images. More specialized settings, such as severe misalignment or sensor-specific degradation, may require additional preprocessing or domain-specific constraints. Representative cases and further discussion are provided in the supplementary material.

## 6 Conclusion

Hybrid Fusion separates the learnable fusion policy from image synthesis: a U-Net predicts source allocation and a fixed Laplacian pyramid constructs the output. The experiments show where this simpler design is useful. It reaches a competitive operating point after one or two epochs, gives the strongest reported downstream detection in our comparison, and applies to medical and video inputs without changing the synthesis rule. The unified benchmark also clarifies the trade-off: several baselines obtain good fusion scores through larger priors, multi-stage initialization, or iterative inference, whereas the proposed formulation favors a smaller dependency stack and rapid retraining.

## AI Use Statement

Generative AI tools were used to aid and polish writing, retrieve and discover related work, assist with research ideation and execution, and draft sections of this paper. They also assisted with adapting the manuscript to the ICLR 2027 template and checking figure and table placement and PDF layout. The authors take responsibility for the final text, citations, experiments, and claims.

## REFERENCES


Hessah Albarqaan, Rongjun Qin, and Yang Teng. Image Fusion in Remote Sensing: An Overview and Meta-Analysis. *Photogrammetric Engineering & Remote Sensing*, 90(12):755–775, December 2024. doi: 10.14358/PERS.24-00110R1. URL https://doi.org/10.14358/PERS.24-00110R1.

Jun Fu, Weisheng Li, Jiao Du, and Bin Xiao. Multimodal medical image fusion via laplacian pyramid and convolutional neural network reconstruction with local gradient energy strategy. *Computers in Biology and Medicine*, 126:104048, 11 2020. doi: 10.1016/j.compbiomed.2020.104048.

Jun Fu, Weisheng Li, Jiao Du, and Yuping Huang. A multiscale residual pyramid attention network for medical image fusion. *Biomedical Signal Processing and Control*, 66:102488, 2021. ISSN 1746-8094. doi: https://doi.org/10.1016/j.bspc.2021.102488. URL https://www.sciencedirect.com/science/article/pii/S1746809421000859.

Yu Han, Yunze Cai, Yin Cao, and Xiaoming Xu. A new image fusion performance metric based on visual information fidelity. *Information Fusion*, 14(2):127–135, 2013. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2011.08.002. URL https://www.sciencedirect.com/science/article/pii/S156625351100056X.

Fayez Lahoud and Sabine Süsstrunk. Zero-learning fast medical image fusion. In *2019 22th International Conference on Information Fusion (FUSION)*, pp. 1–8, 2019. doi: 10.23919/FUSION43075.2019.9011178.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5802–5811, 2022.

Jiayi Ma, Yong Ma, and Chang Li. Infrared and visible image fusion methods and applications: A survey. *Information Fusion*, 45:153–178, 2019a. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2018.02.004. URL https://www.sciencedirect.com/science/article/pii/S1566253517307972.

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, 2019b. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2018.09.004. URL https://www.sciencedirect.com/science/article/pii/S1566253518301143.

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.

G. Piella and H. Heijmans. A new quality metric for image fusion. In *Proceedings 2003 International Conference on Image Processing (Cat. No.03CH37429)*, volume 3, pp. III–173, 2003. doi: 10.1109/ICIP.2003.1247209.

J. Wesley Roberts, Jan A. van Aardt, and Fethi Babikker Ahmed. Assessment of image fusion procedures using entropy, image quality, and multispectral classification. *Journal of Applied Remote Sensing*, 2(1):023522, 2008. doi: 10.1117/1.2945910. URL https://doi.org/10.1117/1.2945910.

Linfeng Tang, Yuxin Deng, Yong Ma, Jun Huang, and Jiayi Ma. Superfusion: A versatile image registration and fusion network with semantic awareness. *IEEE/CAA Journal of Automatica Sinica*, 9(12):2121–2137, 2022a.

Linfeng Tang, Jiteng Yuan, Hao Zhang, Xingyu Jiang, and Jiayi Ma. MSRS: Multi-Spectral Road Scenarios for Practical Infrared and Visible Image Fusion. https://github.com/Linfeng-Tang/MSRS, 2022b.

Linfeng Tang, Yeda Wang, Meiqi Gong, Zizhuo Li, Yuxin Deng, Xunpeng Yi, Chunyu Li, Han Xu, Hao Zhang, and Jiayi Ma. Videofusion: A spatio-temporal collaborative network for multi-modal video fusion and restoration. *arXiv preprint arXiv:2503.23359*, 2025.

Wu Wang, Liang-Jian Deng, and Gemine Vivone. A general image fusion framework using multi-task semi-supervised learning. *Information Fusion*, pp. 102414, 2024.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19539–19548, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Xia_RegionFuse_Region-Adaptive_Pixel_Distribution_Learning_for_Infrared_and_Visible_Image_CVPR_2026_paper.html.

Han Xu and Jiayi Ma. Emfusion: An unsupervised enhanced medical image fusion network. *Information Fusion*, 2021.

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, 2020a.

Han Xu, Jiayi Ma, Zhuliang Le, Junjun Jiang, and Xiaojie Guo. Fusiondn: A unified densely connected network for image fusion. In *proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020b.

Jiaxin Yao, Yongqiang Zhao, Yuanyang Bu, Seong G. Kong, and Jonathan Cheung-Wai Chan. Laplacian pyramid fusion network with hierarchical guidance for infrared and visible image fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(9):4630–4644, 2023. doi: 10.1109/TCSVT.2023.3245607.

Xunpeng Yi, Han Xu, Hao Zhang, Linfeng Tang, and Jiayi Ma. Text-if: Leveraging semantic text guidance for degradation-aware and interactive image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Ran Zhang, Xuanhua He, Ke Cao, Liu Liu, Li Zhang, Man Zhou, and Jie Zhang. Distilling textual priors from llm to efficient image fusion, 2025. URL https://arxiv.org/abs/2504.07029.

Yu Zhang, Yu Liu, Peng Sun, Han Yan, Xiaolin Zhao, and Li Zhang. Ifcnn: A general image fusion framework based on convolutional neural network. *Information Fusion*, 54:99–118, 2020.

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 IEEE/CVF conference on computer vision and pattern recognition*, pp. 5906–5916, 2023a.

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 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8082–8093, October 2023b.

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