Adding Thermal Awareness to Visual Systems in Real-Time via Distilled Diffusion Models

Yuchen Guo[1*] Junli Gong[2] Wenjun Dong[1]
[1]Northwestern University [2]Northeastern University
yuchenguo2027@u.northwestern.edu

Abstract

Purely RGB-based vision models often fail to provide reliable cues in challenging scenarios such as nighttime and fog, leading to degraded performance and safety risks. Infrared imaging captures heat-emitting sources and provides critical complementary information, but existing high-fidelity fusion methods suffer from prohibitive latency, rendering them impractical for real-time edge deployment. To address this, we propose ***FusionProxy****, a real-time image fusion module designed as a fully independent, plug-and-play component with diffusion level quality. FusionProxy exploits two complementary statistics of a teacher sample ensemble: per-pixel variance in raw image space, used to weight pixel-level supervision, and per-pixel variance inside frozen foundation backbones, used to route feature-level alignment spatially. Once trained, FusionProxy can be directly integrated into any visual perception system without joint optimization. Extensive experiments demonstrate that our method achieves superior performance on static recognition tasks and significantly enhances robustness in dynamic tasks, including closed-loop autonomous driving. Crucially, FusionProxy achieves real-time inference speeds on diverse platforms, from high-end GPUs to commodity hardware, providing a flexible and generalizable solution for all-day perception. The source code will be available.*

1. Introduction

Human visual perception, and most of the machine vision we have built (*e.g.*, autonomous vehicles and surveillance), operates within the visible spectrum (*i.e.*, RGB imaging), which leaves both vulnerable in conditions of degraded illumination such as nighttime, fog and glare. To compensate, modern perception systems augment RGB cameras with active sensors such as LiDAR, radar, and sonar, which emit signals and measure their return. Active sensing provides additional information about the surrounding scene for situational awareness, but it has a fundamental scalability limitation: emitted signals interfere with one another as the density of intelligent embodied agents scales up [18, 27]; cost, power, and form-factor budgets per unit further restrict deployment. A scalable substitute should therefore be *passive*, drawing information from signals that already exist in the environment rather than ones it generates [2].

Thermal infrared radiation is the natural candidate. Every object above absolute zero emits thermal radiation, making heat an omnipresent signal that requires no transmission, no synchronization, and no coordination across agents. Thermal infrared is also uniquely suited to integration with the existing visual stack. Unlike LiDAR's point clouds, radar's range-Doppler maps, or event cameras' asynchronous spike streams, all of which demand new processing pipelines, thermal cameras produce a 2D pixel grid, the same format as an RGB image. Image fusion [1, 44] is the natural mechanism that exploits this shared format alignment, it aims to produce a single fused image that combines the thermal information of infrared image I_{IR} with the structural and textural information of visible image I_{VIS} [13, 38]. Because the output remains an RGB-format image, it serves a dual role: *machine-side*, it is consumed as a drop-in replacement for I_{VIS} by any frozen RGB-pretrained model, including detectors, segmenters, vision-language models, and driving policies; *human-side*, it is directly viewable on the same displays, dashboards, and monitors that already mediate human-RGB interaction. This dual-use property is what makes image fusion the right integration mechanism for adding thermal awareness to the existing RGB-based visual stack.

Despite this conceptual fit, no current image fusion method makes the integration practical. The frontier of fusion quality is occupied by diffusion-based methods such as DDFM [43], Mask-DiFuser [25], Text-IF [35], and ControlFusion [26], which produce visually faithful fusions through iterative sampling. Iterative sampling, however, places these methods at one frame every several seconds on server-class GPUs. Although one-step diffusion [30] minimizes inference steps, the distilled generator retains

*Corresponding author.

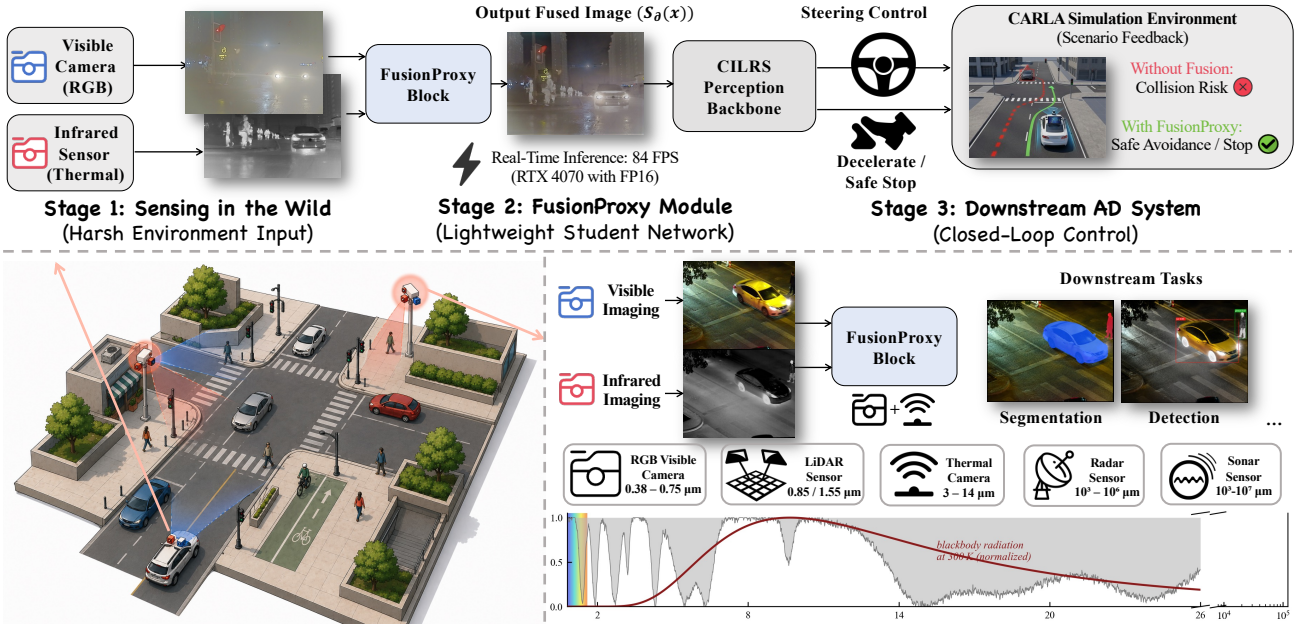

Figure 1. **FusionProxy: a real-time, plug-and-play fusion module that adds thermal awareness to any frozen RGB-pretrained visual system.** *Top:* End-to-end closed-loop autonomous driving pipeline. Under degraded visibility (e.g., fog, glare, night), the unmodified policy avoids collisions that would occur with RGB-only input, lifting closed-loop driving success. *Bottom-left:* Thermal infrared is a passive sensing modality that scales without interference and reveals heat-emitting objects invisible to RGB under low light or scattered illumination. *Bottom-right:* The fused output is consumable by any frozen downstream model such as detectors, segmenters.

the heavy architecture of original diffusion models, which is incompatible with any real-time perception pipeline. At the opposite end, real-time methods such as TarDAL [11] sacrifice fusion quality substantially, falling well behind the diffusion-quality fusions that motivate the integration in the first place. We therefore ask: ***can we combine the fusion quality of diffusion models with the inference speed required for commodity deployment?***

We answer this question with **FusionProxy**, a plug-and-play and real-time fusion module with respect to downstream perception models. On the one hand, we construct a diffusion teacher ensemble by drawing multiple samples from two complementary pre-trained diffusion fusion models per training image, yielding per-pixel sample statistics in both image space and foundation feature space. On the other hand, we introduce a dual-signal distillation loss that turns these statistics into supervision: image-space variance weights pixel-level matching against the ensemble mean, while feature-space variance routes feature-level alignment across a panel of frozen foundation backbones, so that each region is supervised by the backbone locally most informative about it. Based on the above, we distil the ensemble into a deterministic single-pass student whose architecture is decoupled from the teacher's denoising network, allowing lightweight backbones to inherit diffusion-grade fusion quality. Once trained, FusionProxy can be deployed alongside an infrared sensor as a perception front-end, enabling immediate integration with downstream visual systems without requiring any retraining. Our contributions are as follows:

- We propose **FusionProxy**, a plug-and-play fusion module that distills a diffusion teacher into a lightweight student, achieving diffusion-level fusion quality under strict real-time constraints on commodity hardware.
- We introduce a dual-signal distillation framework in which a single diffusion teacher ensemble simultaneously drives uncertainty-weighted pixel supervision (via image-space sample variance) and spatially-adaptive multi-foundation alignment, with both signals derived from a single set of cached teacher forward passes.
- We validate FusionProxy through extensive image fusion benchmarks, plug-and-play deployment on frozen RGB-pretrained detection and segmentation models, and an end-to-end closed-loop autonomous driving pipeline.

2. Related Works

Image Fusion. Image Fusion aims to integrate complementary information from different imaging modalities into a single image [1, 4, 38, 44]. Autoencoder based methods focus on the elaborated reconstruction and fusion loss functions [5, 41, 42], while GAN-based image fusion methods aim to contrast two discriminators for fused im-

ages and both source images[11, 14]. Recently, a few studies have explored incorporating diffusion models into image[26, 34, 43] and introduced video fusion[5, 46]. However, existing approaches predominantly focus on static image fusion or handle video fusion with prohibitively slow inference, limiting their practical usability. In contrast, we aim for real-time video fusion with plug-and-play deployment capability.

Diffusion Models. Recently, diffusion models have achieved amazing performance across a broad range of tasks, including image synthesis[39], video generation[17, 40], and image restoration[10, 37]. Moreover, recent efforts on accelerating diffusion models through distillation[15, 23] have significantly improved their practicality. One-step diffusion for image fusion model[30] reduces the number of inference steps, but the distilled generator still inherits the computationally heavy architecture, resulting in non-negligible latency that hinders real-time deployment.

3. Methodology

3.1. Problem Formulation

Let $x = (I_{\mathrm{IR}}, I_{\mathrm{VIS}}) \in \mathcal{X}$ denote an aligned infrared-visible input pair, where $I_{\mathrm{IR}} \in \mathbb{R}^{H\times W}$ and $I_{\mathrm{VIS}} \in \mathbb{R}^{H\times W\times 3}$. A pre-trained diffusion fusion model T defines a conditional implicit distribution $p_T(y \mid x)$ over fused images $y \in \mathcal{Y} \subset \mathbb{R}^{H\times W\times 3}$, from which high-quality samples $y_T \sim p_T(y \mid x)$ can be drawn via iterative denoising, at a cost prohibitive for real-time deployment.

Our goal is to learn a lightweight, deterministic student $S_\theta : \mathcal{X} \to \mathcal{Y}$ that approximates the teacher in a single forward pass:

$$\hat{y} \;=\; S_\theta(x) \;\approx\; \mathbb{E}_{y\sim p_T(y|x)}[y]. \tag{1}$$

The expectation form encodes the deterministic-student requirement: a single input must map to a single, stable output, so that $\hat{y}$ is consumable by frozen RGB-pretrained downstream models $\mathcal{F}$ in place of I_{VIS}. We further require S_θ to run at $r_{\min} \geq 30$ FPS on commodity hardware.

Direct optimization of Eq. (1) is infeasible: the expectation has no closed form. We instead approximate it via Monte Carlo by drawing $\{y_T^{(n)}\}_{n=1}^{N} \overset{\text{iid}}{\sim} p_T(y \mid x)$ and minimizing

$$\theta^\star \;=\; \arg\min_\theta \; \mathbb{E}_x \; \mathcal{L}\big(S_\theta(x),\, \{y_T^{(n)}\}_{n=1}^{N}\big). \tag{2}$$

Section 3.2 instantiates the teacher and Section 3.3 develops $\mathcal{L}$.

3.2. Diffusion Teacher Ensemble

We instantiate the teacher in Eq. (2) as an ensemble drawn from two pre-trained diffusion fusion models, addressing the bias of any single teacher and providing the structured statistics that drive our distillation loss in Sec. 3.3.

Dual teachers. The single-teacher formulation in Eq. (2) is straightforwardly extended to multiple teachers by drawing samples from each in turn. We use two diffusion teachers with complementary training distributions. *DDFM* [43] is a Bayesian conditional diffusion model trained directly on IR-VIS pairs and provides domain-grounded thermal radiation modeling. *Mask-DiFuser* [25] is a modality-agnostic model trained on natural images via dual masked restoration and provides broad image priors and texture fidelity. Distilling from a single teacher would inherit that teacher's distributional bias; the two together cover complementary failure modes.

Sampling protocol. For each input $x = (I_{\mathrm{IR}}, I_{\mathrm{VIS}})$ we draw N DDIM samples [22] from each teacher, yielding the ensemble

$$\mathcal{Y}_T(x) \;=\; \{y_T^{(n)}\}_{n=1}^{2N}, \qquad y_T^{(n)} \sim p_{T(n)}(y \mid x), \tag{3}$$

where $T(n) \in \{\text{DDFM}, \text{Mask-DiFuser}\}$ indexes the source teacher of the n-th sample. We use $N = 4$ samples per teacher throughout (ablation in Sec. 4.5).

Ensemble statistics. The pointwise ensemble mean

$$\bar{y}_T(x) \;=\; \frac{1}{2N}\sum_{n=1}^{2N} y_T^{(n)}(x) \tag{4}$$

serves as a low-variance Monte Carlo estimate of the teacher expectation $\mathbb{E}_{y\sim p_T(y|x)}[y]$ targeted in Eq. (1). The per-pixel sample variance

$$\sigma_T^2(x,p) \;=\; \mathrm{Var}_n\big[y_T^{(n)}(x,p)\big] \tag{5}$$

quantifies the ambiguity of this estimate at pixel p: high variance flags pixels where the teacher distribution is multimodal in raw image space.

In addition to image-space statistics, we measure the same ensemble inside frozen foundation backbones $\{\Phi_k\}_{k=1}^{K}$. Let $\Phi_k(y_T^{(n)})_p$ denote the feature of Φ_k at spatial location p for the n-th teacher sample. The per-backbone, per-pixel feature variance is

$$v_k(x,p) \;=\; \mathrm{Var}_n\big[\Phi_k(y_T^{(n)})_p\big]. \tag{6}$$

This quantity captures how strongly backbone Φ_k responds to fusion ambiguity at pixel p, and serves as the routing signal for multi-foundation alignment in Sec. 3.3.

Caching. All teacher samples $\mathcal{Y}_T(x)$ and foundation features $\{\Phi_k(y_T^{(n)})\}$ are precomputed once per training image and cached, since both teachers and foundation backbones are frozen. Teacher inference and foundation extraction are therefore one-time amortized costs and do not appear in the per-iteration training budget.

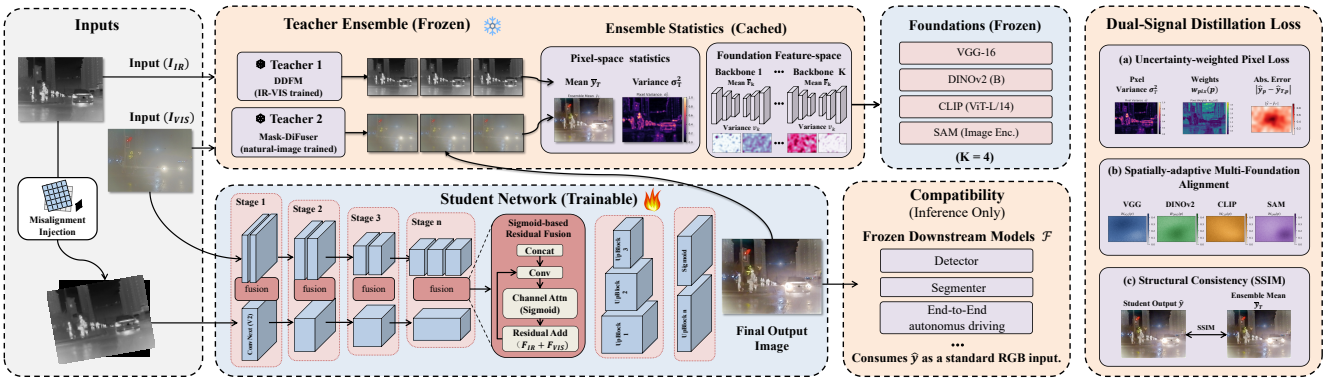

Figure 2. The FusionProxy framework. Dual diffusion teachers generate a sample ensemble whose mean and per-pixel variance drive uncertainty-weighted pixel supervision. The student is aligned to a panel of frozen foundation models via spatially-adaptive weights derived from per-backbone informativeness.

3.3. Dual-Signal Distillation Loss

The teacher ensemble defined in Sec. 3.2 provides a single pseudo-target $\bar{y}_T$ together with two complementary statistics: the pixel-space variance σ_T^2 and the feature-space variance v_k. We use the first to weight pixel-level supervision, the second to route feature-level supervision across multiple foundation backbones. Both signals are derived from the same cached forward passes through the teachers and incur no additional inference at training time.

Uncertainty-weighted pixel supervision. The pseudo-target $\bar{y}_T$ is reliable where the teacher ensemble agrees and unreliable where it disagrees. Forcing the student to match $\bar{y}_T$ uniformly would commit it to teacher noise in high-variance regions. We therefore weight pixel-level supervision by the inverse of the pixel-space variance $\sigma_T^2(x,p)$ from Eq. (5):

$$\begin{aligned} \mathcal{L}_{\text{pix}} &= \sum_p w_{\text{pix}}(x,p) \cdot \big\| S_\theta(x)_p - \bar{y}_T(x)_p \big\|_1, \\ w_{\text{pix}}(x,p) &= \frac{1/\big(\sigma_T^2(x,p)+\epsilon\big)}{\sum_{p'} 1/\big(\sigma_T^2(x,p')+\epsilon\big)}, \end{aligned} \tag{7}$$

where $\epsilon = 10^{-3}$ stabilizes the inverse and the weights are normalized over the spatial domain Ω so that $\sum_p w_{\text{pix}}(x,p) = 1$. Pixels where the teacher ensemble is internally consistent contribute proportionally more gradient; pixels where it is multimodal contribute less.

Spatially-adaptive multi-foundation alignment. Compatibility with downstream RGB-pretrained models (Sec. 3.1) requires that $S_\theta(x)$ remain consistent with the teacher ensemble in the feature spaces those models use. Since these feature spaces are heterogeneous, we align the student to a panel of K frozen foundation backbones $\{\Phi_k\}_{k=1}^K$: VGG-16 [21], DINOv2 [16], CLIP [20], and SAM image encoder [8]. All foundation features are bilinearly resampled to a common spatial grid Ω and rescaled per-channel to unit variance: $\tilde{\Phi}_k = \Phi_k / \hat{\sigma}_k$, with $\hat{\sigma}_k$ the channel-wise standard deviation of Φ_k over the training set.

A uniform sum across $\{\Phi_k\}$ is dominated by whichever backbone has the largest local response, and ignores that different backbones are informative in different regions: SAM near object boundaries, CLIP in semantically rich regions, VGG in textured areas, DINOv2 in mid-level structural regions. We therefore weight each backbone's contribution at each pixel by its local informativeness, measured as its feature variance across teacher samples. To prevent backbones with globally larger feature variance from dominating, we normalize each backbone's variance by its training-set mean before computing routing weights:

$$\begin{aligned} \tilde{v}_k(x,p) &= \frac{v_k(x,p)}{\mathbb{E}_{x'}\big[\overline{v_k}(x')\big] + \epsilon}, \\ W_k(x,p) &= \frac{\exp\big(\tilde{v}_k(x,p)/\tau\big)}{\sum_{k'} \exp\big(\tilde{v}_{k'}(x,p)/\tau\big)}, \end{aligned} \tag{8}$$

where $\overline{v_k}(x') = \frac{1}{|\Omega|}\sum_p v_k(x',p)$ is the spatial mean of v_k on image x', and τ is a temperature controlling routing sharpness. Setting $\tau \to 0$ assigns each pixel to its single most informative backbone; $\tau \to \infty$ recovers uniform weighting. We use $\tau = 1.0$ throughout. The alignment target in foundation feature space is computed as the mean of features extracted from individual teacher samples, which preserves high-frequency feature responses that are smoothed out by averaging in pixel space:

$$\bar{F}_k(x)_p = \frac{1}{2N}\sum_{n=1}^{2N} \tilde{\Phi}_k\big(y_T^{(n)}(x)\big)_p. \tag{9}$$

The multi-foundation alignment loss is then

$$\mathcal{L}_{\text{MFM}} = \frac{1}{|\Omega|}\sum_{k=1}^{K}\sum_{p\in\Omega} W_k(x,p)\cdot\big\|\tilde{\Phi}_k(S_\theta(x))_p - \bar{F}_k(x)_p\big\|_2^2. \tag{10}$$

Structural consistency. Pixel and feature losses against the ensemble mean $\bar{y}_T$ are vulnerable to over-smoothing, since $\bar{y}_T$ averages out high-frequency details that vary across teacher samples. We add a structural term based on SSIM [29], which is computed over local windows and is invariant to such averaging in its low-order statistics:

$$\mathcal{L}_{\text{ssim}} = 1 - \text{SSIM}\big(S_\theta(x),\, \bar{y}_T(x)\big). \tag{11}$$

Total objective. The full training loss is a weighted sum of the three terms:

$$\mathcal{L} = \lambda_{\text{pix}}\mathcal{L}_{\text{pix}} + \lambda_{\text{MFM}}\mathcal{L}_{\text{MFM}} + \lambda_{\text{ssim}}\mathcal{L}_{\text{ssim}}. \tag{12}$$

We use $\lambda_{\text{pix}} = 1.0$, $\lambda_{\text{MFM}} = 0.5$, $\lambda_{\text{ssim}} = 0.2$ throughout. Only S_θ is updated during training; teachers, foundation backbones, and all cached statistics are frozen. At inference time, only S_θ runs; teachers, foundation backbones, and the cached statistics are not invoked.

3.4. Student Architecture

The student S_θ must satisfy the real-time constraint specified in Sec. 3.1, while retaining receptive field large enough to capture cross-modal context between I_{IR} and I_{VIS}, a property standard mobile CNNs lack at this parameter scale. We adopt a ConvNeXt V2 [31] dual-encoder U-Net as the default student: two depthwise-separable encoders process I_{IR} and I_{VIS} independently with 7×7 depthwise convolutions and Global Response Normalization (GRN), and a U-Net decoder produces $\hat{y}$ at full resolution. Encoder features at each scale are merged through a residual fusion head:

$$F_{\text{out}} = \text{Attn}(F_{\text{cat}}) \odot F_{\text{cat}} + F_{\text{IR}} + F_{\text{VIS}}, \tag{13}$$

where $F_{\text{cat}} = [F_{\text{IR}}; F_{\text{VIS}}]$, $\text{Attn}(\cdot)$ is a sigmoid channel attention, and the additive paths preserve direct modality contributions without the zero-sum constraint of softmax fusion. The framework is not tied to this backbone; we verify in Sec. 4.5 that it remains effective across mobile-CNN, mobile-Transformer, and ultra-lightweight alternatives.

3.5. Robustness to Sensor Misalignment

Real-world IR and visible sensors are rarely perfectly co-registered: small inter-sensor offsets, baseline parallax, and lens distortion mismatch produce sub-pixel to several-pixel misalignment in deployment. A fusion module trained on perfectly aligned pairs will degrade on such inputs. We address this with a training-time intervention that adds zero inference cost.

Misalignment injection. During training, we apply a random affine perturbation $\mathcal{T}$ to I_{IR} before passing it to S_θ, while supervising against the teacher ensemble $\mathcal{Y}_T$ computed on the unperturbed pair:

$$S_\theta\big(\mathcal{T}(I_{\text{IR}}),\, I_{\text{VIS}}\big) \longleftarrow \bar{y}_T\big(I_{\text{IR}},\, I_{\text{VIS}}\big), \tag{14}$$

where $\mathcal{T}$ samples translation $\Delta t \in [-10, 10]$ px and rotation $\theta \in [-2°, 2°]$. The teacher ensemble is always computed on aligned inputs and remains cacheable; only the student's input is perturbed, forcing S_θ to implicitly compensate for misalignment when reconstructing the aligned fusion target. At inference, no perturbation is applied and no parameters or latency are added to S_θ.

4. Experiments

4.1. Setup

Training. We train on MSRS [24] with teacher samples generated offline by Mask-DiFuser [25] and DDFM [43]. We use $N = 4$ DDIM samples per teacher per training image (8 total). The student is trained from scratch for 160 epochs on a single H100 with batch size 8 at resolution 256×256. Loss weights are $\lambda_{\text{pix}} = 1.0$, $\lambda_{\text{MFM}} = 0.5$, $\lambda_{\text{ssim}} = 0.2$, selected on a held-out validation split. We benchmark across two deployment tiers: *(i) Server* (A100/H100, used only for training and teacher inference; not in latency tables); *(ii) Commodity hardware* (RTX 4070 desktop GPU, RTX 3060 desktop GPU, Apple M3 laptop).

Foundation panel. $\Phi_k \in \{$VGG-16 [21], DINOv2-ViT-B/14 [16], CLIP-ViT-L/14 [20], SAM-ViT-B image encoder [8]$\}$. All four are frozen and used only at training time. We extract features at one mid-block per backbone, upsample to a common 64×64 grid, and apply training-set normalization $\hat{\sigma}_k$.

4.2. Fusion Quality

We compare against representative IVIF methods spanning diffusion (DDFM [43], Mask-DiFuser [25], ControlFusion [26], Text-IF [35]), one-step diffusion (RFfusion [30]), AE-based (CDDFuse [42], FILM [45]), GAN-based (TarDAL [11]), and unified methods (U2Fusion [33], SegMIF [12]). Following the requirements in Sec. 3.1, we prioritize learning-based IQA metrics (MUSIQ[7], CLIP-IQA[28], DeQA[36]), which correlate with perceptual quality, and downstream-task metrics (mAP, mIoU), which directly reflect plug-and-play compatibility. Traditional fusion metrics (EN, MI, SF, Q_{abf})[13] are reported alongside in Table 1 for completeness. Methods are partitioned by inference latency (Table 1). FusionProxy is the only method to reach ≥ 30 FPS while remaining within 0.5 mIoU of the best non-real-time baseline (65.4 vs. 65.9 of Mask-DiFuser, which is $\sim 10^3\times$ slower) and achieving the second-best learned-IQA scores in the entire table after ControlFusion (which runs at 0.9 FPS).

4.3. Plug-and-Play Deployment

Main results in Table 1 establish that FusionProxy's fused output is competitive with diffusion teachers on frozen downstream metrics. Here we isolate the plug-and-play

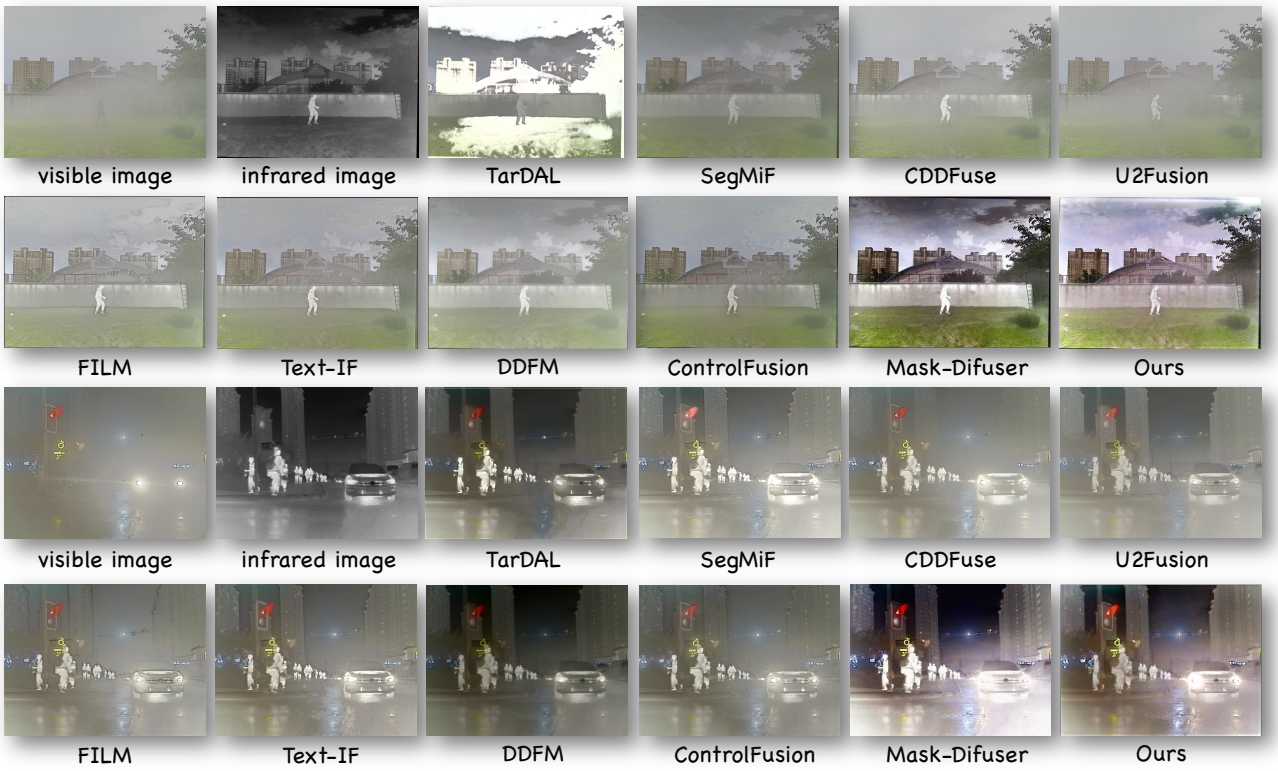

Figure 3. Qualitative comparison of image fusion results with state-of-the-art methods.

claim itself: *does inserting FusionProxy into a frozen perception stack actually improve downstream behavior?* We answer this at two levels of integration. **(i) Perception-level:** we feed the unmodified visible image I_{VIS} versus the fused output $S_\theta(x)$ to frozen YOLOv8 [6] (detection on M3FD [11]) and SegFormer-B2 [32] (segmentation on MSRS), with no fine-tuning of either model. **(ii) System-level:** we run a frozen RGB-pretrained CILRS driving policy in CARLA [3] as shown in Figure 4 under degraded visibility (dense fog and night-glare scenarios in unseen Town02), with FusionProxy inserted between the IR/VIS sensors and the policy without modifying any component of the existing autonomous driving stack. The simulated thermal modality in CARLA is a semantic-segmentation-derived approximation rather than physically calibrated infrared (real-world generalization is established by the M3FD/MSRS results in row (i), which use authentic IR-VIS sensor data.

The lift is consistent across both integration levels. At the perception level, swapping I_{VIS} for $S_\theta(x)$ improves frozen YOLOv8 by $+17.0$ mAP and frozen SegFormer-B2 by $+16.3$ mIoU. At the system level, the same swap raises closed-loop driving success from 52.4% to 86.5% on CARLA scenarios designed to stress visible-only perception. In both cases the downstream models receive no modification, no fine-tuning, and no signal that fusion has occurred—the only change is the source of the RGB-format input. This is the direct empirical signature of plug-and-play deployment: thermal awareness is added by a swap at the input layer alone.

4.4. Real-Time Inference on Commodity Hardware

A core requirement of FusionProxy is real-time inference outside server-tier hardware: prior IVIF methods either ignore inference latency entirely or report only desktop/server numbers, leaving the deployability question unanswered. We measure end-to-end latency on three commodity platforms spanning desktop GPU and mobile chip deployments. Table 3 reports latency at two common perception resolutions.

To our knowledge, no prior diffusion-derived fusion method has reported real-time throughput on consumer-grade desktop GPU or mobile-chip hardware. The RTX 3060 result ($\sim$48 FPS at 480×640) and Apple M3 result ($\sim$38 FPS at 480×640) directly establish that FusionProxy can be deployed in commodity perception stacks without specialized accelerator hardware.

4.5. Ablation Study

We ablate three axes of the FusionProxy design: (i) the teacher ensemble (Block A), (ii) the dual-signal supervision (Block B), and (iii) the student backbone (Block C). Within

Table 1. Main comparison on MSRS, partitioned by latency tier (FP32, RTX 4070). FusionProxy uniquely combines diffusion-grade quality, plug-and-play compatibility with frozen RGB-pretrained models (YOLOv8, SegFormer-B2), and real-time inference. Red : best within tier.

Dataset: MSRS		**Learning-based IQA**			**Traditional Metrics**				**Frozen Downstream**	
Method	FPS↑	MUSIQ↑	CLIP-IQA↑	DeQA↑	EN↑	MI↑	SF↑	Q_{abf} ↑	mAP↑	mIoU↑
Tier 1 – Very High Latency (diffusion sampling, (∗): requiring orders of magnitude more inference time, ≪1 FPS)										
DDFM [43]	∗	38.16	0.32	2.07	6.38	12.21	4.87	0.13	71.5	62.5
Mask-DiFuser [25]	∗	32.72	0.46	2.56	7.75	12.39	13.20	0.17	77.2	65.9
Tier 2 – High Latency (<10 FPS)										
RFfusion [30]	0.35	42.81	0.31	2.20	6.96	11.04	8.80	0.47	68.2	56.2
Text-IF [35]	0.85	43.46	0.44	2.25	6.22	12.19	6.46	0.16	74.5	63.8
ControlFusion [26]	0.90	51.72	0.32	2.43	6.45	11.82	11.39	0.34	75.9	65.1
CDDFuse [42]	1.14	44.81	0.42	2.23	6.28	12.14	10.89	0.41	74.8	64.3
SegMIF [12]	1.18	39.09	0.27	2.46	5.79	11.68	7.90	0.23	70.1	59.5
FILM [45]	2.13	36.20	0.32	2.03	6.34	12.14	7.17	0.18	72.4	62.9
U2Fusion [33]	2.50	35.49	0.44	2.09	5.88	10.35	8.31	0.28	69.8	61.7
Tier 3 – Medium Latency (10–30 FPS, workstation-only)										
TarDAL [11]	16.11	25.44	0.21	1.62	6.16	6.93	6.40	0.12	65.8	58.3
Tier 4 – Real-time (≥30 FPS, commodity-deployable)										
Ours (FusionProxy)	**32.16**	48.22	0.37	2.38	6.57	7.64	9.33	0.32	76.5	65.4

Table 2. Plug-and-play lift over RGB-only input. All downstream models (YOLOv8, SegFormer-B2, CILRS) use unmodified pretrained weights; only the input image source changes between rows.

	Perception (M3FD/MSRS)		**Closed-loop driving (CARLA)**		
Input to frozen model	Det. mAP↑	Seg. mIoU↑	Success↑	Collision↓	Lane Infr.↓
I_{VIS} (RGB-only)	58.2	49.1	52.4	32.0	4.82
$S_\theta(x)$ **(Ours)**	**75.2**	**65.4**	**86.5**	**3.5**	**0.74**
Δ (lift from fusion)	+17.0	+16.3	+34.1	−28.5	−4.08

Table 3. Inference latency on commodity hardware (FP16 PyTorch, median over 1000 runs). FusionProxy reaches ≥30 FPS at 480 × 640 on RTX 3060 and Apple M3.

Platform	**Resolution**	**ms↓**	**FPS↑**
RTX 4070	480 × 640	11.91	83.98
	768 × 1024	19.32	51.75
RTX 3060	480 × 640	20.95	47.7
	768 × 1024	68.11	14.7
Apple M3	480 × 640	26.40	37.8
	768 × 1024	82.50	12.1

each block, the remaining components are held at the default FusionProxy configuration (dual teacher, full dual-signal supervision, ConvNeXt V2-based student); rows (c), (h), and (l) thus correspond to the same default setup viewed from different ablation axes.

Block A: teacher ensemble. Increasing the sample budget from $N{=}1$ to $N{=}4$ on a single teacher (a)→(b) lifts MUSIQ by +3.8 and mIoU by +2.3, indicating that ensemble averaging suppresses sample-specific teacher noise. Adding a complementary teacher with the same total budget (b)→(c) yields a further +1.7 MUSIQ and +1.6 mIoU, showing that teacher diversity matters even when sample count is held fixed.

Block B: dual-signal supervision. The largest single jump in the entire ablation comes from multi-foundation alignment over VGG-only perceptual loss ((d)→(h): +5.7 MUSIQ, +4.4 mIoU), confirming that plug-and-play compatibility with heterogeneous frozen models requires alignment across multiple foundation feature spaces. Within multi-foundation variants, spatially-adaptive routing outperforms uniform weighting ((e)→(h): +2.5 MUSIQ) and learned global weights ((f)→(h): +1.3 MUSIQ), confirm-

Table 4. Ablation across teacher ensemble, supervision design, and student backbone. Within each block, all other components are held at the default FusionProxy configuration; rows (c), (h), and (l) correspond to the same default setup. Red : best in block.

ID	Variant	4070 FPS↑	MUSIQ↑	DeQA↑	Det. mAP↑	Seg. mIoU↑
Block A: Teacher ensemble (default backbone, default supervision)						
(a)	Single teacher (Mask-DiFuser), $N{=}1$	84	42.7	2.10	72.0	61.5
(b)	Single teacher, $N{=}4$ ensemble mean	84	46.5	2.28	74.1	63.8
(c)	**Dual teacher (M-D + DDFM), $N{=}4$ each**	84	48.2	2.38	75.2	65.4
Block B: Dual-signal supervision (default backbone, dual teacher)						
(d)	VGG-only perceptual loss	84	42.5	2.09	71.8	61.0
(e)	Multi-FM, uniform weighting	84	45.7	2.21	73.5	63.2
(f)	Multi-FM, learned global weights	84	46.9	2.30	74.4	64.3
(g)	w/o uncertainty weighting (uniform pixel weights)	84	47.0	2.31	74.5	64.6
(h)	**Full dual-signal (Eq. 8)**	84	48.2	2.38	75.2	65.4
Block C: Student backbone (dual teacher, full dual-signal supervision)						
(i)	Ultra-light (custom CNN)	198	41.5	1.78	63.1	51.6
(j)	EfficientFormer V2-S1 (mobile-Transformer)	101	46.8	2.32	73.8	63.9
(k)	MobileNetV4-Conv-M (mobile-CNN)	122	46.1	2.28	72.5	62.7
(l)	**ConvNeXt V2-based (default)**	84	48.2	2.38	75.2	65.4

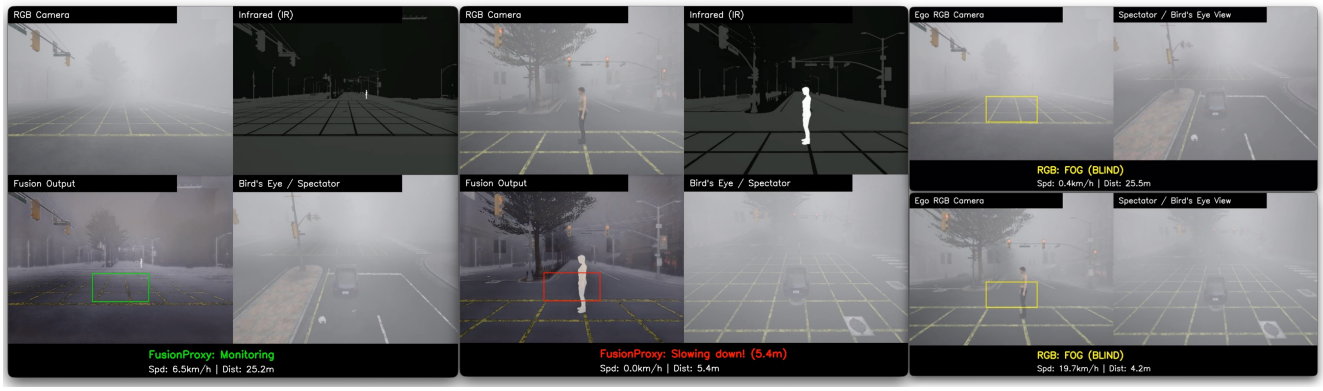

Figure 4. Closed-loop autonomous driving in CARLA: visual examples of degraded-visibility scenarios where FusionProxy (left) reveals pedestrians/vehicles missed by RGB-only (right), enabling correct steering decisions by the frozen CILRS policy.

ing that the most informative backbone genuinely varies across image regions. Uncertainty weighting (g) vs. (h) provides a smaller but consistent improvement, validating pixel-space sample variance as a supervision signal.

Block C: student backbone. We instantiate FusionProxy with three lighter alternatives spanning the mobile design space: an ultra-lightweight custom CNN, MobileNetV4-Conv-M [19] representing the mobile-CNN family, and EfficientFormer V2-S1 [9] representing the mobile-Transformer family. The two mobile-class backbones (rows (j) and (k)) reach within 2.7 mIoU of the default ConvNeXt V2-based student while running 1.2–1.5× faster on the same hardware—both still well above the $\geq$ 30 FPS real-time threshold while preserving over 95% of the default's segmentation performance.

5. Conclusion

We present **FusionProxy**, a plug-and-play, real-time fusion module that distills a diffusion teacher ensemble into a deterministic single-pass student. Experiments across image fusion benchmarks, frozen RGB-pretrained perception models, and a closed-loop autonomous driving pipeline confirm that FusionProxy delivers diffusion-grade quality in real time on commodity hardware.

Limitations. FusionProxy decouples the student architecture from the teacher's denoising network, which enables lightweight backbones to inherit diffusion-grade fusion quality at real-time speeds. The trade-off is that the student does not retain the teachers' generative diversity: a single input deterministically maps to a single output, so applications requiring multiple plausible fusions would need to revisit this design choice.

References

[1] R Archana and PS Eliahim Jeevaraj. Deep learning models for digital image processing: a review. *Artificial Intelligence Review*, 57(1):11, 2024. 1, 2

[2] Fanglin Bao, Xueji Wang, Shree Hari Sureshbabu, Gautam Sreekumar, Liping Yang, Vaneet Aggarwal, Vishnu N Boddeti, and Zubin Jacob. Heat-assisted detection and ranging. *Nature*, 619(7971):743–748, 2023. 1

[3] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 1–16, 2017. 6

[4] Yuchen Guo and Weifeng Su. Fuse4seg: Image-level fusion based multi-modality medical image segmentation. *arXiv preprint arXiv:2409.10328*, 2024. 2

[5] Yuchen Guo, Ruoxiang Xu, Rongcheng Li, and Weifeng Su. Dae-fuse: An adaptive discriminative autoencoder for multi-modality image fusion. *arXiv preprint arXiv:2409.10080*, 2024. 2, 3

[6] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolov8, 2023. 6

[7] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5148–5157, 2021. 5

[8] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 4, 5

[9] Yanyu Li, Ju Hu, Yang Wen, Georgios Evangelidis, Kamyar Salahi, Yanzhi Wang, Sergey Tulyakov, and Jian Ren. Rethinking vision transformers for mobilenet size and speed. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16889–16900, 2023. 8

[10] Xinqi Lin, Fanghua Yu, Jinfan Hu, Zhiyuan You, Wu Shi, Jimmy S Ren, Jinjin Gu, and Chao Dong. Harnessing diffusion-yielded score priors for image restoration. *ACM Transactions on Graphics (TOG)*, 44(6):1–21, 2025. 3

[11] 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*, pages 5802–5811, 2022. 2, 3, 5, 6, 7

[12] 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 *International Conference on Computer Vision*, 2023. 5, 7

[13] Jiayi Ma, Yong Ma, and Chang Li. Infrared and visible image fusion methods and applications: A survey. *Information fusion*, 45:153–178, 2019. 1, 5

[14] Jiayi Ma, Han Xu, Junjun Jiang, Xiaoguang Mei, and Xiao-Ping Zhang. Ddcgan: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion. *IEEE Transactions on Image Processing*, 29:4980–4995, 2020. 3

[15] Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14297–14306, 2023. 3

[16] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 4, 5

[17] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 3

[18] Gerald B Popko, Thomas K Gaylord, and Christopher R Valenta. Interference measurements between single-beam, mechanical scanning, time-of-flight lidars. *Optical Engineering*, 59(5):053106–053106, 2020. 1

[19] Danfeng Qin, Chas Leichner, Manolis Delakis, Marco Fornoni, Shixin Luo, Fan Yang, Weijun Wang, Colby Banbury, Chengxi Ye, Berkin Akin, et al. Mobilenetv4: Universal models for the mobile ecosystem. In *European conference on computer vision*, pages 78–96. Springer, 2024. 8

[20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 4, 5

[21] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4, 5

[22] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3

[23] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. 2023. 3

[24] Linfeng Tang, Jiteng Yuan, Hao Zhang, Xingyu Jiang, and Jiayi Ma. Piafusion: A progressive infrared and visible image fusion network based on illumination aware. *Information Fusion*, 83-84:79–92, 2022. 5

[25] Linfeng Tang, Chunyu Li, and Jiayi Ma. Mask-difuser: A masked diffusion model for unified unsupervised image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 1, 3, 5, 7

[26] Linfeng Tang, Yeda Wang, Zhanchuan Cai, Junjun Jiang, and Jiayi Ma. Controlfusion: A controllable image fusion framework with language-vision degradation prompts. *arXiv preprint arXiv:2503.23356*, 2025. 1, 3, 5, 7

[27] Simon Verghese. Self-driving cars and lidar. In *CLEO: Applications and Technology*, pages AM3A–1. Optica Publishing Group, 2017. 1

[28] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2555–2563, 2023. 5

[29] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 5

[30] Zirui Wang, Jiayi Zhang, Tianwei Guan, Yuhan Zhou, Xingyuan Li, Minjing Dong, and Jinyuan Liu. Efficient rectified flow for image fusion. *arXiv preprint arXiv:2509.16549*, 2025. 1, 3, 5, 7

[31] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16133–16142, 2023. 5

[32] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34: 12077–12090, 2021. 6

[33] 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, 2020. 5, 7

[34] Xunpeng Yi, Linfeng Tang, Hao Zhang, Han Xu, and Jiayi Ma. Diff-if: Multi-modality image fusion via diffusion model with fusion knowledge prior. *Information Fusion*, 110:102450, 2024. 3

[35] 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*, pages 27026–27035, 2024. 1, 5, 7

[36] Zhiyuan You, Xin Cai, Jinjin Gu, Tianfan Xue, and Chao Dong. Teaching large language models to regress accurate image quality scores using score distribution. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14483–14494, 2025. 5

[37] Fanghua Yu, Jinjin Gu, Zheyuan Li, Jinfan Hu, Xiangtao Kong, Xintao Wang, Jingwen He, Yu Qiao, and Chao Dong. Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 25669–25680, 2024. 3

[38] Hao Zhang, Han Xu, Xin Tian, Junjun Jiang, and Jiayi Ma. Image fusion meets deep learning: A survey and perspective. *Information Fusion*, 76:323–336, 2021. 1, 2

[39] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 3

[40] Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. *arXiv preprint arXiv:2406.19680*, 2024. 3

[41] Zixiang Zhao, Shuang Xu, Chunxia Zhang, Junmin Liu, Pengfei Li, and Jiangshe Zhang. Didfuse: Deep image decomposition for infrared and visible image fusion. *arXiv preprint arXiv:2003.09210*, 2020. 2

[42] 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 (CVPR)*, pages 5906–5916, 2023. 2, 5, 7

[43] 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*, pages 8082–8093, 2023. 1, 3, 5, 7

[44] Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Kai Zhang, Shuang Xu, Dongdong Chen, Radu Timofte, and Luc Van Gool. Equivariant multi-modality image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2

[45] Zixiang Zhao, Lilun Deng, Haowen Bai, Yukun Cui, Zhipeng Zhang, Yulun Zhang, Haotong Qin, Dongdong Chen, Jiangshe Zhang, Peng Wang, et al. Image fusion via vision-language model. *arXiv preprint arXiv:2402.02235*, 2024. 5, 7

[46] Zixiang Zhao, Haowen Bai, Bingxin Ke, Yukun Cui, Lilun Deng, Yulun Zhang, Kai Zhang, and Konrad Schindler. A unified solution to video fusion: From multi-frame learning to benchmarking. *arXiv preprint arXiv:2505.19858*, 2025. 3