PhysDNet: Physics-Driven Gradient Amplification for Real-Time Image Dehazing

Rayen Dhahri[1] **Lei Sun**[2,†] **Victor Javier Kartsch Morinigo**[3]
Alice Pagnoux[3] **Luca Benini**[3,4] **Danda Pani Paudel**[2] **Luc Van Gool**[2]

[1]Independent Researcher
[2]INSAIT, Sofia University "St. Kliment Ohridski"
[3]ETH Zurich
[4]Department of Electrical, Electronic and Information Engineering (DEI), University of Bologna

Abstract

Modern dehazing networks are often computationally expensive and overfit to synthetic haze, limiting real-time deployment. This underscores an urgent demand for efficient, high-fidelity restoration methods that can operate on resource-constrained edge devices in practical, dynamic environments. We introduce PhysDNet, a lightweight multimodal architecture that fuses NIR imagery with sparse LiDAR depth and is trained with a physics-constrained dual-head objective. Differentiating through a fixed Koschmieder inversion induces *density-adaptive gradient amplification*, acting as an implicit resource allocation mechanism that concentrates learning on heavily degraded regions without increasing inference cost. PhysDNet delivers superior restoration with minimal capacity and enhanced training efficiency, facilitating real-time deployment on edge devices without increasing inference overhead. On the Seeing Through Fog benchmark, PhysDNet outperforms competitors by **1.67 dB** at **5.2**$\times$ lower compute and recovering **54.9%** of fog-induced detection loss. Ablations and cross-domain evaluations show that physics-guided training improves generalization including under non-homogeneous haze that violates the Koschmieder assumption. PhysDNet further demonstrates compatibility with fixed-function NPUs, running at **29/45/99 FPS** at $<$**4 W** across L/M/S variants. Code, deployment scripts and pre-trained checkpoints are available at `https://github.com/Heatdh/PhysDnet`.

1 Introduction

Image dehazing is foundational for outdoor perception under adverse weather, particularly autonomous driving where reduced visibility compromises detection and tracking [1]. Modern autonomous platforms already carry the sensors needed: time-gated NIR cameras that reject nearby backscatter, and LiDAR providing weather-invariant depth [2]. Active time-gated NIR provides a hardware-level bypass to RGB's failure under dense scattering: synchronized exposure with emitted laser pulses rejects atmospheric backscatter and recovers structural semantics where RGB is effectively blind. Yet most learned dehazing methods [3–7] process RGB alone and chase quality by scaling capacity, deeper backbones, wider attention, iterative refinement, inflating inference cost and relying on operators such as self-attention and layer normalization, which are incompatible with fixed-function edge accelerators [8]. Consequently, a practical dehazing solution must be lightweight enough for deployment on low-power edge accelerators while seamlessly exploiting the multi-modal sensor suites already ubiquitous on autonomous platforms.

† Correspondence to: Lei Sun (`lei.sun@insait.ai`, `rayendhahri@gmail.com`).

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

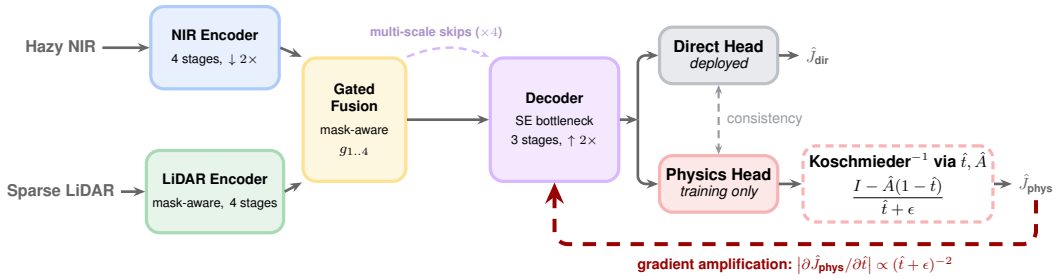

Figure 1: **PhysDNet architecture.** Dual encoders extract NIR and sparse-depth features, merged by mask-aware gating with Squeeze-and-Excitation at the bottleneck. The shared representation feeds a direct head (deployed at inference) and a physics head (training only). Backpropagating through the Koschmieder inversion scales gradients by $(\hat{t}+\epsilon)^{-2}$, concentrating optimization on heavily degraded regions at zero inference cost.

We introduce **PhysDNet**, a lightweight multimodal dehazing network for real-time edge deployment. PhysDNet fuses gated NIR imagery with sparse LiDAR depth through mask-aware multi-scale gating and is trained with a *physics-constrained dual-head framework*: a deployable direct restoration head and a parameter-free auxiliary head that inverts the Koschmieder scattering model [9].

Our central insight is that the auxiliary physics head improves learning not by constraining the final output, but by shaping the *optimization dynamics*. Backpropagating through the fixed physical inversion induces transmission-dependent gradient scaling proportional to $(t+\epsilon)^{-2}$, automatically amplifying supervision in heavily degraded regions. We call this *density-adaptive gradient amplification*: a hyperparameter-free implicit curriculum that prioritizes the hardest pixels during training at zero inference cost. **To our knowledge, this is the first analysis of differentiation through a physical inversion as a mechanism for shaping optimization via structured gradient scaling.** This mechanism improves performance through more efficient use of gradients during training rather than increased model capacity. In contrast to scaling architectures, which increases inference cost, PhysDNet achieves its gains by reallocating optimization focus, leaving the deployment footprint unchanged.

We deliver the following contributions:

- **Density-adaptive gradient amplification.** A dual-head training framework where a fixed inversion induces $(t+\epsilon)^{-2}$ gradient scaling, focusing learning on degraded regions without explicit weighting or extra inference cost.
- **State-of-the-art accuracy at low compute.** On STF, PhysDNet reaches 36.46 dB PSNR at 18.98 GMACs, outperforming FFA-Net by $+1.67$ dB at $5.2\times$ lower compute, and recovers 54.9% of YOLOv8 detection loss under aggressive fog.
- **Edge deployability.** The only evaluated method combining competitive restoration quality with fixed-function NPU compatibility, running at 29-99 FPS at <4 W under standard INT8 quantization.

2 Method

We adopt the Koschmieder scattering model [9]: for a pixel x, the hazy intensity $I(x)$ is a convex combination of scene radiance $J(x)$ and global airlight A,

$$I(x) = J(x)\,t(x) + A\big(1 - t(x)\big), \qquad t(x) = e^{-\beta d(x)}, \tag{1}$$

with transmission $t(x) \in (0, 1]$, scattering coefficient β, and depth $d(x)$. Single-image recovery is ill-posed in both t and A [10]; PhysDNet breaks this ambiguity by fusing sparse LiDAR depth (weather-invariant) with time-gated NIR imagery (rejects nearby backscatter) [2]. Extended background in Appendix B.

We present **PhysDNet** (**Phys**ics **D**ual-head Dehazing **Net**work), a lightweight architecture for LiDAR-guided image dehazing. A fixed, non-learnable physics inversion head operates in parallel with a

data-driven restoration head, both sharing a common encoder-decoder backbone. Backpropagation through the physics inversion induces $(\hat{t}+\epsilon)^{-2}$ gradient scaling, concentrating optimization on severely degraded regions without additional parameters.

The model receives as input a hazy image $I \in [0,1]^{3\times H\times W}$, sparse LiDAR depth $D_s \in \mathbb{R}^{1\times H\times W}$, and a binary validity mask $M \in \{0,1\}^{1\times H\times W}$, and jointly predicts a restored image, a per-pixel transmission map, and a global airlight vector.

2.1 Network Architecture

PhysDNet is a dual-branch encoder-decoder (fig. 1) with gated cross-modal fusion and two output heads. Two modality-specific encoder branches process NIR and depth independently at four spatial scales, merging features at each scale via a learned gate. A single decoder upsamples the fused representations. Unlike a standard U-Net [11], each branch learns modality-specific representations before fusion, with gating adapting to local signal availability.

Dual encoder and decoder. Two modality-specific encoder branches follow a four-stage downsampling structure using depthwise-separable (DWSep) convolutions [12], BatchNorm, and ReLU6 [13], with residual shortcuts and $2\times$ max-pooling per stage for edge efficiency. The image branch accepts 3-channel NIR input; the depth branch receives the sparse depth D_s concatenated with the validity mask M. A Squeeze-and-Excitation (SE) channel attention block [14] is applied once at the bottleneck after gated fusion. The decoder applies three bilinear upsampling stages with skip connections from each encoder scale. Full layer-wise specifications are in Appendix D.

Gated fusion. At each of the four encoder scales, the image features $\mathbf{f}_{\text{img}}$ and depth features $\mathbf{f}_{\text{depth}}$ are combined via a learned gate:

$$g = \sigma\big(W_g[\mathbf{f}_{\text{img}};\, \mathbf{f}_{\text{depth}}]\big), \qquad \mathbf{f}_{\text{fused}} = g \odot \mathbf{f}_{\text{img}} + (1-g)\odot \mathbf{f}_{\text{depth}}, \tag{2}$$

where $[\cdot\,;\cdot]$ denotes channel-wise concatenation and W_g is a 1×1 convolution following the gated multimodal fusion paradigm [15, 16]. This mechanism is inherently mask-aware: where LiDAR returns are absent ($M{=}0$), the depth features carry no geometric signal and the gate learns to rely on the image branch; where depth is present, geometric cues help disambiguate haze density from scene content.

Model family. Scaling the base channel width $c \in \{32, 64, 96\}$ yields three variants (S/M/L) spanning the accuracy, efficiency Pareto front (table 10). Efficiency metrics are presented in table 1; even PhysDNet-L (18.98 GMACs) remains lower-cost than evaluated baselines.

2.2 Physics-Constrained Dual-Head Design

The shared decoder features $\mathbf{z} \in \mathbb{R}^{c\times H\times W}$ feed two complementary output heads. This dual-head design is the core mechanism through which physical knowledge is injected into the learning process.

Direct Head (data-driven). A convolutional projection with sigmoid activation maps decoder features directly to the image domain, yielding $\hat{J}_{\text{dir}}(x) \in [0,1]^3$. This unconstrained estimator makes no assumption about the physical image formation process and serves as the deployed output, providing robustness against Koschmieder-model violations (spatially varying airlight, multiple scattering, sensor artifacts).

Physics Head (structured). Two auxiliary sub-heads predict the atmospheric scattering parameters from the same shared features:

$$\hat{t}(x) = t_{\min} + (t_{\max} - t_{\min})\,\sigma\big(W_t * \mathbf{z}(x) + b_t\big) \;\in\; [t_{\min},\, t_{\max}], \tag{3}$$

$$\hat{A} = \sigma\big(W_A \cdot \text{GAP}(\mathbf{z}_{\text{bot}}) + b_A\big) \;\in\; [0,1]^3, \tag{4}$$

where W_t is a 1×1 convolution producing per-pixel transmission, W_A is a fully connected layer applied after global average pooling of the bottleneck features $\mathbf{z}_{\text{bot}}$, and $t_{\min}{=}0.05$, $t_{\max}{=}0.95$ prevent numerical instability.

The physics-restored image is obtained by inverting Eq. (1) via its closed-form solution:

$$\hat{J}_{\text{phys}}(x) = \text{clamp}\left(\frac{I(x) - \hat{A}\,(1-\hat{t}(x))}{\hat{t}(x)+\epsilon},\; 0,\; 1\right), \tag{5}$$

with $\epsilon = 10^{-4}$. This inversion contains **zero learnable parameters**: it is a fixed, differentiable function whose output quality is determined solely by the accuracy of $\hat{t}$ and $\hat{A}$. During training, gradients flow backwards through this fixed computation into the shared encoder, producing the structured gradient signal analysed in §2.3.

Inference behaviour. At test time only the direct head is deployed; the physics head operates solely during training. This ensures reported gains reflect improved features rather than test-time physics exploitation.

2.3 Density-Adaptive Gradient Amplification

We now derive the gradient signal the physics head propagates into the shared encoder. The structure reveals an implicit curriculum that emerges from the scattering inversion itself, without explicit sample weighting or scheduling.

Taking partial derivatives of Eq. (5) (within the valid range, before clamping) with respect to the predicted scattering parameters:

$$\frac{\partial \hat{J}_{\text{phys}}}{\partial \hat{t}} = \frac{\hat{A} - I(x)}{(\hat{t} + \epsilon)^2}, \tag{6}$$

$$\frac{\partial \hat{J}_{\text{phys}}}{\partial \hat{A}} = \frac{\hat{t} - 1}{\hat{t} + \epsilon}, \tag{7}$$

where we neglect the $O(\epsilon)$ term in the numerator of Eq. (6). In practice, $\epsilon = 10^{-4}$ remains several orders of magnitude smaller than $t_{\min} = 0.05$, ensuring the approximation holds even in the lowest-transmission regime.

Transmission gradient: implicit hard-example mining. The factor $(\hat{t} + \epsilon)^{-2}$ in Eq. (6) implies that gradient magnitude scales *inversely with the square of transmission*. In dense fog ($\hat{t} \to t_{\min} = 0.05$), gradients are amplified by up to $(t_{\min} + \epsilon)^{-2} \approx 400\times$ relative to clear-air pixels ($\hat{t} \approx 1$), concentrating learning on heavily degraded regions. This scaling is empirically confirmed in fig. 3(a): PhysDNet-M follows the $(\hat{t} + \epsilon)^{-2}$ curve and exceeds the non-physics baseline by up to $\sim 694\times$ in dense haze (fig. 3(b)).

This behaviour acts as an implicit resource allocation mechanism during training, analogous to hard-example mining [17], where degraded pixels dominate the gradient signal. Unlike explicit weighting schemes (e.g., focal loss [18]), this effect requires *no additional hyperparameters* and arises directly from differentiating through the physics inversion.

Variance correspondence. The same t^{-2} structure governs the inversion's noise variance, $\mathrm{Var}[\hat{J}_{\text{phys}}] = \sigma^2/t^2 + O(\delta)$ (Appendix F, Prop. 1), making the physics head the structural inverse of classical heteroscedastic regression [19], which *down*-weights high-variance samples. Full derivations, airlight-gradient analysis, bounded-gradient discussion, and per-pixel spatial validation on real STF fog are in Appendix F. Empirically, on 300 STF validation frames with known synthetic sensor noise ($\sigma \approx 0.01$), the squared error of the inversion binned by ground-truth transmission follows σ^2/t^2 across two orders of magnitude (log-log r=0.981, 24 bins, clamp-saturated pixels excluded; App. F).

Scope of the mechanism. The inversion induces transmission-dependent Jacobian scaling, but its restoration benefit depends on whether the inverted parameter is *anchored* during training. Without transmission supervision (RGB-only RESIDE-6K [20], LOL-v1 [21]), physics and NoPhy variants are within 0.15 dB. Where a measurement anchors the parameter, the gain reappears beyond dehazing: depth-anchored underwater restoration ($+0.38$ dB overall, $+0.82$ dB mid-transmission) and Retinex [22] low-light enhancement on LOL-v2-Real [23] ($+0.5$–0.6 dB), the $A \to 0$ case of Eq. (5) (App. R). The anchored STF setting yields the largest gain (§3.3).

2.4 Physics-Anchored Residual Learning

The physics-reconstruction loss $\mathcal{L}_{\text{pr}}$ anchors the shared latent representation in a physically meaningful space, while the standard reconstruction loss $\mathcal{L}_{\text{rec}}$ lets the direct head learn residual corrections for non-homogeneous fog and sensor effects beyond Koschmieder, ensuring the network is physics-anchored but not physics-limited. A contrastive term $\mathcal{L}_{\text{contra}}$ pushes Direct Head features away from

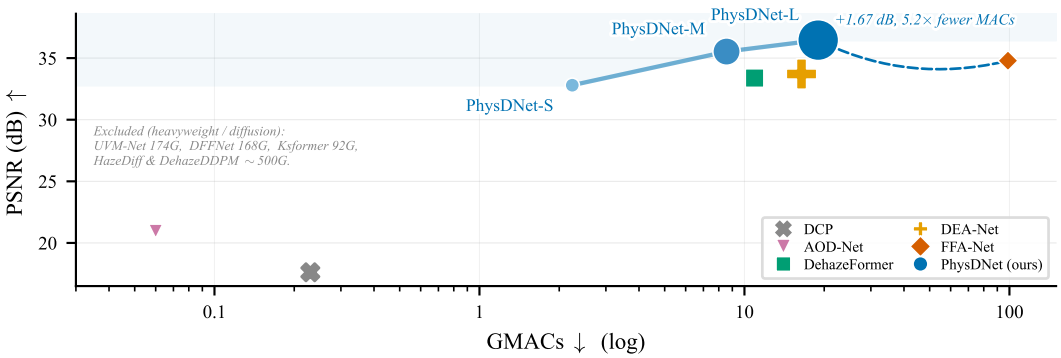

Figure 2: **Accuracy vs. computational cost on STF.** PhysDNet establishes a new Pareto front in the edge-deployable region. Baselines are plotted at their best-performing resolution (Table 1). PhysDNet-L outperforms FFA-Net by $+1.67$ dB at $5.2\times$ lower compute; PhysDNet-S retains competitive accuracy at ultra-low compute. Aside from PhysDNet, AOD-Net compiles to the NPU but its restoration quality is insufficient for deployment. Heavyweight and diffusion baselines excluded (App. C).

the hazy input, activated late in training. The full objective is organized in three families (data-driven, physics, auxiliary) with every equation in App. F.1.

3 Experiments

We evaluate PhysDNet along three axes. First, **quality** on STF (§3.2–§3.3), out-of-distribution fog (§3.4), and real STF fog (§3.8). Second, **utility** for downstream detection (§3.5) and zero-shot cross-domain transfer (§3.6). Third, **deployability** on two automotive edge platforms (§3.9). Appendices report LiDAR-sparsity robustness (App. O) and no-reference perceptual quality (App. P).

3.1 Experimental Setup

We train and evaluate on the Seeing Through Fog (STF) dataset [1], using 4 212 training frames and 1 299 validation frames with on-the-fly synthetic haze generated via Eq. (1). Sparse LiDAR is used unmodified at both train and test time. Dataset details in App. S.

We compare against Dark Channel Prior (DCP) [24], AOD-Net [3], FFA-Net [4], DehazeFormer-T [5], and DEA-Net [6]. All PhysDNet variants are trained with AdamW at 256×256 using only NPU/TensorRT-friendly operators (depthwise-separable convolutions, BatchNorm, ReLU6, bilinear upsampling). All reported results correspond to the unconstrained direct head, ensuring gains cannot stem from trivial consistency with the synthetic generation process. Recent heavyweight (UVM-Net [25], DFFNet [26], Ksformer [27]) and diffusion-based (HazeDiff [28], DehazeDDPM [29], FrDiff [7]) methods are excluded as they exceed real-time edge constraints; selection rationale in App. C.

3.2 Restoration Quality on STF

Table 1 compares PhysDNet against four learned baselines and the classical DCP [24]. At matched 256×256 resolution, PhysDNet-L reaches 36.46 dB PSNR, outperforming FFA-Net+LiDAR at its preferred 512×512 resolution (34.79 dB) by $+1.67$ dB at $5.2\times$ lower compute. PhysDNet-M (2.95 M, 8.56 GMACs) surpasses DEA-Net by $+1.82$ dB at roughly half the cost. The gap persists across resolutions, and §3.3 confirms the physics-constrained training regime, not backbone capacity, is the primary driver. Qualitative comparisons in Appendix T.

3.3 Physics Head Ablation

To isolate the effect of physics-constrained training, we remove the physics head (*NoPhy*) from an identical ch64 model, disabling transmission/airlight prediction and all physics losses. This reduces parameters by only 0.04 M (2.95→2.91 M) but causes a large drop in performance (Table 2).

Table 1: **Quantitative comparison on the STF validation set.** PSNR/SSIM for the *direct* restoration head. PhysDNet-M reports mean±SE over 3 seeds. Best in **bold**, second underlined. GMACs at the indicated resolution.

Method	Input	Res.	Params (M)	GMACs	PSNR (dB) ↑	SSIM ↑
DCP [24]	NIR	256	0	—	17.62	0.5992
AOD-Net [3]	NIR	512	0.002	0.23	18.62	0.7123
AOD-Net	NIR	256	0.002	0.06	20.69	0.7499
AOD-Net	LiDAR+Gated	512	0.002	0.24	17.05	0.6669
AOD-Net	LiDAR+Gated	256	0.002	0.06	20.95	0.7329
DehazeFormer-T [5]	NIR	512	1.36	10.92	33.38	0.9348
DehazeFormer-T	NIR	256	1.36	2.73	32.22	0.9398
DehazeFormer-T	NIR+LiDAR	512	1.36	10.97	29.91	0.9092
FFA-Net [4]	NIR	512	0.77	98.57	33.90	0.9474
FFA-Net	NIR	256	0.77	24.64	33.80	0.9504
FFA-Net	NIR+LiDAR	512	0.77	98.73	34.79	0.9516
FFA-Net	NIR+LiDAR	256	0.77	24.68	33.23	0.9476
DEA-Net [6]	NIR	256	3.65	16.42	33.71	0.9588
DEA-Net	LiDAR	256	3.65	16.44	33.53	0.9572
PhysDNet-S	NIR+LiDAR	256	0.72	2.24	32.80	0.9331
PhysDNet-M	NIR+LiDAR	256	2.95	8.56	35.55±0.07	0.9566±0.0009
PhysDNet-L	NIR+LiDAR	256	6.59	18.98	**36.46**	**0.9606**

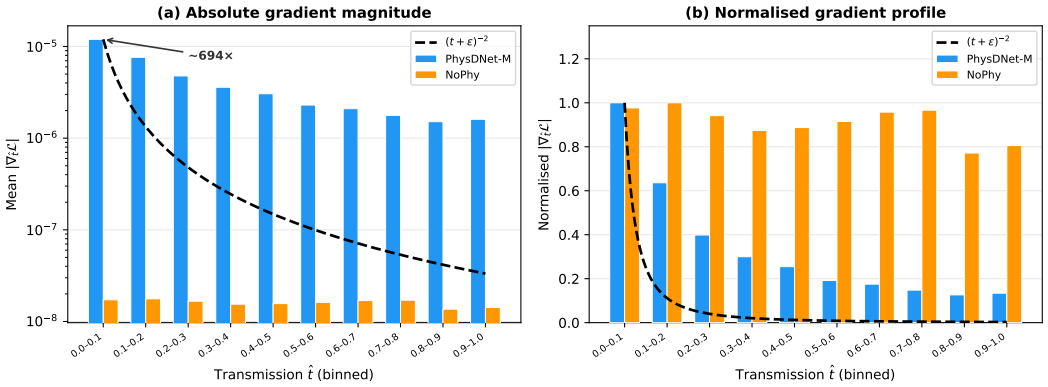

Figure 3: **Density-adaptive gradient amplification.** PhysDNet-M follows the $(t+\epsilon)^{-2}$ scaling and concentrates gradients on low-transmission regions, unlike the flat NoPhy baseline.

Table 2: **Physics head ablation.** Removing the physics head degrades PSNR by 5.51 dB and reduces downstream mAP recovery.

Variant	Physics Head	Params (M)	PSNR (dB)	SSIM	mAP (%)	Δ mAP
PhysDNet-M	✓	2.95	**35.55**	**0.9569**	**11.70**	**+3.05**
PhysDNet-M NoPhy	×	2.91	30.02	0.9263	10.70	+2.06
Difference		−0.04	−5.51	−0.0306	−1.00	−0.99

Removing the physics head reduces PSNR by **5.51 dB** and mAP recovery from $+3.05$ to $+2.06$, indicating that gains arise from the training regime rather than architecture alone.

Gradient amplification. Figure 3 confirms the $(\hat{t}+\epsilon)^{-2}$ scaling: PhysDNet-M gradients follow the theoretical curve and exceed NoPhy by up to $\sim 694\times$ in dense haze, reflecting implicit hard-example mining.

3.4 Out-of-Distribution Fog Robustness

We sweep $\beta \in [0.005, 0.10]$ at A=0.8 on the full validation set (1 299 frames per β), with no retraining (see Appendix U for the full performance curve); values $\beta \geq 0.05$ lie outside the training

Table 3: **Downstream object detection** on 426 STF val frames with synthetic fog (β=0.06). Detector: YOLOv8-M (conf$\geq$0.25, IoU$\geq$0.5). ΔmAP is relative to the foggy baseline; recovery = ΔmAP/|fog degradation|.

Condition	AP_{veh}	AP_{pers}	AP_{cyc}	mAP (%)	ΔmAP
Clean (upper bound)	33.95	7.97	1.05	14.33	—
Foggy (β=0.06)	18.09	5.42	2.42	8.64	−5.68
AOD-Net (NIR)	15.14	4.66	1.27	7.02	−1.62
AOD-Net (LiDAR+Gated)	14.40	4.23	1.49	6.70	−1.94
DehazeFormer-T (NIR)	16.37	5.74	1.65	7.92	−0.72
DehazeFormer-T (LiDAR)	15.62	5.24	1.21	7.36	−1.29
FFA-Net (NIR)	24.37	6.30	1.32	10.66	+2.02
FFA-Net (LiDAR)	23.09	5.35	1.39	9.94	+1.30
DEA-Net (NIR)	25.82	5.82	1.66	11.10	+2.45
DEA-Net (LiDAR)	25.52	5.72	2.00	11.08	+2.44
PhysDNet-S	24.49	5.94	1.22	10.55	+1.91
PhysDNet-M	**27.65**	6.45	0.99	**11.70**	**+3.05**
PhysDNet-L	27.60	**6.70**	0.99	**11.76**	**+3.12**

Recovery: PhysDNet-L=**54.9%**, PhysDNet-M=53.7%, DEA-Net (NIR)=43.1%, FFA-Net (NIR)=35.6%

range. PhysDNet-M maintains the highest PSNR across the entire range and degrades smoothly beyond the training boundary, retaining a >6 dB advantage over FFA-Net at β=0.06. We attribute this robustness to the physics-constrained encoder: because the inversion formula (Eq. 5) is domain-invariant, the learned (t, A, J) factorisation generalises to unseen β values without retraining.

3.5 Downstream Object Detection

Pixel-level metrics on synthetic data do not guarantee that semantically meaningful structure has been recovered. We therefore measure the impact of dehazing on downstream object detection using YOLOv8-M [30] (pretrained on COCO, conf$\geq$0.25, IoU$\geq$0.5) on STF validation frames with synthetically added *aggressive* fog (β=0.06, out-of-distribution). Full protocol details are in section K. We report relative recovery rather than absolute performance, as the detector is not fine-tuned for STF.

PhysDNet-L recovers 3.12 pp (**54.9%** of the degradation) and PhysDNet-M 3.05 pp (53.7%), while the best modern baseline, DEA-Net (NIR), recovers only 2.45 pp (43.1%). This indicates that PhysDNet restores semantically relevant structure rather than only improving pixel-level metrics. A representative qualitative example is shown in fig. 4.

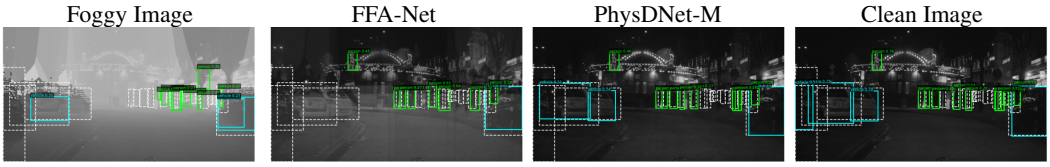

Figure 4: **Downstream detection after dehazing** (β=0.06). PhysDNet-M recovers scene contrast and enables more accurate detections compared to FFA-Net. Dashed white: ground truth; solid coloured: predictions.

3.6 Zero-Shot Cross-Domain Generalization and Non-Homogeneous Fog

We evaluate zero-shot on **O-HAZE** [31] and **I-HAZE** [32], imposing a quadruple shift (NIR→RGB, synthetic→real, LiDAR→none, outdoor→indoor). No fine-tuning. Due to near-achromatic outputs of the monochrome-trained model, we report luminance (Y) metrics following prior work; colour is learned via adaptation in §3.7.

PhysDNet remains competitive without depth input. PhysDNet-M achieves the best overall performance, improving over DEA-Net by +0.53 dB on O-HAZE and +0.62 dB on I-HAZE, while PhysDNet-S matches or exceeds all baselines on I-HAZE. The same advantage holds on NH-HAZE (non-homogeneous outdoor haze, App. N): PhysDNet-M outperforms DEA-Net by +0.15 dB PSNR-

Table 4: **Zero-shot cross-domain generalization** on O-HAZE and I-HAZE. All models trained on STF; no LiDAR at test time.

	O-HAZE [31] (outdoor)				**I-HAZE** [32] (indoor)			
Method	PSNR-Y ↑	SSIM-Y ↑	PSNR-RGB ↑	SSIM-RGB ↑	PSNR-Y ↑	SSIM-Y ↑	PSNR-RGB ↑	SSIM-RGB ↑
AOD-Net	15.11	0.5000	14.88	0.4931	14.69	0.5824	14.51	0.5724
FFA-Net	19.20	0.7177	18.49	0.6968	15.82	0.6521	15.48	0.6402
DehazeFormer-T	19.48	0.6839	18.81	0.6641	15.86	0.6142	15.56	0.6050
DEA-Net	19.92	0.7042	19.20	0.6854	16.31	0.6455	16.02	0.6362
PhysDNet-S	19.89	0.7190	19.21	0.6989	16.81	0.6895	16.49	0.6734
PhysDNet-M	**20.45**	0.7461	**19.65**	0.7253	**16.93**	**0.6947**	**16.59**	**0.6795**
PhysDNet-M (NoPhy)	18.92	0.7358	18.38	0.7174	16.33	0.6907	16.06	0.6746
PhysDNet-L	20.31	**0.7488**	19.51	**0.7279**	16.57	0.6887	16.24	0.6756



Figure 5: **Label-free colour adaptation on real haze.** Taillights and lane markings regain colour without paired data.

Y; removing the physics head reduces PSNR-Y by 0.54 dB, confirming that physics-guided training transfers despite the test distribution violating the homogeneous-scattering assumption.

The NoPhy ablation isolates the physics head: removing it reduces PSNR-Y by 1.53 dB on O-HAZE and 0.60 dB on I-HAZE. While smaller than the in-domain gap, this confirms that physics-guided training improves cross-domain generalization beyond backbone capacity.

Additional qualitative results and chrominance recovery are provided in App. L. Robustness to LiDAR sparsity is shown in App. O, where PhysDNet-M retains a $+4.10$ dB advantage at 0% depth.

3.7 RGB-Only Operation and Self-Supervised Colour Adaptation

RGB-only. Trained from scratch on RESIDE-6K [20] ($D_s{=}M{=}0$, no transmission supervision), PhysDNet-M reaches 28.4 dB on SOTS-mix (NoPhy: 28.6 dB, consistent with §2.3). For RGB-only deployment, the depth branch folds into constants (bit-exact), reducing compute from 8.56 to 6.33 GMACs (-26%).

Learning colour without paired data. A model trained on monochrome NIR produces near-achromatic RGB outputs. Rather than transferring chrominance from the hazy input, which carries the haze's colour cast, we learn colour through label-free adaptation on 4 803 unpaired URHI images (5 epochs, no architecture change). The Physics Head is closed-form algebra on the coloured input, so its detached output serves as a colour teacher for the Direct Head (App. M). On 1 297 real RTTS images [20], colourfulness [33] rises from 0.03 to 12.8 with the best no-reference scores of all conditions (NIQE 4.74, BRISQUE 33.4 vs. 5.35/38.3 for the hazy input; Fig. 5). Adaptation specializes the model to the target domain (-5.1 dB on O-HAZE), so Table 4 reports the unadapted model.

3.8 Real-World Fog Validation

Since STF real-fog frames lack ground truth, we evaluate via **dual-head consistency**, i.e., agreement between $\hat{J}_{\text{dir}}$ and $\hat{J}_{\text{phys}}$ (Table 5).

Both heads show near-perfect agreement (PSNR >30 dB, $r > 0.99$), confirming that the physics inversion produces outputs consistent with the direct head. Consistency is higher in dense fog (37.48 vs. 33.83 dB), aligning with the density-adaptive gradient amplification (§2.3).

We further validate real-haze performance using no-reference metrics (App. P): PhysDNet-M achieves the best NIQE on O-HAZE (3.92 vs. 4.64 for DEA-Net), confirming transfer beyond synthetic data.

Table 5: **Dual-head consistency on real STF fog.** High PSNR and correlation indicate strong agreement.

Metric	Dense Fog	Light Fog	Interpretation
MSE	0.0002	0.0006	Very low gap
PSNR (dB) ↑	37.48	33.83	High agreement
Pearson r ↑	0.9947	0.9912	Near-perfect match

Table 6: **Unified edge-deployment benchmark** (256×256, batch 1, 50 warmup + 300 timed iterations on the Jetson Orin Nano). Fixed-function NPU uses INT8 post-training quantization; Jetson uses TensorRT FP16. ✗ indicates compilation failure due to unsupported operators. Full CUDA FP32 numbers and power profiles in Appendix G.

Model	Params (M)	GMACs	Fixed-function NPU		Jetson Orin Nano	
			Compile	FPS	FPS	$\overline{P}$ (W)
AOD-Net LiDAR	0.002	0.06	✓	**1220**	327.0	4.94
FFA-Net LiDAR	0.77	24.68	✗	—	46.9	8.21
DehazeFormer-T LiDAR	1.36	2.73	✗	—	47.1	12.38
DEA-Net LiDAR	3.65	16.44	✗	—	28.5	10.17
PhysDNet-S (ours)	0.72	2.24	✓	**99**	**108.3**	13.45
PhysDNet-M (ours)	2.95	8.56	✓	45	58.5	12.38
PhysDNet-L (ours)	6.59	18.98	✓	29	38.8	11.80

Across all official STF weather splits, including rain and snow, agreement remains high and is again highest in dense fog ($r\geq0.993$ throughout; App. Q).

3.9 Edge Deployment

PhysDNet's restriction to NPU-friendly operators avoids softmax-attention, LayerNorm, and deformable convolutions, which block baseline compilation.

We validate deployability on two representative edge targets: (i) a fixed-function 26 TOPS NPU (<4 W TDP), representative of strict automotive perception SoCs, and (ii) the NVIDIA Jetson Orin Nano Super (8 GB), representative of more permissive ADAS compute platforms. All models are exported to ONNX from the same trained checkpoints and evaluated at 256×256. Table 6 reports compilation status, throughput, and power.

Fixed-function NPU. All models are compiled via vanilla INT8 post-training quantization (64-sample calibration, no QAT). The $\sim$1.5 dB PSNR drop (App. G.1) is a conservative lower bound on quantization robustness. fig. 6 shows that within the training range, the INT8 output is visually indistinguishable from the FP32 reference.

Under our evaluation protocol (standard ONNX export and vendor compiler with INT8 post-training quantization), only PhysDNet and AOD-Net successfully compile, but AOD-Net's restoration quality is insufficient for deployment (20.95 dB, §3.2). All other baselines fail due to unsupported operators (e.g., attention, LayerNorm). PhysDNet is thus the only method that combines competitive accuracy with NPU compatibility, while exceeding the 10 Hz LiDAR rate.

Jetson Orin Nano. All primary baselines compile under CUDA+TensorRT. PhysDNet-M runs $1.25\times$ faster than FFA-Net and $2.05\times$ faster than DEA-Net; PhysDNet-L remains $1.36\times$ faster than DEA-Net. AOD-Net's 327 FPS lacks deployable quality (§3.2).

4 Related Work

Physics-Guided Dehazing Classical dehazing combines the atmospheric scattering model with image priors such as the Dark Channel Prior [24]. Learned physics-guided methods incorporate atmospheric structure through model-based parameterizations or consistency objectives [3, 34, 35]. PhysDNet emphasizes the optimization effect of differentiating through a fixed inversion: its transmission-dependent Jacobian shapes supervision in the shared backbone. The Direct Head remains unconstrained by the inversion at inference, allowing residual corrections beyond the as-

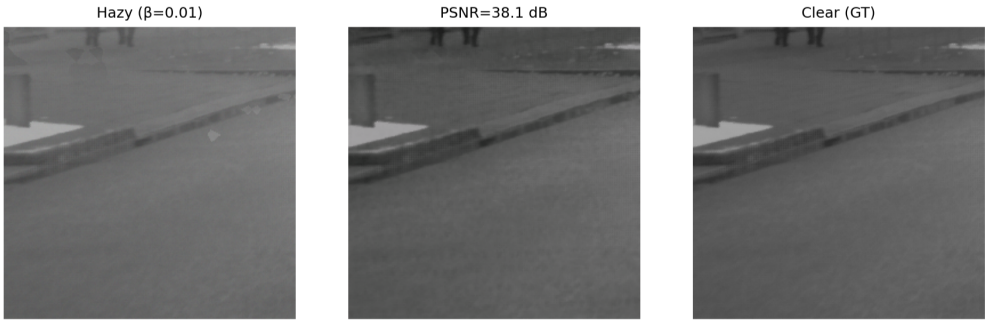

Figure 6: **INT8 NPU inference on a representative STF sample** (β=0.01). Hazy input, PhysDNet-M restored on the fixed-function NPU (38.1 dB), and ground truth. INT8 deployment preserves restoration quality within the training range; full β-sweep in App. G.1.

sumed image-formation model. This separates the role of physics during training from the form of the deployed restoration output.

Image-Only and Depth-Guided Dehazing Image-only CNN and transformer methods infer restoration cues from appearance [4–6]. Depth-aware approaches additionally exploit binocular disparity, monocular depth estimation, or LiDAR-derived geometry [36–39]. PhysDNet directly combines gated NIR imagery with sparse LiDAR covering approximately 3% of image pixels, using mask-aware multi-scale gating without a depth-completion stage. LiDAR is treated as a weather-attenuated geometric prior: we assume retained strongest returns provide useful range measurements, while missing returns are represented explicitly by a validity mask. Neither complete coverage nor immunity to fog is assumed. The zero-depth experiments in Appendix O quantify operation when this prior is unavailable.

Edge Deployment and Baseline Selection We target single-pass restoration within 33 ms and 15 W, compatible with fixed-function INT8 accelerators [8], which motivates our comparison with compact, toolchain-supported architectures. Recent heavyweight [25–27] and diffusion-based [40, 7, 29, 41, 28, 42] methods trade compute for quality; we measure rather than assume their cost (App. C). The fastest, Ksformer [27], reaches 2.8 FPS in TensorRT FP16, while DehazeDDPM [29] and RSHazeDiff [41] need an estimated 3.1 and 17.0 s per image. PhysDNet confines the inversion to training, adding no cost to the deployed Direct Head.

5 Conclusion

We introduced PhysDNet, a lightweight multimodal dehazing architecture that uses a physics-constrained dual-head objective to shape optimization dynamics. Differentiating through a fixed Koschmieder inversion induces density-adaptive gradient amplification, concentrating learning on severely degraded regions at zero inference cost. This mechanism yields state-of-the-art performance on Seeing Through Fog, consistent gains under cross-domain evaluation, and robustness to LiDAR sparsity. PhysDNet is also the only evaluated method that combines strong restoration quality with fixed-function low-power NPU deployment, achieving 29-99 FPS at $<$4 W under INT8 quantization.

Beyond dehazing, our results suggest that physical models can serve not only as priors on outputs, but as mechanisms for shaping the learning signal itself, offering a new paradigm for efficient and deployable perception under physical constraints.

Limitations. PhysDNet is trained on synthetic Koschmieder haze; NoPhy ablations on three real-fog benchmarks (O-HAZE, I-HAZE, NH-HAZE) confirm transfer, including under non-homogeneous fog (App. N). Broader cross-platform multi-modal validation beyond STF remains open. Without color supervision, the model produces near-achromatic RGB outputs requiring heuristic chrominance recovery (§3.6).

Acknowledgments

This work was partially supported by the Ministry of Education and Science of Bulgaria, as part of the Bulgarian National Roadmap for Research Infrastructure (support for INSAIT) and the European Union's Horizon Europe – the Framework Programme for Research and Innovation, under grant agreement 101168521.
Rayen Dhahri's research was self-funded, including all computational resources used for this work.

References

[1] Mario Bijelic, Tobias Gruber, Fahim Mannan, Florian Kraus, Werner Ritter, Klaus Dietmayer, and Felix Heide. Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 11679–11689. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.01170. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Bijelic_Seeing_Through_Fog_Without_Seeing_Fog_Deep_Multimodal_Sensor_Fusion_CVPR_2020_paper.html.

[2] Tobias Gruber, Mariia Kokhova, Werner Ritter, Norbert Haala, and Klaus Dietmayer. Learning super-resolved depth from active gated imaging. *CoRR*, abs/1912.02889, 2019. URL http://arxiv.org/abs/1912.02889.

[3] Boyi Li, Xiulian Peng, Zhangyang Wang, Jizheng Xu, and Dan Feng. Aod-net: All-in-one dehazing network. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 4780–4788. IEEE Computer Society, 2017. doi: 10.1109/ICCV.2017.511. URL https://doi.org/10.1109/ICCV.2017.511.

[4] Xu Qin, Zhilin Wang, Yuanchao Bai, Xiaodong Xie, and Huizhu Jia. Ffa-net: Feature fusion attention network for single image dehazing. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 11908–11915. AAAI Press, 2020. doi: 10.1609/AAAI.V34I07.6865. URL https://doi.org/10.1609/aaai.v34i07.6865.

[5] Yuda Song, Zhuqing He, Hui Qian, and Xin Du. Vision transformers for single image dehazing. *IEEE Transactions on Image Processing*, 32:1927–1941, 2023. ISSN 1941-0042. doi: 10.1109/tip.2023.3256763. URL http://dx.doi.org/10.1109/TIP.2023.3256763.

[6] Zixuan Chen, Zewei He, and Zhe-Ming Lu. Dea-net: Single image dehazing based on detail-enhanced convolution and content-guided attention. *IEEE Trans. Image Process.*, 33:1002–1015, 2024. doi: 10.1109/TIP.2024.3354108. URL https://doi.org/10.1109/TIP.2024.3354108.

[7] Chengxu Liu, Lu Qi, Jinshan Pan, Xueming Qian, and Ming-Hsuan Yang. Frequency domain-based diffusion model for unpaired image dehazing. *arXiv preprint arXiv:2507.01275*, 2025. doi: 10.48550/arXiv.2507.01275. URL https://arxiv.org/abs/2507.01275.

[8] Rayen Dhahri and Steffen Urban. Quant-trim in practice: Improved cross-platform low-bit deployment on edge npus, 2025. URL https://arxiv.org/abs/2511.15300.

[9] Earl J. McCartney. *Optics of the Atmosphere: Scattering by Molecules and Particles*. John Wiley & So ns, New York, 1976.

[10] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence*, 33(12):2341–2353, 2010.

[11] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells III, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention*

- MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III, Lecture Notes in Computer Science, pages 234–241. Springer, 2015. doi: 10.1007/978-3-319-24574-4_28. URL https://doi.org/10.1007/978-3-319-24574-4_28.

[12] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications, 2017. URL https://arxiv.org/abs/1704.04861.

[13] Tao Sheng, Chen Feng, Shaojie Zhuo, Xiaopeng Zhang, Liang Shen, and Mickey Aleksic. A quantization-friendly separable convolution for mobilenets. In *2018 1st Workshop on Energy Efficient Machine Learning and Cognitive Computing for Embedded Applications (EMC2)*, page 14–18. IEEE, March 2018. doi: 10.1109/emc2.2018.00011. URL http://dx.doi.org/10.1109/EMC2.2018.00011.

[14] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. *CoRR*, abs/1709.01507, 2017. URL http://arxiv.org/abs/1709.01507.

[15] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A. González. Gated multimodal networks. *Neural Comput. Appl.*, 32(14):10209–10228, July 2020. ISSN 0941-0643. doi: 10.1007/s00521-019-04559-1. URL https://doi.org/10.1007/s00521-019-04559-1.

[16] Yanhua Cheng, Rui Cai, Zhiwei Li, Xin Zhao, and Kaiqi Huang. Locality-sensitive deconvolution networks with gated fusion for rgb-d indoor semantic segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1475–1483, 2017. doi: 10.1109/CVPR.2017.161.

[17] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining, 2016. URL https://arxiv.org/abs/1604.03540.

[18] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection, 2018. URL https://arxiv.org/abs/1708.02002.

[19] Vincent Mai, Waleed Khamies, and Liam Paull. Batch inverse-variance weighting: Deep heteroscedastic regression, 2021. URL https://arxiv.org/abs/2107.04497.

[20] Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *IEEE Transactions on Image Processing*, 28 (1):492–505, 2019. doi: 10.1109/TIP.2018.2867951.

[21] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

[22] Edwin H. Land. The retinex theory of color vision. *Scientific American*, 237(6):108–128, 1977. doi: 10.1038/scientificamerican1277-108.

[23] Wenhan Yang, Wenjing Wang, Haofeng Huang, Shiqi Wang, and Jiaying Liu. Sparse gradient regularized deep retinex network for robust low-light image enhancement. *IEEE Transactions on Image Processing*, 30:2072–2086, 2021. doi: 10.1109/TIP.2021.3050850.

[24] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 1956–1963. IEEE Computer Society, 2009. doi: 10.1109/CVPR.2009.5206515. URL https://doi.org/10.1109/CVPR.2009.5206515.

[25] Zhuoran Zheng and Chen Wu. U-shaped vision mamba for single image dehazing, 2024. URL https://arxiv.org/abs/2402.04139.

[26] Huazhong Jin, Zhang Chen, Zhina Song, and Kaimin Sun. Dffnet: A dual-domain feature fusion network for single remote sensing image dehazing. *Sensors*, 25(16), 2025. ISSN 1424-8220. doi: 10.3390/s25165125. URL https://www.mdpi.com/1424-8220/25/16/5125.

[27] Lihan TONG, Weijia LI, Qingxia YANG, Liyuan CHEN, and Peng CHEN. Vision transformer with key-select routing attention for single image dehazing. *IEICE Transactions on Information and Systems*, E107.D, 07 2024. doi: 10.1587/transinf.2024EDL8043.

[28] Xiaoxia Lin, Zhengao Li, Dawei Huang, Wancheng Feng, XinJun An, Lin Sun, Niuzhen Yu, Yan Li, and Chunwei Leng. HazeDiff: A training-free diffusion-based image dehazing method with pixel-level feature injection. *PLOS One*, 20(10):e0329759, 2025. doi: 10.1371/journal.pone.0329759.

[29] Hu Yu, Jie Huang, Kaiwen Zheng, and Feng Zhao. High-quality image dehazing with diffusion model. *arXiv preprint arXiv:2308.11949*, 2023.

[30] Muhammad Yaseen. What is yolov8: An in-depth exploration of the internal features of the next-generation object detector, 2024. URL https://arxiv.org/abs/2408.15857.

[31] Codruta O. Ancuti, Cosmin Ancuti, Radu Timofte, and Christophe De Vleeschouwer. O-haze: a dehazing benchmark with real hazy and haze-free outdoor images, 2018. URL https://arxiv.org/abs/1804.05101.

[32] Cosmin Ancuti, Codruta O. Ancuti, Radu Timofte, and Christophe De Vleeschouwer. I-HAZE: A dehazing benchmark with real hazy and haze-free indoor images. In Jacques Blanc-Talon, David Helbert, Wilfried Philips, Dan C. Popescu, and Paul Scheunders, editors, *Advanced Concepts for Intelligent Vision Systems, 19th International Conference, ACIVS 2018, Poitiers, France, September 24-27, 2018, Proceedings*, Lecture Notes in Computer Science, pages 620–631. Springer, 2018. doi: 10.1007/978-3-030-01449-0_52. URL https://doi.org/10.1007/978-3-030-01449-0_52.

[33] David Hasler and Sabine E. Suesstrunk. Measuring colorfulness in natural images. In *Human Vision and Electronic Imaging VIII*, volume 5007 of *Proceedings of SPIE*, pages 87–95, 2003. doi: 10.1117/12.477378.

[34] Yiming Yue, Long Ma, and Peng Li. PEDNet: Physics-guided encoder–decoder network for image dehazing. *Journal of Imaging Science and Technology*, 69(6):060503–1–060503–7, 2025. doi: 10.2352/J.ImagingSci.Technol.2025.69.6.060503. URL https://library.imaging.org/jist/articles/69/6/060503.

[35] Shan Liang, Tao Gao, Ting Chen, Qianxi Zhang, Xiao Wang, and Xiaowei Li. A physics prompt-based network for interpretable and effective image dehazing. *Expert Systems with Applications*, 291:128511, 2025. doi: 10.1016/j.eswa.2025.128511. URL https://www.sciencedirect.com/science/article/abs/pii/S095741742502130X.

[36] Yanwei Pang, Jing Nie, Jin Xie, Jungong Han, and Xuelong Li. Bidnet: Binocular image dehazing without explicit disparity estimation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 5930–5939. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.00597. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Pang_BidNet_Binocular_Image_Dehazing_Without_Explicit_Disparity_Estimation_CVPR_2020_paper.html.

[37] Yudong Liang, Shaoji Li, De Cheng, Wenjian Wang, Deyu Li, and Jiye Liang. Image dehazing via self-supervised depth guidance. *Pattern Recognit.*, 158:111051, 2025. doi: 10.1016/J.PATCOG.2024.111051. URL https://doi.org/10.1016/j.patcog.2024.111051.

[38] Won Young Chung, Sun Young Kim, and Chang Ho Kang. Image dehazing using lidar generated grayscale depth prior. *Sensors*, 22(3):1199, 2022. doi: 10.3390/S22031199. URL https://doi.org/10.3390/s22031199.

[39] Le-Anh Tran and Dong-Chul Park. Encoder-decoder network with guided transmission map: Architecture - extended abstract. *CoRR*, abs/2212.05936, 2022. doi: 10.48550/ARXIV.2212.05936. URL https://doi.org/10.48550/arXiv.2212.05936.

[40] Yunwei Lan, Zhigao Cui, Chang Liu, Jialun Peng, Nian Wang, Xin Luo, and Dong Liu. Exploiting diffusion prior for real-world image dehazing with unpaired training. *arXiv preprint arXiv:2503.15017*, 2025. doi: 10.48550/arXiv.2503.15017. URL https://arxiv.org/abs/2503.15017.

[41] Jiamei Xiong, Xuefeng Yan, Yongzhen Wang, Wei Zhao, Xiao-Ping Zhang, and Mingqiang Wei. RSHazeDiff: A unified fourier-aware diffusion model for remote sensing image dehazing. *IEEE Transactions on Intelligent Transportation Systems*, 26(1):1055–1070, 2025.

[42] Ziwei Luo, Fredrik K. Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B. Schön. Image restoration with mean-reverting stochastic differential equations. In *Proceedings of the 40th International Conference on Machine Learning*, pages 23045–23066. PMLR, 2023.

[43] Junbao Zheng, Chenke Xu, Wei Zhang, and Xu Yang. Single image dehazing using global illumination compensation. *Sensors*, 22(11), 2022. ISSN 1424-8220. doi: 10.3390/s22114169. URL https://www.mdpi.com/1424-8220/22/11/4169.

[44] Simone Bianco, Luigi Celona, and Flavio Piccoli. Single image dehazing by predicting atmospheric scattering parameters. volume 2020, 09 2020. doi: 10.2352/issn.2694-118X.2020.LIM-11.

[45] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics, 2018. URL https://arxiv.org/abs/1705.07115.

[46] Aaron Fisher. Inverse-variance weighting for estimation of heterogeneous treatment effects. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 13706–13733. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/fisher24a.html.

[47] Vien V. Mai and Mikael Johansson. Stability and convergence of stochastic gradient clipping: Beyond lipschitz continuity and smoothness, 2021. URL https://arxiv.org/abs/2102.06489.

[48] Dongruo Zhou, Jinghui Chen, Yuan Cao, Ziyan Yang, and Quanquan Gu. On the convergence of adaptive gradient methods for nonconvex optimization, 2024. URL https://arxiv.org/abs/1808.05671.

[49] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In Andrea Pohoreckyj Danyluk, Léon Bottou, and Michael L. Littman, editors, *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, ACM International Conference Proceeding Series, pages 41–48. ACM, 2009. doi: 10.1145/1553374.1553380. URL https://doi.org/10.1145/1553374.1553380.

[50] Codruta Orniana Ancuti, Cosmin Ancuți, and Radu Timofte. Nh-haze: An image dehazing benchmark with non-homogeneous hazy and haze-free images. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1798–1805, 2020.

[51] Anish Mittal, Rajiv Soundararajan, and Alan Conrad Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2012.

[52] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *Trans. Img. Proc.*, 21(12):4695–4708, December 2012. ISSN 1057-7149. doi: 10.1109/TIP.2012.2214050. URL https://doi.org/10.1109/TIP.2012.2214050.

[53] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *Computer Vision – ECCV 2012*, pages 746–760, 2012. doi: 10.1007/978-3-642-33715-4_54.

[54] Derya Akkaynak and Tali Treibitz. A revised underwater image formation model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6723–6732, 2018.

A Broader impact

PhysDNet improves the ability of resource-constrained edge systems to perceive in adverse weather, with direct applications to autonomous driving, robotics, and ADAS. By recovering semantically meaningful structure and improving downstream object detection in dense fog, it may help reduce weather-related accidents and improve road safety. Its design also supports *Green AI* by enabling high-quality, physics-grounded restoration on ultra-low-power fixed-function NPUs ($<$ **4 W**), reducing reliance on energy-intensive GPU inference.

At the same time, this technology has dual-use potential and could support unconsented surveillance or military deployment in conditions where poor visibility would otherwise provide cover. It also introduces safety risks in critical systems if exposed to out-of-distribution weather, such as heterogeneous smoke or severe occlusions that violate the assumed Koschmieder model. For this reason, PhysDNet should be used only as part of a sensor-fusion stack, together with uncertainty estimation and hardware fail-safes.

B Extended Background

Ill-posed nature of monocular dehazing. Recovering $J(x)$ from a single $I(x)$ is fundamentally ill-posed: both $t(x)$ and A are unknown, and hand-crafted priors such as the Dark Channel Prior [10] frequently collapse under non-homogeneous real-world fog.

Active sensing and gated imaging mechanics. Modern autonomous platforms combine sparse LiDAR, which provides weather-invariant depth measurements $d(x)$, with active time-gated NIR imaging [2]. By delaying sensor exposure after emitting a synchronised NIR laser pulse, a gated camera physically rejects nearby backscatter and integrates only photons from a target depth slice, drastically improving the signal-to-noise ratio in dense fog. This gated imagery combined with sparse LiDAR forms the physical foundation of PhysDNet's multi-modal input.

C Baseline Selection Rationale

Table 7 measures every candidate under our edge protocol (Jetson Orin Nano Super, TensorRT FP16 where export succeeds). The primary baselines meet the single-pass deployment criterion. The fastest recent heavyweight method that runs, Ksformer, reaches 2.8 FPS, UVM-Net cannot build an engine, and the diffusion methods need 3.1–17 s per image. Even PhysDNet-L runs at 38.8 FPS within the latency and power targets.

D Architecture Details

Each encoder stage consists of two DWSep blocks with a residual shortcut, followed by $2\times$ max-pooling; the four stages produce feature maps at H, $H/2$, $H/4$, and $H/8$ (bottleneck). The image branch accepts a 3-channel input (monochrome NIR replicated or colour); the depth branch accepts 2 channels (sparse depth D_s and binary validity mask M). A Squeeze-and-Excitation (SE) channel attention block [14] with reduction ratio 4 is applied once at the bottleneck after gated fusion, recalibrating channel responses before decoding. The decoder consists of three upsampling stages, each performing bilinear $2\times$ upsampling followed by channel-wise concatenation with the corresponding fused skip features and refinement through a DWSep residual block. The bottleneck fused features ($8c$ channels at $H/8$) serve as the decoder input; skip connections from scales 3, 2, and 1 progressively reintroduce spatial detail. The head projections use 1×1 convolutions throughout; airlight is predicted via global average pooling of the bottleneck followed by a 3-unit fully-connected layer.

E Training Details and Hyperparameters

All PhysDNet variants share a common training recipe and loss weights, differing only in base channel width c. The **L** variant additionally enables mixed-precision (AMP) to fit within VRAM at c=96.

Table 7: **Measured Edge Deployment and Baseline Selection.** Compute is profiled at 256×256, batch 1. Jetson results report latency, throughput, and mean module power. Successful exported models use TensorRT FP16; DFFNet uses PyTorch eager FP32. Diffusion per-image costs are derived from measured network-call latency and the stated number of network evaluations. Unmeasured related methods are listed separately.

Method	Params (M)	Compute (GMACs/image)	Jetson latency (ms/image)	FPS	Power (W)	Explanation
Primary Comparison Baselines						
DCP [24]	0	—	—	—	—	Classical physical-prior baseline.
AOD-Net [3]	0.002	0.06	3.1	327.0	4.94	Ultra-lightweight model-based CNN.
DehazeFormer-T [5]	1.36	2.73	21.2	47.1	12.38	Compact transformer baseline.
DEA-Net [6]	3.65	16.44	35.1	28.5	10.17	Competitive CNN baseline; exceeds the latency target.
FFA-Net [4]	0.77	24.68	21.3	46.9	8.21	Feature-fusion attention baseline.
Measured Heavyweight Baselines						
UVM-Net [25]†	1003.9	250.9	—	—	—	TensorRT engine construction fails: out-of-memory in the tested implementation.
DFFNet [26]†	17.8	61.3	157.0	6.4	—	ONNX export fails on FFT operations; latency measured in eager FP32.
Ksformer [27]†	5.5	109.0	360.0	2.8	19.8	TensorRT FP16; exceeds both latency and power targets.
Measured Diffusion Backbones and Sampling Costs						
DehazeDDPM [29]‡	97.8	4 509	≈3 100	0.32	20.6	50 network evaluations per image.
RSHazeDiff [41]‡	109.7	13 523	≈17 000	0.06	16.0	10 steps across 169 overlapping patches: 1,690 network evaluations per image.
PhysDNet Deployment Reference						
PhysDNet-L	6.59	18.98	**25.8**	**38.8**	11.80	TensorRT FP16; Even our largest variant satisfies the stated Jetson latency and power targets.
Other Related Methods: Qualitative Positioning						
HazeDiff [28]	—	—	—	—	—	Iterative generative restoration approaches outside the primary single-pass comparison. Retained for related-work positioning; not included in the measured benchmark above.
IR-SDE [42]	—	—	—	—	—	
FrDiff [7]	—	—	—	—	—	

Protocol. Jetson Orin Nano Super, MAXN, batch 1; engine construction excluded from inference timing; mean module power recorded with `tegrastats`. Primary learned-baseline rows use the LiDAR-augmented configurations from Table 6. Their displayed latency is calculated as $1000/\mathrm{FPS}$.

†DFFNet and Ksformer are reimplementations matched to published architecture descriptions and parameter budgets; their random weights support compute and latency profiling, not restoration-quality comparisons. UVM-Net uses an official-code port with a pure-PyTorch selective-scan fallback; its failure applies to this tested path.

‡DehazeDDPM: 90.2 GMACs and 62.7 ms per call, using a 50-call benchmark setting. RSHazeDiff: 8.0 GMACs and approximately 10.0 ms per call, using 10×169 patch evaluations. Per-image compute and latency multiply per-call measurements by these evaluation counts; latency excludes sampler and patch-aggregation overhead. Totals use unrounded measurements. Dashes indicate unavailable or unmeasured results.

Table 8: **Training recipe.** Shared across all variants except where noted. † AMP enabled for PhysDNet-L only.

Hyperparameter	PhysDNet-S	PhysDNet-M	PhysDNet-L
Base channel width c	32	64	96
SE attention	✓	✓	✓
Epochs	500	500	500
Batch size	16	16	8†
Optimizer	AdamW		
Learning rate	2×10^{-4}		
Weight decay	1×10^{-4}		
LR scheduler	Cosine annealing $\rightarrow 10^{-6}$		
Warmup epochs	10		
Gradient clipping	$\|\nabla\|_2 \leq 1.0$		
Crop size	256×256		
Data augmentation	Random flip + crop		
Mixed precision (AMP)	×	×	✓
β range (haze synthesis)	$\mathcal{U}(0.005,\ 0.04)$		
A_k range (airlight per channel)	$\mathcal{U}(0.7,\ 1.0)$		
Max LiDAR depth	120 m		

Table 9: **Loss component weights (λ).** All variants share the same loss weights and contrastive activation schedule (epoch 185); they differ only in base channel width c. Data-driven / Physics / Auxiliary shading indicates loss family.

Loss	Symbol	Supervises	S	M	L	Role
Reconstruction	$\mathcal{L}_{\text{rec}}$	$\hat{J}_{\text{dir}}$ vs. J_{gt}	1.0	1.0	1.0	Primary restoration
Perceptual	$\mathcal{L}_{\text{perc}}$	VGG-16 features	0.05	0.05	0.05	Feature-space similarity
Gradient	$\mathcal{L}_{\text{grad}}$	$\nabla\hat{J}_{\text{dir}}$ vs. ∇J_{gt}	0.5	0.5	0.5	Edge preservation
Physics recon.	$\mathcal{L}_{\text{pr}}$	$\hat{J}_{\text{phys}}$ vs. J_{gt}	1.0	1.0	1.0	Physics-head fidelity
Transmission	$\mathcal{L}_{t}$	$\hat{t}$ vs. t_{gt}	2.0	2.0	2.0	LiDAR-supervised t
Smoothness	$\mathcal{L}_{\text{sm}}$	$\nabla\hat{t}$	0.1	0.1	0.1	Edge-aware regularization
Consistency	$\mathcal{L}_{\text{con}}$	$\hat{J}_{\text{dir}}$ vs. $\hat{J}_{\text{phys}}$	0.2	0.2	0.2	Dual-head agreement
FFT	$\mathcal{L}_{\text{fft}}$	Frequency domain	0.1	0.1	0.1	High-frequency detail
Contrastive	$\mathcal{L}_{\text{contra}}$	Latent features	0.1	0.1	0.1	Physics-aware bound
start epoch			185	185	185	Delayed activation

Table 10: **PhysDNet model family.** Scaling the base channel width c yields three variants spanning the accuracy–efficiency Pareto front. All variants share the same dual-encoder, gated-fusion, SE-decoder topology (§2.1) and dual-head design (§2.2). Physics head parameters are excluded from the inference count as only the direct head is deployed.

Variant	c	SE Attn.	Params (M)	GMACs	Bottleneck	Target
PhysDNet-S	32	✓	0.72	2.24	$256 \times \frac{H}{8} \times \frac{W}{8}$	Ultra-low-power NPU
PhysDNet-M	64	✓	2.95	8.56	$512 \times \frac{H}{8} \times \frac{W}{8}$	Edge GPU / mid-range NPU
PhysDNet-L	96	✓	6.59	18.98	$768 \times \frac{H}{8} \times \frac{W}{8}$	High-perf. edge accelerator

Convergence and Training Budgets. All models were trained until performance plateaued under their respective training configurations, subject to a fixed computational budget. In practice, we observe that PSNR improvements saturate after approximately 80 epochs for most baselines, with additional training yielding only marginal gains (typically < 0.3 dB). This behavior is consistent across all evaluated methods, indicating that extended training primarily increases compute cost without significantly affecting relative performance rankings. PhysDNet variants were trained for 500 epochs to ensure the complex multi-term objective (including late-activated contrastive losses) fully converged, though primary restoration metrics similarly plateaued early in the schedule. Experiments require 24 GB VRAM GPU.

F Gradient Weighting under the Scattering Model: Full Derivation

Section 2.3 empirically characterised the $(\hat{t}+\epsilon)^{-2}$ gradient structure induced by differentiating the Koschmieder inversion and briefly noted its structural correspondence to the inversion's noise variance. This appendix provides the formal derivations of this variance-gradient relationship, alongside the airlight-gradient analysis, the bounded-gradient discussion, and a scope statement delineating what the analysis does and does not claim.

Setup. We adopt the standard atmospheric scattering model with additive sensor noise [24, 43, 44]:

$$I(x) = J(x)\,t(x) + A\big(1-t(x)\big) + \xi(x), \qquad \xi(x) \sim \mathcal{N}(0,\sigma^2), \tag{8}$$

with $\xi(x)$ independent across pixels. Given estimates $\hat{t}(x), \hat{A}$, the physics head computes the ϵ-stabilised inversion

$$\hat{J}_{\text{phys}}(x) \;=\; \frac{I(x) - \hat{A}\big(1-\hat{t}(x)\big)}{\hat{t}(x)+\epsilon}, \tag{9}$$

consistent with Eq. (5) in the main text.

Proposition 1 (Transmission-Scaled Noise Variance). *Under Eq.* (8) *and first-order estimates* $\hat{t}(x) = t(x) + \delta_t$, $\hat{A} = A + \delta_A$ *with* $|\delta_t|, |\delta_A| = O(\delta)$ *and* $\delta \ll t_{\min}$, *the inverted radiance admits*

$$\hat{J}_{\text{phys}}(x) \;=\; J(x) + \frac{\xi(x)}{t(x)} + O(\delta), \qquad \text{Var}\big[\hat{J}_{\text{phys}}(x)\big] \;=\; \frac{\sigma^2}{t(x)^2} + O(\delta). \tag{10}$$

Proof. Substituting Eq. (8) into Eq. (9):

$$\hat{J}_{\text{phys}} \;=\; \frac{Jt + A(1-t) + \xi - (A+\delta_A)\big(1-(t+\delta_t)\big)}{(t+\delta_t)+\epsilon}.$$

Expanding the bracket in the numerator gives $(A+\delta_A)(1-t-\delta_t) = A(1-t) - A\delta_t + \delta_A(1-t) - \delta_A\delta_t$, so the $A(1-t)$ terms cancel and

$$\text{numerator} \;=\; Jt + \xi + A\delta_t - \delta_A(1-t) + O(\delta^2).$$

For the denominator, since $\delta, \epsilon \ll t$, we have the first-order expansion

$$\frac{1}{t+\delta_t+\epsilon} \;=\; \frac{1}{t}\cdot\frac{1}{1+(\delta_t+\epsilon)/t} \;=\; \frac{1}{t} - \frac{\delta_t+\epsilon}{t^2} + O(\delta^2,\epsilon^2).$$

Multiplying numerator and denominator and retaining terms up to $O(\delta)$:

$$\begin{aligned}
\hat{J}_{\text{phys}} \;&=\; \big[Jt + \xi + A\delta_t - \delta_A(1-t)\big]\cdot\tfrac{1}{t} \;-\; Jt\cdot\tfrac{\delta_t+\epsilon}{t^2} \;+\; O(\delta^2,\epsilon) \\
&=\; J + \tfrac{\xi}{t} + \tfrac{A\delta_t - \delta_A(1-t)}{t} \;-\; \tfrac{J\delta_t}{t} + O(\delta^2,\epsilon) \\
&=\; J + \tfrac{\xi}{t} - \tfrac{(J-A)\delta_t + (1-t)\delta_A}{t} + O(\delta^2,\epsilon),
\end{aligned}$$

where the cross-terms (e.g. $\xi\cdot O(\delta,\epsilon)$, $A\delta_t\cdot O(\delta,\epsilon)$) all collapse into $O(\delta^2,\epsilon)$. The deterministic perturbation terms are $O(\delta)$; only ξ/t contributes stochastic variance. Since $\xi \sim \mathcal{N}(0,\sigma^2)$, $\text{Var}[\hat{J}_{\text{phys}}] = \sigma^2/t^2 + O(\delta)$. □

Quantitative validation of Proposition 1. On 300 STF validation frames (clear day/night splits) with synthetic haze and Gaussian sensor noise ($\sigma\approx0.01$), we bin the per-pixel squared error of the parameter-free inversion of the released PhysDNet-M by ground-truth transmission (24 bins, clamp-saturated pixels excluded). The empirical curve follows σ^2/t^2 across two orders of magnitude (log-log r=0.981; Fig. 7); the small uniform offset is consistent with the $O(\delta)$ estimation term. This is an MSE trend consistent with the predicted variance law, not a direct variance measurement on natural fog. It also identifies the low-transmission noise visible in Physics Head panels (Fig. 13) as the predicted structure, which the Direct Head does not inherit.

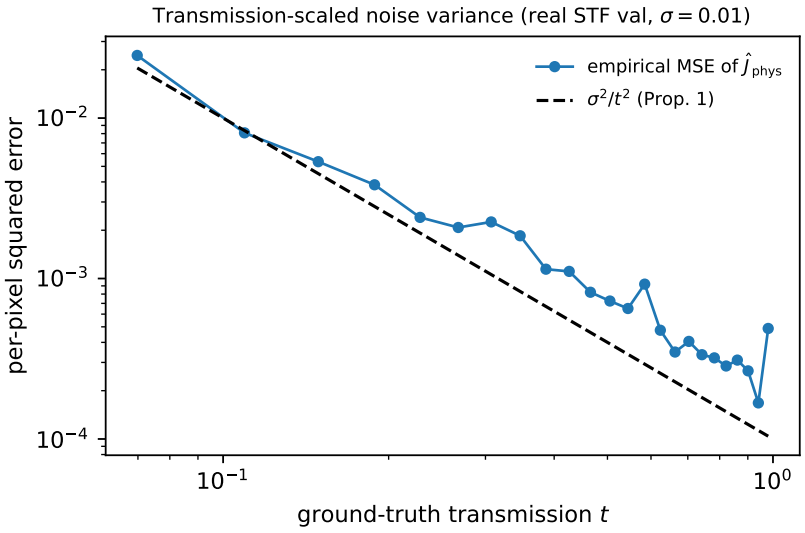

Figure 7: **Empirical test of Proposition 1.** Binned squared error of $\hat{J}_{\text{phys}}$ vs. ground-truth transmission, overlaid with σ^2/t^2.

Proposition 2 (Structured Jacobian Scaling). The Jacobian of the inversion with respect to transmission is

$$\frac{\partial \hat{J}_{\text{phys}}}{\partial \hat{t}}(x) = \frac{\hat{A} - I(x)}{(\hat{t}(x) + \epsilon)^2}, \tag{11}$$

treating $(I, \hat{A})$ as fixed with respect to $\hat{t}$.

Derivation. Write $\hat{J}_{\text{phys}} = u/v$ with $u(\hat{t}) = I - \hat{A}(1 - \hat{t}) = I - \hat{A} + \hat{A}\hat{t}$ and $v(\hat{t}) = \hat{t} + \epsilon$. Then $\partial u/\partial \hat{t} = \hat{A}$, $\partial v/\partial \hat{t} = 1$, and the quotient rule yields

$$\frac{\partial \hat{J}_{\text{phys}}}{\partial \hat{t}} = \frac{\hat{A}\,(\hat{t} + \epsilon) - \big[I - \hat{A}(1 - \hat{t})\big]}{(\hat{t} + \epsilon)^2} = \frac{\hat{A} - I + \hat{A}\,\epsilon}{(\hat{t} + \epsilon)^2} = \frac{\hat{A} - I}{(\hat{t} + \epsilon)^2} + O(\epsilon),$$

which gives Eq. (11) after dropping the $O(\epsilon)$ term consistent with the main text. □

This expression reveals a *structural* inverse-square dependence on transmission through the denominator $(\hat{t} + \epsilon)^{-2}$. Consequently, the computational graph intrinsically amplifies gradients in low-transmission (dense fog) regions. The effective gradient magnitude additionally depends on the residual term $(\hat{A} - I(x))$, which introduces a data-dependent modulation. Under the scattering model, $(\hat{A} - I(x)) \approx t(x)(A - J(x))$ when estimates are accurate, partially attenuating the scaling in well-reconstructed regions. Nevertheless, the dominant $(\hat{t} + \epsilon)^{-2}$ structure ensures that low-transmission pixels receive substantially larger gradient magnitudes throughout training. Importantly, this amplification arises directly from differentiating the physical inversion and does not rely on explicit weighting heuristics, auxiliary parameters, or curriculum schedules.

Corollary (Loss-gradient scaling). Under $\mathcal{L}_{\text{pr}} = \|\hat{J}_{\text{phys}} - J_{\text{gt}}\|_1$,

$$\frac{\partial \mathcal{L}_{\text{pr}}}{\partial \hat{t}(x)} = \operatorname{sign}\big(\hat{J}_{\text{phys}}(x) - J_{\text{gt}}(x)\big) \cdot \frac{\hat{A} - I(x)}{(\hat{t}(x) + \epsilon)^2}, \tag{12}$$

and under an ℓ_2 loss the sign is replaced by $2\big(\hat{J}_{\text{phys}} - J_{\text{gt}}\big)$. In both cases the amplification propagates unchanged into the encoder weights via $\partial \hat{t}/\partial \theta_{\text{enc}}$.

Airlight gradient: physically correct supervision routing. The magnitude of the airlight gradient is $|\partial \hat{J}_{\text{phys}}/\partial \hat{A}| = |1 - \hat{t}|/(\hat{t} + \epsilon)$, which scales as $\hat{t}^{-1}$ for small $\hat{t}$ and vanishes as $\hat{t} \to 1$. The network

therefore learns to estimate airlight primarily from foggy regions, where atmospheric light dominates the observation, and ignores clear-air pixels that carry no information about A. This supervision routing matches the physical reality: airlight is only observable where scattering is present.

Empirical spatial validation. Figure 8 confirms the $(\hat{t}+\epsilon)^{-2}$ scaling holds on real (non-synthetic) STF fog frames at the per-pixel level: the analytically predicted gradient magnitude map spatially aligns with low-transmission regions in the predicted $\hat{t}$, with high-gradient activations concentrated precisely where the physics model is most sensitive.

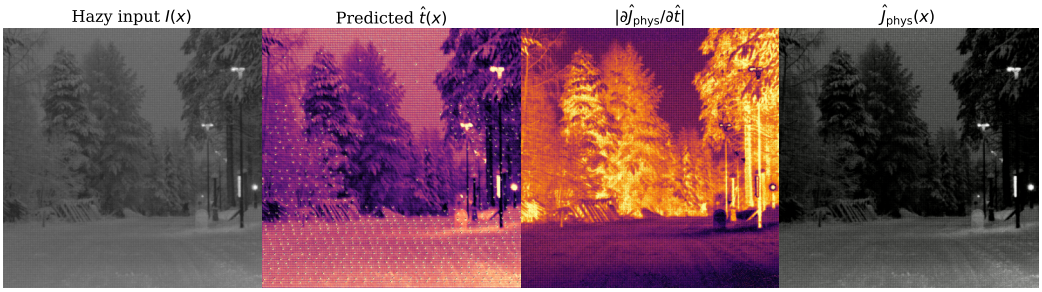

Figure 8: Empirical gradient map on a real STF fog frame. Left to right: hazy input, predicted transmission $\hat{t}$, analytical gradient magnitude map $|\partial \hat{J}_{\text{phys}}/\partial \hat{t}|$, and physics-restored output. High-gradient regions align spatially with low-transmission regions, empirically validating the density-adaptive behaviour predicted by Propositions 1–2.

Contrast with the direct-head gradient. The direct head's gradient through the shared encoder, $\partial \hat{J}_{\text{dir}}/\partial \mathbf{z} = \sigma'(\cdot)\, W_{\text{dir}}$, is spatially uniform and has no notion of local fog density. The physics head provides spatially structured, density-adaptive magnitudes. Both signals flow into the same encoder, but the physics path shapes features toward physically meaningful representations even though only the direct head is used at inference.

Interpretive connection to inverse-variance weighting. Proposition 1 establishes that the stochastic sensitivity of the inversion scales as t^{-2}. Classical heteroscedastic regression [45, 19, 46] exploits this inversely, multiplying each sample's loss by $1/\operatorname{Var}[\hat{J}] \propto t^2$ to down-weight unreliable pixels. Differentiating through Eq. (9) applies the structural inverse, up-weighting low-transmission regions. The two schemes are not in conflict; they operate on different signal models. Inverse-variance weighting is optimal when the signal is dominated by irreducible noise. Under the scattering model, the numerator $(\hat{A} - I(x)) = -(J - A)\, t + O(\xi)$ is dominated by the deterministic $(J - A)$ term in degraded regions where $(J - A)$ is large, so larger updates reduce estimation error rather than fit noise. We interpret the physics head as performing implicit hard-example mining along the transmission axis, empirically supported by the NoPhy ablation (§3.3).

Bounded gradients and stability. To ensure numerical stability, the transmission is constrained to $\hat{t} \in [t_{\min}, t_{\max}]$ with $t_{\min} = 0.05$ (Eq. 3). This bounds the Jacobian:

$$\left| \frac{\partial \hat{J}_{\text{phys}}}{\partial \hat{t}}(x) \right| \leq \frac{|\hat{A} - I(x)|}{(t_{\min} + \epsilon)^2} \leq \frac{1}{(t_{\min} + \epsilon)^2}, \tag{13}$$

since $\hat{A}, I(x) \in [0, 1]$. In practice, this upper bound is approximately 400, and is further controlled via gradient clipping (Appendix E). As a result, the optimization remains within a bounded-gradient regime compatible with standard convergence guarantees for stochastic gradient methods [47, 48].

Implicit curriculum. The transmission-dependent gradient scaling induces an implicit curriculum [49] over training samples. Early in training, low-transmission (high-degradation) regions dominate the gradient signal due to their larger magnitude. As the model improves and transmission estimates stabilize, the gradient distribution naturally rebalances without requiring explicit scheduling. This

behavior emerges directly from the physics-based formulation and does not introduce additional hyperparameters.

Scope. The analysis assumes (i) homogeneous single-scattering, (ii) Gaussian i.i.d. sensor noise, and (iii) first-order accuracy of $(\hat{t}, \hat{A})$. Real fog may violate (i) through non-uniform airlight and multiple scattering; low-light imagery may violate (ii) through Poisson-dominated shot noise; heavily degraded regions may violate (iii) when $\delta \not\ll t_{\min}$. The structural amplification arises directly from the $1/\hat{t}$ factor in Eq. (9), not from Gaussianity—and is therefore robust under (ii); the prefactor may shift under (i) and (iii). The cross-domain results on O-HAZE and I-HAZE (§3.6), where the NoPhy ablation degrades PSNR-Y by 1.53 and 0.60 dB on real non-synthetic haze, provides empirical evidence that the mechanism transfers beyond the assumed regime.

F.1 Loss Definitions

Let $\mathrm{Ch}(x, y) = \frac{1}{N}\sum \sqrt{(x-y)^2 + \varepsilon^2}$ (ε=10^{-6}), $\partial_{x,y}$ finite differences, ϕ_l frozen VGG-16 features at relu1_2, relu2_2, relu3_3, and ψ VGG-16 up to relu2_2.

Data-driven family (supervises the deployed output). $\mathcal{L}_{\text{rec}} = \mathrm{Ch}(\hat{J}_{\text{dir}}, J)$; $\mathcal{L}_{\text{perc}} = \sum_l \|\phi_l(\hat{J}_{\text{dir}}) - \phi_l(J)\|_1$; $\mathcal{L}_{\text{grad}} = \sum_{k\in\{x,y\}} \|\partial_k \hat{J}_{\text{dir}} - \partial_k J\|_1$.

Physics family (shapes optimization through the inversion). $\mathcal{L}_{\text{pr}} = \mathrm{Ch}(\hat{J}_{\text{phys}}, J)$, the term carrying the $(\hat{t}+\epsilon)^{-2}$ amplification; $\mathcal{L}_t = \|\hat{t} - t\|_1$ with $t = e^{-\beta \tilde{d}}$ from the synthesis depth, the anchor that localizes the amplified gradients; $\mathcal{L}_{\text{sm}} = \sum_{k\in\{x,y\}} \overline{e^{-10|\partial_k I|}\,|\partial_k \hat{t}|}$, with $|\partial_k I|$ averaged over channels; $\mathcal{L}_{\text{con}} = \mathrm{Ch}(\hat{J}_{\text{dir}}, \hat{J}_{\text{phys}})$.

Auxiliary family. $\mathcal{L}_{\text{fft}} = \big\||\mathcal{F}(\hat{J}_{\text{dir}})| - |\mathcal{F}(J)|\big\|_1$ (orthonormal real 2-D FFT); $\mathcal{L}_{\text{contra}} = \max\big(0, \|\psi(\hat{J}_{\text{dir}}) - \psi(J)\|_1 - \|\psi(\hat{J}_{\text{dir}}) - \psi(I)\|_1 + m\big)$, m=0.3, enabled from epoch 185.

The objective is $\mathcal{L} = \sum_i \lambda_i \mathcal{L}_i$ with the weights in Table 9, shared across S/M/L. NoPhy removes the entire physics family.

G Edge Deployment: Full Results

This appendix provides the full deployment measurements summarized in §3.9: INT8 quantization quality on the fixed-function NPU, qualitative comparison under quantization, and the full embedded-GPU benchmark including CUDA FP32 reference. numbers and onboard power profiles.

G.1 INT8 Quantization Quality on the Fixed-Function NPU

To verify that INT8 post-training quantization preserves restoration quality, we study PhysDNet-M in detail. We use the vendor compiler's default calibration pipeline with 64 real STF samples (β=0.02, A=0.8). **We do not apply quantization-aware training, mixed-precision layer policies, custom activation-range search, or any other vendor-specific optimization**; the reported gap is therefore representative of naive post-training quantization rather than a tuned result. Evaluation covers 200 held-out samples spanning $\beta \in \{0.005, \ldots, 0.10\}$ (Table 11).

Within the training range ($\beta \le 0.04$), INT8 incurs only a $\sim$1.5 dB average PSNR drop vs. FP32; at β=0.01 the quantized model comes within 0.8 dB of FP32. Out-of-distribution degradation ($\beta > 0.04$) follows the same graceful decline as FP32, confirming that the physics-guided head is not destabilized by aggressive quantization.

G.2 Embedded-GPU Full Benchmark

Table 12 reports the full benchmark under both CUDA FP32 and TensorRT FP16 execution providers on the Jetson Orin Nano. The TensorRT FP16 path yields 3–5$\times$ speedups over CUDA FP32 across the board with no measurable PSNR degradation, and is therefore the reference deployment configuration in the main text.

Table 11: **PhysDNet-M INT8 quantization quality** on the fixed-function NPU. PSNR drop vs. FP32 is $\sim$1.5 dB within the training β range. $\pm$ values are 1σ over n samples.

β	n	INT8 PSNR (dB)	FP32 PSNR (dB)
0.005	29	35.30±4.9	37.95±4.9
0.010	29	36.38±3.9	37.19±5.1
0.020	29	34.08±5.8	35.58±5.9
0.040	29	31.42±6.1	32.34±6.5
In-distribution avg		34.30	35.77
0.060	28	22.78±7.9	25.91±8.4
0.080	28	20.07±6.1	20.02±6.4
0.100	28	14.31±5.2	16.18±6.0
Overall avg		27.89	29.31

Table 12: **Full embedded-GPU benchmark** (256×256, 50 warmup + 300 timed iterations, MAXN mode). Mean and median latency, FPS, and onboard rail power reported for both CUDA (FP32) and TensorRT FP16 providers.

Model	Params	Provider	Mean (ms)	Median (ms)	FPS	$\overline{P}$ (W)	$P_{\max}$ (W)
AOD-Net LiDAR	0.002 M	TRT_FP16	3.06	3.05	327.0	4.94	5.21
AOD-Net LiDAR	0.002 M	CUDA	12.36	12.35	80.9	5.13	5.66
FFA-Net LiDAR	0.77 M	TRT_FP16	21.31	21.30	46.9	8.21	10.09
FFA-Net LiDAR	0.77 M	CUDA	83.67	83.65	12.0	13.39	15.44
DehazeFormer-T LiDAR	1.36 M	TRT_FP16	21.25	21.24	47.1	12.38	12.93
DehazeFormer-T LiDAR	1.36 M	CUDA	62.61	62.60	16.0	11.47	13.17
DEA-Net LiDAR	3.65 M	TRT_FP16	35.12	35.12	28.5	10.17	12.42
DEA-Net LiDAR	3.65 M	CUDA	89.93	89.93	11.1	12.45	13.05
PhysDNet-S (ours)	0.72 M	TRT_FP16	9.24	9.23	108.3	13.45	13.48
PhysDNet-S (ours)	0.72 M	CUDA	40.72	40.70	24.6	10.63	13.54
PhysDNet-M (ours)	2.95 M	TRT_FP16	17.09	17.09	58.5	12.38	12.42
PhysDNet-M (ours)	2.95 M	CUDA	79.08	79.07	12.6	12.78	14.97
PhysDNet-L (ours)	6.59 M	TRT_FP16	25.80	25.80	38.8	11.80	13.01
PhysDNet-L (ours)	6.59 M	CUDA	120.52	120.54	8.3	13.33	16.12

Hardware disclosure. The NVIDIA Jetson Orin Nano Super (8 GB, MAXN) is named explicitly. The fixed-function NPU is characterised by its specifications (26 TOPS INT8, $<$4 W TDP, vendor dataflow compiler with default post-training quantization); all throughput and quality figures are as measured.

H Baseline Descriptions

We compare PhysDNet against both established dehazing architectures and architectural ablations to quantify the contribution of each design choice.

Published baselines. **AOD-Net** [3]: a lightweight all-in-one CNN that directly regresses the clean image, representing early physics-reformulated dehazing. **FFA-Net** [4]: a feature-fusion attention CNN; strong attention-based baseline. **DehazeFormer-T** [5]: a Transformer-based restoration model representing modern attention architectures. **DEA-Net** [6]: a difference-aware encoder-decoder network. We include it to benchmark our physics-guided feature extraction against, purely data-driven detail enhancement.

I Model Efficiency Breakdown

Table 13 provides a detailed efficiency comparison across all models. PhysDNet variants are measured at 256×256; baselines at 512×512 (their best-performing resolution per Table 1).

Table 13: **Full efficiency breakdown.** Latency and FPS on a single NVIDIA GPU, batch size 1. †Measured at 256×256. We report the PyTorch numbers (no additional optimization is conducted)

Model	Config	Params (M)	GMACs	GFLOPs	Lat. (ms)	FPS
AOD-Net	NIR	0.002	0.23	0.46	0.3	3 753
AOD-Net	LiDAR	0.002	0.24	0.47	0.3	3 550
DehazeFormer-T	NIR	1.36	10.92	21.83	7.0	143
DehazeFormer-T	LiDAR	1.36	10.97	21.95	7.0	143
FFA-Net	NIR	0.77	98.57	197.15	11.5	87
FFA-Net	LiDAR	0.77	98.73	197.45	11.5	87
DEA-Net	RGB	3.65	16.42	32.84	8.9	112.6
DEA-Net	LiDAR	3.65	16.44	32.88	8.9	112.6
PhysDNet-S†	ch32	0.72	2.24	4.47	1.5	661
PhysDNet-M†	ch64, SE	2.95	8.56	17.13	2.3	429
PhysDNet-L†	ch96, SE	6.59	18.98	37.97	3.6	275

Table 14: **Auxiliary loss ablation.** PhysDNet-M (ch64) trained for 200 epochs at 512×512. Each row enables one auxiliary loss on top of the base (reconstruction + perceptual + gradient) objective. FFT contributes the largest PSNR gain; Contrastive contributes the largest SSIM gain.

Auxiliary loss	PSNR (dB) ↑	SSIM ↑
Base objective only	32.38	0.9315
+ Contrastive	32.97	**0.9406**
+ FFT	**33.02**	0.9401

J Auxiliary Loss Ablation

We ablate the two auxiliary loss components, FFT and Contrastive, that supplement the base objective (reconstruction + perceptual + gradient). We trained PhysDNet-M (ch64) for 200 epochs at 512×512; each row in Table 14 enables one auxiliary loss on top of the base objective. The results motivate the inclusion of both in the complete recipe (table 9), which combines them under the extended 500-epoch schedule used in the main results.

K Downstream Detection Protocol

Fog generation. Synthetic fog is applied using the same Koschmieder model as in training (§3.1), with β=0.06 and A=0.8 to test at out-of-distribution severity.

Detector and classes. YOLOv8-M [30] (COCO-pretrained, conf≥0.25, IoU≥0.5). STF KITTI-format labels are mapped to COCO classes and grouped into three families: *vehicle* (car, bus, truck), *person*, and *cyclist*. AP per family uses 101-point interpolation (COCO style).

Tiled inference. Dehazing at 256×256 with overlapping tiles (stride 192), linearly blended in the overlap to avoid boundary artifacts.

L Cross-Domain Generalization: Additional Details

Training-free alternative: chrominance transfer. The restored output on O-HAZE is structurally strong but desaturated. We use the model's predicted transmission $\hat{t}$ for *transmission-guided chrominance transfer*: convert both restored and hazy images to YCbCr, keep the Y channel from the restored output, and transplant attenuated Cb/Cr from the hazy input:

$$\mathrm{Cb}_{\mathrm{out}} = 0.5 + \hat{t} \cdot (\mathrm{Cb}_{\mathrm{hazy}} - 0.5), \quad \mathrm{Cr}_{\mathrm{out}} = 0.5 + \hat{t} \cdot (\mathrm{Cr}_{\mathrm{hazy}} - 0.5).$$

This attenuates haze-coloured chrominance in heavily fogged regions (low $\hat{t}$) while preserving natural colour where the scene is clear (high $\hat{t}$). Because it copies chrominance from the hazy input, the haze's colour cast persists; we therefore learn colour instead (App. M).

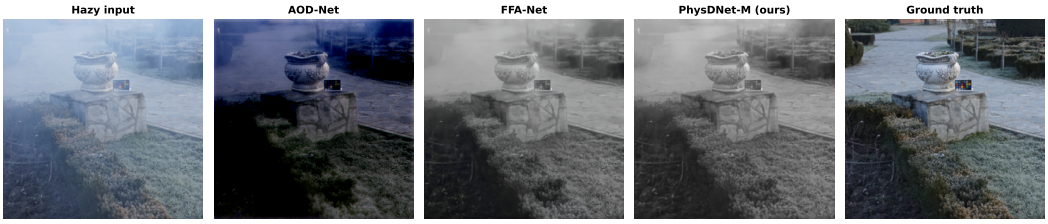

Figure 9: **Zero-shot O-HAZE restoration.** Left to right: hazy input, AOD-Net, FFA-Net, PhysDNet-M (ours), ground truth. All models trained on STF only.

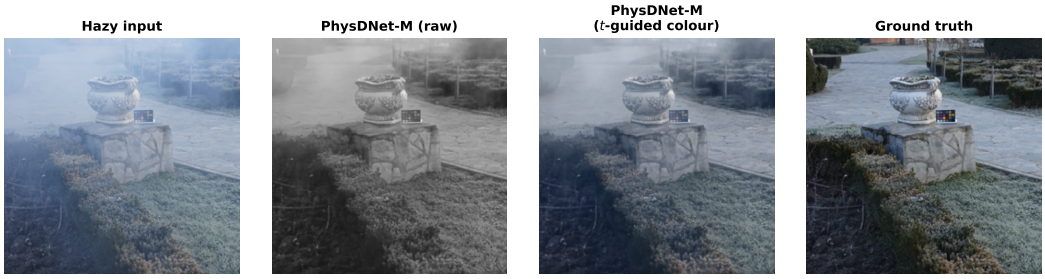

Figure 10: **Transmission-guided chrominance recovery.** Left to right: hazy input, raw PhysDNet-M output, after $\hat{t}$-guided chrominance transfer, ground truth.

M Self-Supervised Colour Adaptation and RTTS

Objective. On unpaired URHI images (4 803, 5 epochs, encoders at reduced learning rate), PhysDNet-M is trained with: (i) recomposition $\mathrm{Ch}(\hat{J}_{\text{dir}}\hat{t} + \hat{A}(1-\hat{t}), I)$; (ii) dual-head consistency $\mathrm{Ch}(\hat{J}_{\text{dir}}, \mathrm{sg}[\hat{J}_{\text{phys}}])$, where the stop-gradient makes the coloured inversion a teacher; (iii) a dark-channel prior on $\hat{J}_{\text{dir}}$; (iv) edge-aware $\hat{t}$ smoothness; (v) supervised synthetic STF replay batches. No paired data, labels, or architecture changes are used.

N Cross-Dataset Generalization: NH-HAZE

NH-HAZE [50] contains 55 *non-homogeneous* real outdoor hazy/clean image pairs (NTIRE 2020). Unlike O-HAZE's uniform fog, NH-HAZE exhibits spatially varying haze density, directly testing whether the physics head transfers despite the Koschmieder model's homogeneous-scattering assumption. As with other cross-dataset benchmarks, we feed $\mathbf{d}{=}\mathbf{0}$, $\mathbf{m}{=}\mathbf{0}$ and resize to $256{\times}256$. No fine-tuning.

PhysDNet-M achieves the highest PSNR-Y (14.74 dB), outperforming DEA-Net by $+0.15$ dB and FFA-Net by $+0.20$ dB; PhysDNet-L attains the best SSIM-Y (0.5159). PhysDNet-S remains competitive (14.70 dB) at $4\times$ fewer parameters than the M variant. Removing the physics head (NoPhy) reduces PSNR-Y by 0.54 dB and drops the model below all modern baselines, confirming that the physics-head contribution transfers to non-homogeneous fog despite the underlying Koschmieder model assuming homogeneous scattering. This evidence directly addresses the model-mismatch concern raised in our Limitations.

O LiDAR Sparsity Robustness

Automotive LiDAR produces inherently sparse depth maps—typically $<5\%$ of pixels carry valid returns, with sparsity varying significantly across weather conditions, surface reflectivity, and sensor

Table 15: **Zero-shot NH-HAZE results at** 256×256**.** All models trained on STF; no NH-HAZE fine-tuning. Best in **bold**.

Method	PSNR-Y ↑	SSIM-Y ↑	PSNR-RGB ↑	SSIM-RGB ↑
AOD-Net	12.21	0.3835	12.12	0.3791
FFA-Net	14.54	0.5037	14.16	0.4889
DehazeFormer-T	14.55	0.4889	14.18	0.4742
DEA-Net	14.59	0.5069	14.24	0.4932
PhysDNet-M (NoPhy)	14.20	0.5294	13.98	0.5165
PhysDNet-S (ours)	14.70	0.4950	14.37	0.4807
PhysDNet-M (ours)	**14.74**	0.5111	**14.37**	0.4965
PhysDNet-L (ours)	14.45	**0.5159**	14.09	**0.5012**

Table 16: **LiDAR sparsity ablation** on STF val (256×256). PSNR (dB) at each keep ratio. At $r=0$ (no LiDAR) the model falls back to NIR-only restoration via the learned gated-fusion mechanism. Training uses $r\approx3\%$ (typical Velodyne HDL-64 density).

Model	**100%**	**75%**	**50%**	**25%**	**10%**	**5%**	**1%**	**0%**
PhysDNet-S	32.80	31.99	30.98	29.77	28.60	28.03	27.29	26.52
PhysDNet-M	35.51	34.44	33.27	32.17	31.16	30.66	30.23	29.70
PhysDNet-L	36.54	34.64	32.50	31.15	30.73	30.55	30.69	30.57
PhysDNet-M (NoPhy)	30.02	29.38	28.73	28.06	27.41	27.00	26.06	25.60

occlusion. To study how PhysDNet's restoration quality degrades as depth information becomes scarcer, we systematically subsample valid LiDAR returns from 100% down to 0% (pure NIR fallback) while keeping all other inputs fixed. This stress-tests the gated-fusion mechanism (§2.1): a well-designed gate should learn to fall back to NIR features when depth is absent ($\mathbf{m}=\mathbf{0}$), rather than catastrophically failing.

Protocol. For each keep ratio $r \in \{1.0, 0.75, 0.50, 0.25, 0.10, 0.05, 0.01, 0.0\}$, we randomly retain r fraction of valid mask pixels in the STF validation set (1 299 images) and zero-out the corresponding depth entries as shown in table 16 and fig. 11. The model receives a hazy image $I \in [0,1]^{3\times H\times W}$, sparse LiDAR depth D_s, and validity mask M, and jointly predicts a restored image, transmission map, and airlight vector.

Three observations. **(i)** All PhysDNet variants degrade smoothly rather than collapsing as LiDAR density decreases, confirming the gated-fusion mechanism (§2.1) successfully learned to attribute weight to NIR features when depth signal is unavailable. **(ii)** Even at 0% LiDAR (pure NIR fallback), PhysDNet-M retains 29.70 dB—still 4.10 dB above the NoPhy ablation at matched 0% density (25.60 dB). This confirms that the physics head's contribution is not merely depth-driven but reshapes the encoder's NIR processing as well. **(iii)** At extreme sparsity ($r=1\%$, well below training density), PhysDNet-L's PSNR (30.69) remains higher than at $r=0\%$ (30.57), indicating that even one valid LiDAR point per 100 pixels provides usable signal. This degradation profile validates deployment under realistic LiDAR conditions where weather can drop point density significantly below nominal.

P No-Reference Perceptual Quality

Models trained on STF synthetic fog produce near-achromatic restorations on cross-dataset images (§3.6), so chrominance-sensitive full-reference metrics underestimate restoration quality. We supplement Table 4 with two established no-reference IQA metrics: **NIQE** [51] and **BRISQUE** [52] (lower = better; both assess naturalness and distortion blindly without ground truth). We evaluate on O-HAZE (45 outdoor pairs, no LiDAR; $\mathbf{d}=\mathbf{0}$).

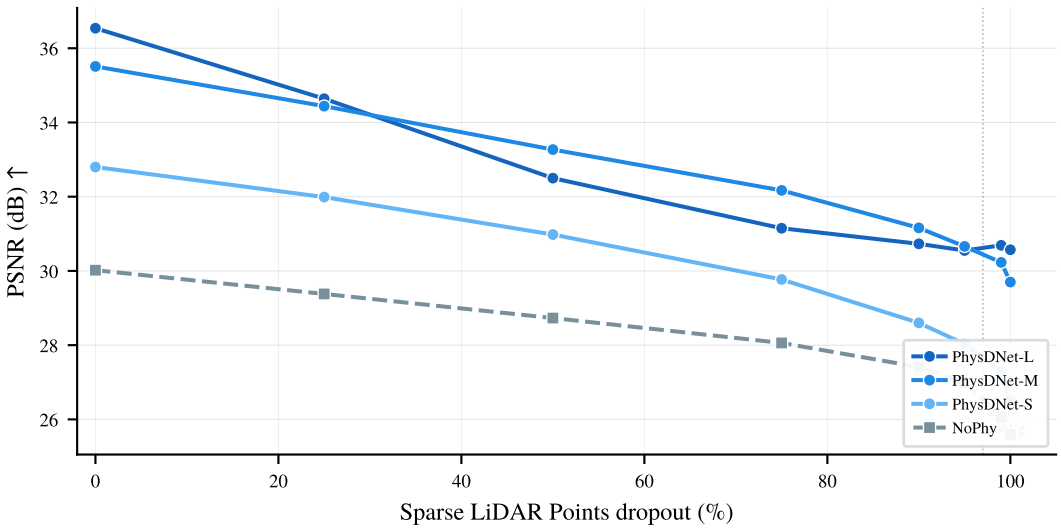

Figure 11: **PSNR vs. LiDAR keep ratio.** All PhysDNet variants degrade gracefully as LiDAR points are removed. Even at 0% density (NIR only), PhysDNet-M retains a $+4.10$ dB advantage over the NoPhy ablation, confirming that the physics-guided representation transfers across input modalities. Vertical dotted line marks the typical training-time density ($\sim$3%).

Table 17: **No-reference perceptual quality on O-HAZE** (NIQE / BRISQUE, lower = better). Numbers in parentheses are 1σ across the 45-image O-HAZE validation set. "Hazy" is the unprocessed input. Best in **bold**.

Method	**NIQE**↓		**BRISQUE**↓	
Hazy input	4.33	(σ=1.14)	7.75	(σ=7.02)
AOD-Net	6.37	(σ=4.18)	20.62	(σ=15.58)
FFA-Net	4.30	(σ=1.41)	12.25	(σ=8.84)
DehazeFormer-T	4.32	(σ=1.25)	12.76	(σ=9.83)
DEA-Net	4.64	(σ=1.62)	11.95	(σ=10.66)
PhysDNet-S	4.21	(σ=1.04)	15.98	(σ=6.92)
PhysDNet-M	**3.92**	(σ=1.08)	11.13	(σ=5.58)
PhysDNet-L	4.06	(σ=1.10)	**9.49**	(σ=6.73)
PhysDNet-M (NoPhy)	3.91	(σ=1.27)	11.96	(σ=6.29)

Discussion. PhysDNet leads or matches the strongest baselines on both no-reference metrics. PhysDNet-M achieves the best NIQE among learned methods (3.92, vs. DEA-Net 4.64 and AOD-Net 6.37); PhysDNet-L achieves the best BRISQUE (9.49, vs. FFA-Net 12.25 and DEA-Net 11.95). Compared to the NoPhy ablation, PhysDNet-M shows a consistent BRISQUE improvement (-0.83), while NIQE differences fall within noise—NIQE is a weaker discriminator than BRISQUE for low-distortion images, where most methods already approach naturalness statistics. Together with the PSNR-Y and SSIM-Y results in Table 4, these no-reference metrics provide independent confirmation that PhysDNet's restoration quality on real haze is not an artifact of synthetic-data circularity.

Q Adverse-Weather Consistency and Failure Modes

Table 18: Dual-head agreement (PSNR / Pearson r) on real STF captures, 250 frames per official split, full resolution, released PhysDNet-M checkpoint.

Condition	Day	Night
Clear	39.15 / 0.998	33.39 / 0.993
Light fog	40.13 / 0.999	34.73 / 0.994
Dense fog	**40.31** / 0.999	34.84 / 0.995
Snow	39.45 / 0.997	35.38 / 0.994
Rain	35.05 / 0.995	—

Agreement is a consistency measure, not accuracy against clean references.

Failure modes and restrictions. **Night:** a uniform $\approx$5 dB consistency reduction ($r \geq 0.993$); active light sources locally violate the global-airlight assumption, and the Direct Head absorbs this. **Over-enhancement:** $\hat{t} \geq t_{\min} = 0.05$ caps the inversion gain at $\approx 20\times$ by design. **Colour:** after adaptation, colour inherits $\hat{A}$, so a mild warm shift can appear under tinted illumination. **Adaptation trade-off:** -5.1 dB on O-HAZE. **Low-transmission noise:** in the Physics Head this is the σ^2/t^2 structure of Prop. 1 (App. F), not inherited by the deployed Direct Head.

R RGB-Only Training and Other Inverse Problems

RESIDE-6K. PhysDNet-M and a matched NoPhy twin are trained from scratch (RGB only, $D_s = M = 0$, $\lambda_t = 0$, same seed and recipe, 600 epochs). On the full 1 000-image SOTS-mix test set they reach 28.45 vs. 28.59 dB, and 21.9 vs. 22.2 dB on the densest-pixel bin. Depth-branch folding is bit-exact (max $|\Delta| = 0$) at 6.33 vs. 8.56 GMACs. Jetson TensorRT FP16 latency is unchanged (13.3 ms, bandwidth-bound), so the saving accrues on MAC-limited NPUs. Zero-shot on O-HAZE, the STF-trained model exceeds this colour-native RESIDE model by $+3.1$ dB PSNR-RGB (19.65 vs. 16.58).

Underwater restoration (anchored). From NYU-v2 RGB-D [53] we synthesise $I_c = J_c e^{-\beta_c d} + B_c(1 - e^{-\beta_c d})$ [54] with $\beta_r \gg \beta_g > \beta_b$ and signal-dependent noise. A 3% sparse-depth prior anchors a 3-channel transmission head, mirroring LiDAR on STF. Comparison axes were fixed before the runs.

Table 19: Underwater (NYU-v2 synthesis, 1 000 epochs, seed 42). PSNR (dB) by ground-truth transmission bin.

	Overall	Dense ($t<0.2$)	Mid	Light	Red ch.
Physics	**28.60**	**24.04**	**31.58**	37.30	**24.28**
NoPhy	28.22	23.83	30.76	**37.37**	23.68

Low-light enhancement. The Retinex inversion $\hat{R} = I/(\hat{L} + \epsilon)$ is the $A \to 0$ case of Eq. (5). On LOL-v2-Real [23] (100 test images), the physics variant leads in both run pairings (20.04/19.64 vs. 19.46/19.15 dB; SSIM 0.906 vs. 0.897). On LOL-v1 [21], the 15-image test set is noise-dominated and the sign is unstable, so the result is inconclusive.

Summary. Amplification helps where the inverted parameter is anchored by a measurement and degradation is dense, and is neutral otherwise. Magnitudes outside STF (0.2–0.8 dB) come from a single seed.

S Dataset Details

The Seeing Through Fog (STF) benchmark [1] provides 10 280 synchronised multi-sensor frames captured over 14 months of real adverse weather across Northern and Central Europe, spanning



Figure 12: **STF sensor overlay per weather condition.** Sparse Velodyne HDL-64 LiDAR returns (centre, right) highlight the multi-modal perception challenge addressed by PhysDNet. Imagery from the monochrome 12-bit gated NIR camera.

a temperature range of $-25.5°$C to $19.7°$C. Each frame includes monochrome 12-bit gated NIR imagery, sparse Velodyne HDL-64 LiDAR returns, RGB stereo, and weather metadata. We use only the gated NIR and LiDAR modalities, as they are the most informative under fog (§B). Table 20 summarises the per-condition split and how each subset is used in our experiments. Figure 12 illustrates the multi-modal perception challenge: LiDAR returns are sparse and non-uniform across weather conditions, motivating our gated-fusion design (§2.1).

T Additional Qualitative Results

We provide additional dual-head visualisations to illustrate the complementarity discussed in §2.2. Figure 13 shows side-by-side outputs of the direct head and the physics head across eight diverse synthetic-fog samples from the STF validation set, alongside the predicted transmission map and sparse LiDAR overlay. The physics head produces sharper outputs in well-conditioned regions

Table 20: **STF dataset split statistics.** 10 280 frames with weather metadata, Feb 2018–May 2019, -25.5°C to 19.7°C.

Condition	Frames	Ratio	Usage
Clear	1 497	14.6%	Synthetic haze training
Overcast	2 715	26.4%	Synthetic haze training
Light Fog	3 788	36.8%	Real-fog eval (Table 5)
Dense Fog	1 219	11.9%	Real-fog eval (Table 5)
Snow/Ice	1 061	10.3%	Consistency eval (App. Q)
Total	**10 280**	**100%**	

but exhibits noise amplification under dense haze (blue boxes in fig. 13), where the $(\hat{t} + \epsilon)^{-2}$ inversion is most ill-conditioned. The unconstrained direct head, trained jointly under the physics-anchored objective (§2.4), successfully suppresses these artifacts while inheriting the density-adaptive representation shaped during training. Only the direct head is deployed at inference.

U Out-of-Distribution Fog Robustness

To evaluate model robustness to out-of-distribution weather severities, we sweep the scattering coefficient $\beta \in [0.005, 0.10]$ at fixed global airlight ($A = 0.8$) across the full 1,299-image validation set. As shown in fig. 14, values $\beta \geq 0.05$ fall entirely outside the training distribution.

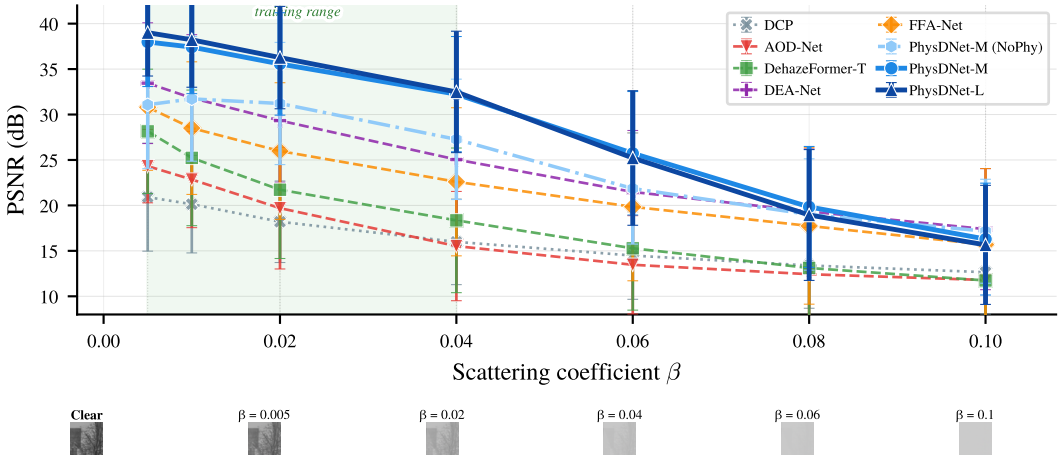

Figure 14: **PSNR vs. fog density β.** All models evaluated at 256×256 with A=0.8. The green band denotes the training range. PhysDNet-M maintains the highest PSNR across all densities, retaining a >6 dB advantage over the best baseline at β=0.06 and degrading gracefully outside the training distribution.

PhysDNet-M degrades smoothly beyond the training boundary, whereas baselines like FFA-Net suffer steeper drop-offs. We attribute this to the physics-constrained encoder: because the inversion formula is domain-invariant, the learned representations generalize significantly better to unseen β values.

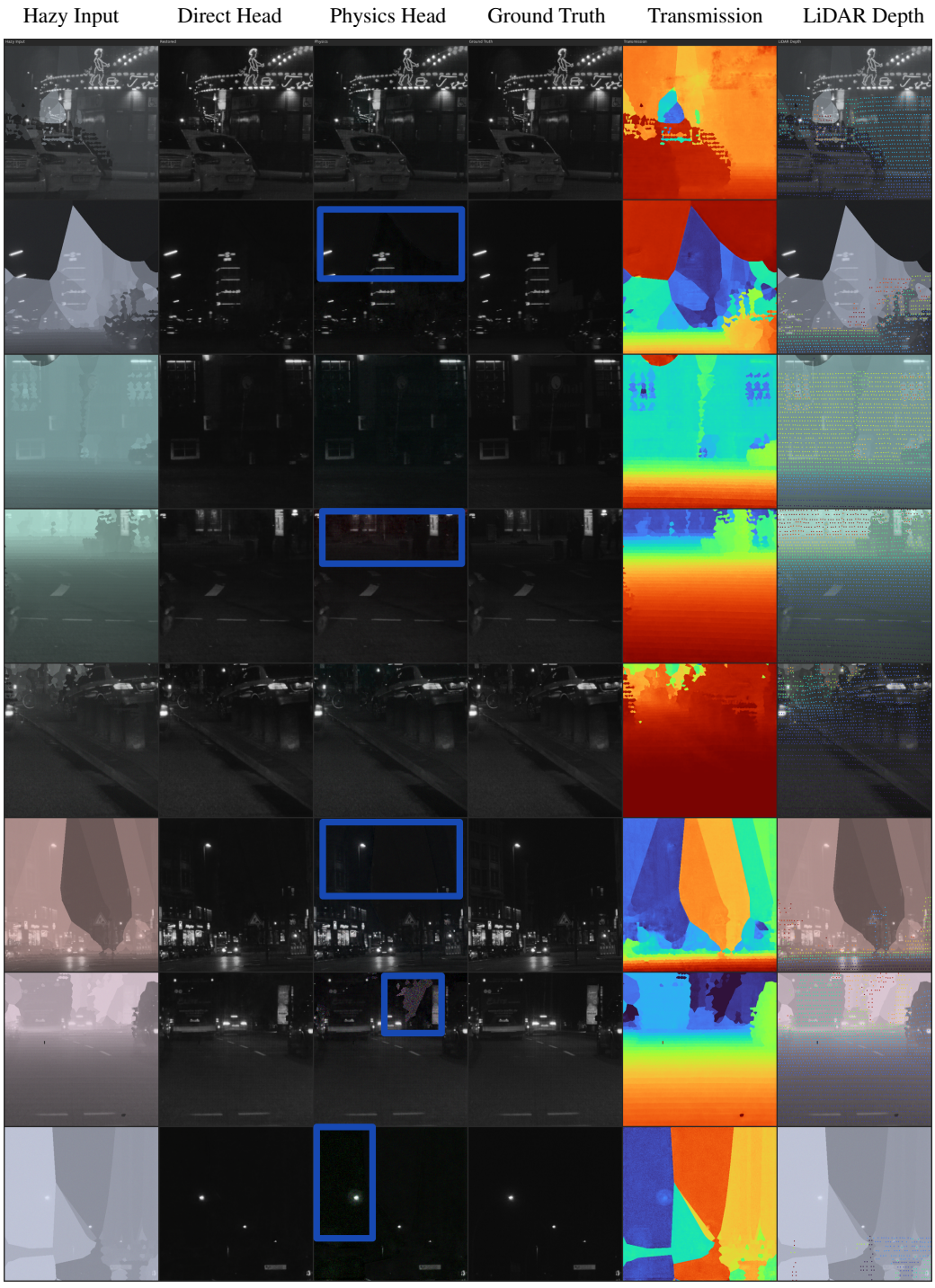

Figure 13: **Dual-head outputs on synthetic STF fog.** Hazy inputs are rendered with Eq. (1) from nearest-neighbour-filled sparse LiDAR depth, so haze density follows the interpolated depth and can look unnatural at depth discontinuities and in regions without LiDAR returns (e.g. sky). Blue boxes mark dense-haze regions where the fixed, parameter-free Physics Head retains these projection artifacts and the σ^2/t^2 noise predicted by Prop. 1; the deployed Direct Head suppresses them, restoring beyond the Koschmieder model.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction state four claims: (i) a density-adaptive gradient amplification mechanism arising from differentiation through a Koschmieder inversion (section 2.3, section F); (ii) state-of-the-art accuracy on Seeing Through Fog at low compute (table 1, fig. 2); (iii) zero-shot generalization to non-homogeneous fog (section 3.6, section N); and (iv) deployability on fixed-function NPUs (table 6, fig. 6). All are quantitatively supported in the corresponding sections and appendices.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: A dedicated Limitations paragraph at the end of the Conclusion discusses (i) synthetic Koschmieder training distribution, (ii) STF-only multi-modal training, (iii) near-achromatic outputs from monochrome NIR supervision, and (iv) Koschmieder-specific theoretical scope. section F also includes an explicit Scope paragraph stating the assumptions of the theoretical analysis (homogeneous single-scattering, Gaussian i.i.d. noise, first-order accuracy of $(\hat{t}, \hat{A})$).

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

   Justification: Theoretical results in section 2.3 (eqs. (6) and (7)) are derived in full in section F, including Proposition 1 (transmission-scaled noise variance with proof), Proposition 2 (Jacobian structure with derivation), the loss-gradient corollary, the airlight-gradient analysis, and the bounded-gradient analysis. Assumptions are stated explicitly in the Scope paragraph.

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: Training recipe (table 8), loss weights and activation schedule (table 9), model variants (table 10), architecture details (section D), dataset splits and statistics (section S, table 20), evaluation protocols for each experiment (sections G, K, O, P and U), and baseline configurations (section H) are documented. Code available under `https://github.com/Heatdh/PhysDnet`.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes]

   Justification: Code is available under `https://github.com/Heatdh/PhysDnet`, as stated in the abstract. STF [1], O-HAZE [31], I-HAZE [32], and NH-HAZE [50] are publicly available from their original authors. Hyperparameters, training recipes, and architectural details are fully documented in sections D and E to enable independent reimplementation.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Optimizer (AdamW), learning rate (2×10^{-4}), schedule (cosine annealing with 10-epoch warmup), batch size, weight decay, gradient clipping, crop size, and mixed-precision use are all specified in table 8. Loss weights and the contrastive activation schedule (epoch 185) are in table 9. Haze synthesis parameters ($\beta \sim \mathcal{U}(0.005, 0.04)$, $A_k \sim \mathcal{U}(0.7, 1.0)$) are specified in the same table. Train/val splits are documented in section S.

Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: PhysDNet-M reports mean±SE over 3 seeds in the main comparison (table 1: 35.55 ± 0.07 dB PSNR, 0.9566 ± 0.0009 SSIM). INT8 NPU quantization quality reports $\pm 1\sigma$ across 28-29 samples per fog density β (table 11). No-reference perceptual metrics on O-HAZE include 1σ across the 45-image validation set (table 17). The fog-density sweep (fig. 14) report per-condition averages over the full 1,299-frame validation set. Cross-domain results (tables 4 and 15) and downstream detection (table 3) report point estimates due to compute constraints; the magnitude of the effects ($+1.67$ dB on STF, -5.51 dB NoPhy ablation, $+0.62$ dB cross-domain) substantially exceeds typical seed variance.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Training compute (500 epochs per variant, single GPU, AMP (FP16) enabled for the L variant) is documented in section E. Edge deployment compute is reported in detail: fixed-function NPU and Jetson Orin Nano latency, FPS, and onboard power in table 6 and table 12; desktop GPU latency in table 13. Power profiles and warmup/timing protocols are specified in the captions of tables 6 and 12.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: Our work uses publicly available datasets collected by their original authors with appropriate research-use licensing. No human-subjects research was conducted.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: section A discusses positive impacts (improved adverse-weather perception for autonomous driving and robotics, reduced weather-related accidents, Green AI benefits from sub-4 W deployment) and negative impacts (dual-use potential for unconsented surveillance or military deployment, safety risks under out-of-distribution weather violating the Koschmieder assumption). Mitigation recommendations (redundant sensor fusion, uncertainty estimation, hardware fail-safes) are included.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

Answer: [N/A]

Justification: This work releases a dehazing model trained on publicly available adverse-weather datasets. It does not pose elevated misuse risks comparable to large-scale generative models or scraped datasets. Dual-use considerations and recommended deployment safeguards are discussed in section A.

Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [Yes]

Justification:All datasets (STF [1], O-HAZE [31], I-HAZE [32], NH-HAZE [50]) are cited and used under their original research-use licenses. Baselines (AOD-Net [3], FFA-Net [4], DehazeFormer-T [5], DEA-Net [6]) and YOLOv8-M [30] are cited and used under their respective open-source licenses. Our code is hosted under `https://github.com/Heatdh/PhysDnet`.

Guidelines:

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

13. **New assets**

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

Answer: [Yes]

Justification: The PhysDNet model and code release will include model architecture, training pipeline, evaluation scripts, configuration files, and pre-trained checkpoints with accompanying documentation. Hyperparameters, architectural specifications, and training schedules are fully documented in the paper (sections D and E).

Guidelines:

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

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

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

    Answer: [N/A]

    Justification: No crowdsourcing or human-subjects research was conducted. All evaluations use publicly available datasets.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: No human-subjects research was conducted.

    Guidelines:

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

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: No large language models are used as a core component of the proposed method.

Guidelines:

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