ShadowGDNet: Gated and Decoupled Fusion of Unreliable Scene Priors for Mask-Free Shadow Removal

Ujjwal Mishra[1] Kailash A. Hambarde[2] Hugo Proença[2]
Sachin Chaudhary[3] Praful Hambarde[4] Amit Shukla[4]

[1]Indian Institute of Information Technology Una, Himachal Pradesh, India
[2]University of Beira Interior, Instituto de Telecomunicações, Covilhã, Portugal
[3]UPES, India [4]Indian Institute of Technology Mandi, Himachal Pradesh, India

ujjwalmishra238@gmail.com, kailash.hambarde@ubi.pt, hugomcp@ubi.pt,
sachin.chaudhary@ddn.upes.ac.in, praful@iitmandi.ac.in, amitshukla@iitmandi.ac.in

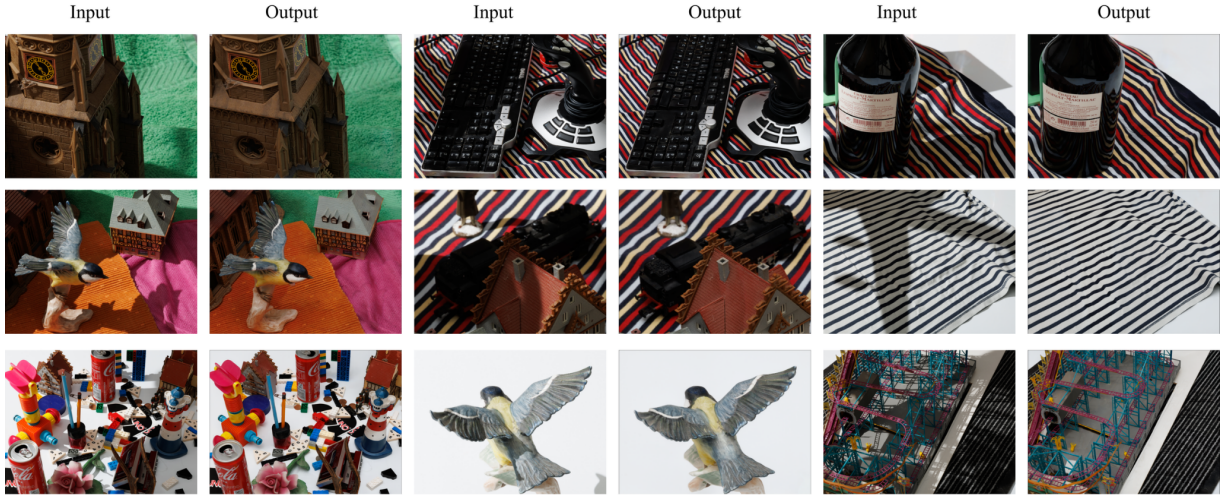


Figure 1. ShadowGDNet results on WSRD+ [35] validation images. Each pair shows the input (left) and our output (right).

Abstract

Mask-free shadow removal methods often take depth, surface normals and semantic features from frozen foundation models as extra inputs. These priors are computed from the shadowed image, so the shadow, a local loss of image quality, affects them as well. In a full-reference consistency test on ISTD+, the depth estimated from a shadowed image differs from the depth of its shadow-free pair 6.9 times more inside the shadow than outside it. Current architectures nevertheless apply the priors uniformly across the image. We propose ShadowGDNet, which makes two changes to the DenseSR backbone. Gated Skip Fusion (GSF) learns a per-pixel, per-channel gate on each skip connection from the reconstruction loss alone. Decoupled Multi-Modal Attention (DMMA) assigns each prior to its own group of attention heads, so within an attention layer an error in one prior no longer affects every head. The two modules add only 0.04M parameters. ShadowGDNet improves on DenseSR on ISTD, ISTD+ and WSRD+ by up to 0.77 dB PSNR, and a swap of the depth model shows that the gain comes mainly from the architecture. The advantage grows when the priors are worse: with Gaussian noise on the geometric priors at test time ($\sigma = 0.2$), our model loses 0.42 dB PSNR, against 1.21 dB for DenseSR. Supplementary material: github.com/Ujjwal238/ShadowGDNet.

1. Introduction

Shadows reduce the accuracy of detection and tracking [32], semantic segmentation [20] and 3D reconstruction [41], so shadow removal is often used as a preprocessing step. Many

learned methods [11, 12, 53] need a shadow mask as input [14, 37], which is hard to obtain when indirect light softens the shadow edge. Mask-free methods [8, 46] work without a mask, but then a dark region can be either a shadow or a dark surface. To resolve this, recent methods add priors from pretrained models [33, 46, 48]. Depth and surface normals describe the scene geometry, and self-supervised features such as DINOv3 indicate which pixels belong to the same material.

Degradations such as noise, blur and low contrast are known to reduce the accuracy of pretrained vision models [7, 15, 18]. A shadow is a local degradation of this kind: the image is darker there and has less contrast and texture, which is what depth estimators and feature extractors rely on. Consequently, the priors intended to guide shadow removal can themselves be degraded by the shadow. We test this with a full-reference consistency test on ISTD+ (Sec. 3), using the shadow-free image as the reference. Inside the shadow, the depth estimated from the shadowed image differs from the reference 6.9 times more than outside it, and normals and semantic features show the same trend. Both images show the same scene, so a large difference means that at least one of the two predictions is wrong. The encoder features of the restoration network are computed from the same shadowed image, which motivates gating them as well.

Current prior-guided models do not account for this. DenseSR [23], a recent mask-free model that we use as our backbone, is a typical example. Its skip connections send encoder features to the decoder at every scale, including features from inside the shadow. Its attention blocks multiply the attention logits by a semantic map and a geometric map, so a near-zero value in one map cancels the other, in every head.

We modify both parts with two modules. Gated Skip Fusion (GSF) computes a per-pixel, per-channel gate from the encoder and decoder features and scales the encoder features before they are merged. Decoupled Multi-Modal Attention (DMMA) splits the attention heads into an unmodulated, a semantic and a geometric group, which confines an erroneous prior to one group within each attention layer. Both are trained with the reconstruction loss alone, reduce to the original design with the gate fixed to one or a single head group, and add only 0.04M parameters. We use DINOv3 [33] for semantic features and DepthPro [3] for depth, and we measure separately how much of the gain comes from the depth model.

ShadowGDNet improves PSNR over DenseSR on all three benchmarks we test, ISTD [37], ISTD+ [19] and WSRD+ [35], by 0.77, 0.45 and 0.14 dB, with the best mask-free PSNR on ISTD and WSRD+ (examples in Figure 1). Training both architectures with both depth models shows that the gain comes mainly from the architecture rather than the stronger depth model. The advantage increases as the priors degrade: with noise added to the geometric priors at test time, DenseSR loses about three times as much PSNR as our model, and the ratio is highest when the noise is inside the shadow.

Our contributions are:

- A full-reference consistency test showing that a shadow, as a local quality degradation, changes the depth, normal and semantic priors from frozen models much more inside the shadow than outside it (Sec. 3).
- Gated Skip Fusion and Decoupled Multi-Modal Attention, two modules that gate the skip features and decouple the priors inside attention, at a cost of 0.04M parameters (Sec. 4).
- Experiments on three benchmarks showing consistent gains over DenseSR, an architectural gain larger than the gain from a better depth model on every benchmark, and a much smaller accuracy loss when the priors are corrupted (Sec. 5).

2. Related Work

Early shadow removal methods relied on hand-crafted cues such as gradient consistency and illumination statistics [10, 13]. Paired datasets with shadow masks, ISTD [37] and SRD [29], enabled supervised mask-guided methods based on multi-context embedding [29], stacked conditional GANs [37], masked adaptive instance normalization [14], auto-exposure fusion [11] and regional relighting [25]. Mask-free methods remove the dependence on an accurate mask through patch matching [5], domain-classifier guidance [16], Fourier-domain restoration [8], or joint modeling of direct and indirect lighting [46]. Transformers are now common, both as general restoration backbones [9, 21, 26, 34, 40, 42, 43, 45, 51] and in shadow-specific designs [12, 44]. To separate shadows from dark surfaces, recent mask-free methods add scene priors: depth and normals describe the 3D structure that determines how shadows form [1, 48], and self-supervised features [33] help distinguish shadows from dark materials. OmniSR [46] combines both kinds of prior in a transformer, and DenseSR [23] injects multi-scale semantic, depth and normal information through a Scene-Integrated Module that multiplies the attention map by every prior map. In all of these methods, the skip connections pass encoder features to the decoder unchanged, and each prior is weighted uniformly across the image.

A related line of work studies how input quality affects pretrained vision models. Dodge and Karam [7] showed that image classifiers are sensitive to blur and noise, and ImageNet-C [15] and RoboDepth [18] benchmark classification and monocular depth estimation under common corruptions. These studies degrade the whole image and evaluate each model on its own task. We study a naturally oc-

curring local degradation and its effect on pretrained models whose outputs guide another network, and we measure it with the full-reference principle behind restoration metrics such as PSNR, SSIM [39] and LPIPS [52], taking the shadow-free image as the reference.

GSF is related to adaptive feature fusion. Feature Pyramid Networks [22] made multi-scale fusion standard, U-Net [31] skip connections are known to carry artifacts between levels, and several methods reweight or resample features instead of concatenating them [24, 38]. Gated convolution [49] learns a per-pixel validity mask for inpainting, where the damaged region is known. In restoration, selective kernel feature fusion [50] and gated fusion networks [30] weight several feature streams with learned gates. Closest to GSF are the attention gates of Attention U-Net [28], which weight encoder skip features with a decoder signal; GSF differs in predicting a separate gate for every channel and in applying it to a prior-guided shadow removal network, where the damaged region is unknown and must be inferred from the damaged features themselves.

3. How Does a Shadow Affect the Priors?

Our method rests on the assumption that the priors are worse inside shadows. We test it on ISTD+ [19], which provides a shadowed image I_s, a shadow-free image I_f and a shadow mask for each scene, captured from the same viewpoint. We apply the frozen models used by our network, DepthPro [3] for depth and normals and DINOv3 [33] for semantic features, to both images and compare their outputs inside and outside the mask, using the output from I_f as the reference. Since ISTD+ has no ground-truth geometry, this full-reference test measures how much a prior changes when a shadow is added, not how accurate it is. The true depth d^* is nevertheless the same in both images, so for scale-aligned predictions d_s from I_s and d_f from I_f the triangle inequality gives $|d_s - d_f| \le |d_s - d^*| + |d_f - d^*|$. The discrepancy is therefore a lower bound on the sum of the two prediction errors, and the same holds for the angle between normals. We use it as a proxy for reliability; the full protocol is in the supplementary material. Two confounds remain uncontrolled: shadows in ISTD+ often fall on weakly textured ground, where depth may be less reliable even without a shadow, and the ISTD+ shadow-free images are color-corrected. The test requires only paired shadowed and shadow-free images, so it applies to any frozen prior model.

All three priors change much more inside shadows (Table 1). The relative depth discrepancy, the median of $|d_s - d_f|/d_f$ over the region, rises from 0.021 to 0.144, the angle between normals from $4.8°$ to $19.6°$, and the cosine similarity of the semantic features drops from 0.972 to 0.883. Depth discrepancy also correlates with shadow intensity ($r = 0.46$) and is higher near soft boundaries than near hard ones (0.169 vs. 0.118). By this proxy, the priors are least reliable where the network needs them most.

Prior	Non-shadow	Shadow	Ratio
Depth: rel. discrepancy ↓	0.021	0.144	6.9×
Normals: angle (°) ↓	4.8	19.6	4.1×
Semantic: cosine sim. ↑	0.972	0.883	0.91×

Table 1. Change of each prior between the shadowed and the shadow-free image of the same ISTD+ [19] scene, inside and outside the shadow mask.

4. Method

4.1. Overview

ShadowGDNet uses the DenseSR [23] backbone, a U-Net [31] built from Swin [26] transformer blocks (Figure 2). Given a shadowed image $I_s \in \mathbb{R}^{H\times W\times 3}$, the network F predicts a residual, and the output is $\hat{I}_f = F(I_s) + I_s$. The priors are computed once by frozen models: depth is concatenated with the RGB input, point and normal maps modulate window attention, and DINOv3 features are concatenated at the bottleneck. Shadowed information thus reaches the decoder along two routes, the skip connections and the prior-modulated attention. GSF (Sec. 4.4) acts on the first and DMMA (Sec. 4.5) on the second; apart from the depth model, which we study in Sec. 5.3, everything else is identical to the baseline.

4.2. Scene Priors

DepthPro [3] predicts a metric depth map $\mathbf{D} \in \mathbb{R}^{H\times W}$. We divide it by its median over valid pixels, so that the geometric prior does not depend on the absolute scale of the scene, and back-project each pixel (x, y) with normalized depth $z = \mathbf{D}[y, x]$ using a pinhole model,

$$x_{3d} = \frac{(x - c_x)\cdot z}{f}, \quad y_{3d} = \frac{(y - c_y)\cdot z}{f}, \tag{1}$$

where $f = \frac{W}{2\tan(\mathrm{FOV}/2)}$ and (c_x, c_y) is the image center. This yields a point map $\mathbf{P} \in \mathbb{R}^{H\times W\times 3}$, from which we compute the surface normals $\mathbf{N} \in \mathbb{R}^{H\times W\times 3}$ as $\mathbf{n} \propto \partial_x\mathbf{P} \times \partial_y\mathbf{P}$, using finite differences and normalizing to unit length. The datasets provide no camera intrinsics, so to back-project every depth model in the same way we assume a fixed $60°$ field of view rather than DepthPro's own estimate. The geometry is approximate but identical for both architectures, and other values change the results very little (Sec. 5.3).

The depth z forms the fourth input channel. $\mathbf{P}$ and $\mathbf{N}$ are bilinearly resized to the resolution of each prior-modulated block, where they define the geometric map M_{geo} (Sec. 4.5),

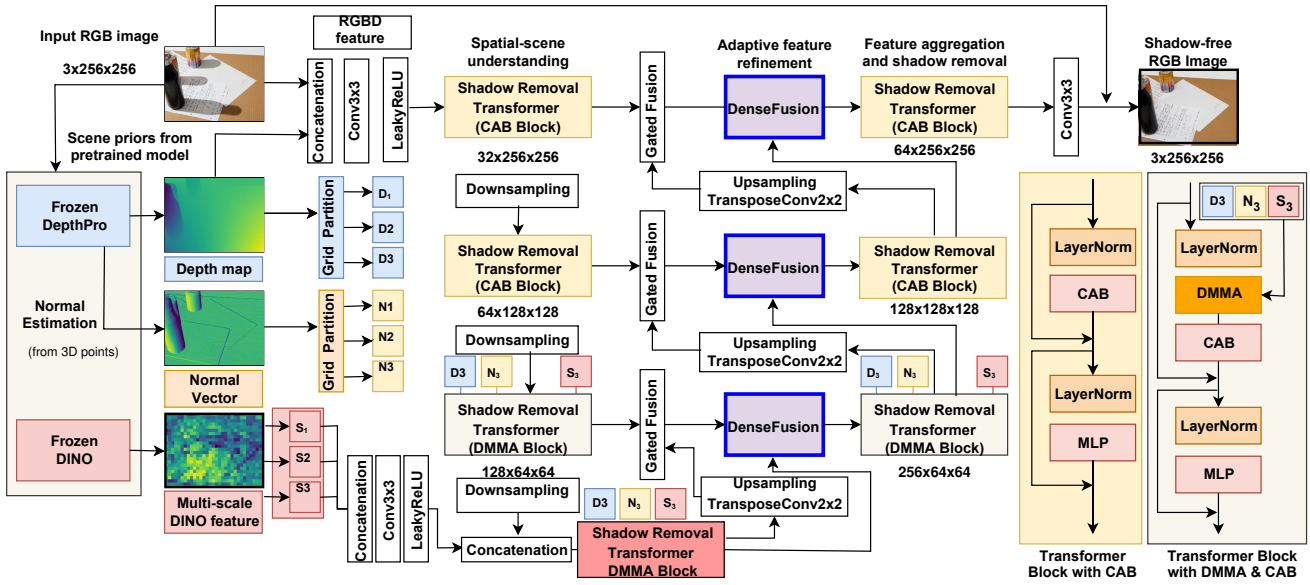


Figure 2. ShadowGDNet on the DenseSR [23] backbone. DepthPro [3] depth is concatenated with RGB as a fourth input channel. Point and normal maps from the same depth modulate attention in the deeper blocks, and DINOv3 [33] features are concatenated at the bottleneck. Shallow stages use channel attention and deeper stages use DMMA. GSF gates the three encoder skips before they are merged with the decoder.

and the DINOv3 [33] features are projected and concatenated with the encoder output at the bottleneck and also define the semantic map M_{sem}.

4.3. Encoder-Decoder Backbone

The first encoder stage works at full resolution with 32 channels. Each following stage halves the resolution with a 4×4 stride-2 convolution and doubles the channels, giving 64 channels at $H/2$ and 128 at $H/4$. A final downsampling yields 256 channels at $H/8$, which are concatenated with the projected DINOv3 features to form the 512-channel bottleneck. The decoder mirrors the encoder and upsamples with transposed convolutions. The stages at H and $H/2$ use channel attention blocks, and the stages at $H/4$ and $H/8$ use prior-modulated window attention, which is where we apply DMMA. At decoder stage l, the upsampled feature $\mathbf{U}^l \in \mathbb{R}^{B\times C\times H_l\times W_l}$ is merged with the encoder skip feature $\mathbf{E}^l$ of the same size. DenseSR concatenates them, $[\mathbf{U}^l; \mathbf{E}^l]$; we replace this step with GSF.

4.4. Gated Skip Fusion (GSF)

Skip connections provide the decoder with high-resolution detail, but in shadow removal they also propagate the shadowed appearance: since $\mathbf{E}^l$ is computed from I_s, at every scale the decoder receives uncorrected features next to its own partly corrected ones. GSF lets the network determine, per location and channel, how much of $\mathbf{E}^l$ to propagate. At each of the three skips (H, $H/2$ and $H/4$), it computes a gate from the encoder and decoder features (Figure 3),

$$\mathbf{G}^l = \sigma\left(\mathrm{BN}\left(\mathrm{Conv}_{1\times1}\left([\mathbf{E}^l;\ \mathbf{U}^l]\right)\right)\right), \tag{2}$$

where $[\cdot;\ \cdot]$ is channel concatenation, the 1×1 convolution maps $2C$ channels to C, BN is batch normalization and σ is the sigmoid, so that $\mathbf{G}^l \in [0,1]^{B\times C\times H_l\times W_l}$ has one value per pixel and channel. The gated encoder features are then concatenated with the decoder features,

$$\mathbf{F}^l = \left[\mathbf{E}^l \odot \mathbf{G}^l;\ \mathbf{U}^l\right], \tag{3}$$

where $\odot$ is element-wise multiplication.

The gate acts only on $\mathbf{E}^l$. $\mathbf{U}^l$ is the decoder's own estimate, already refined by the deeper prior-modulated blocks, whereas $\mathbf{E}^l$ carries the shadowed appearance without correction; gating $\mathbf{U}^l$ would let the network attenuate its primary reconstruction signal, while gating $\mathbf{E}^l$ only controls how much uncorrected detail is added to it. Because the gate is computed from both features, the decoder state determines how much of $\mathbf{E}^l$ is used, and because it is defined per channel, shadow-sensitive channels can be attenuated while the remaining channels at the same location are preserved. The gate is learned from the reconstruction loss alone, without a shadow mask or any signal about prior quality. On ISTD+, its mean value is about 0.3 inside the shadow mask and 0.7 outside at all three scales, and the gap is widest at the coarsest skip (0.26 vs. 0.66 at $H/4$; supplementary material). Since shadows are also darker, the gate may partly

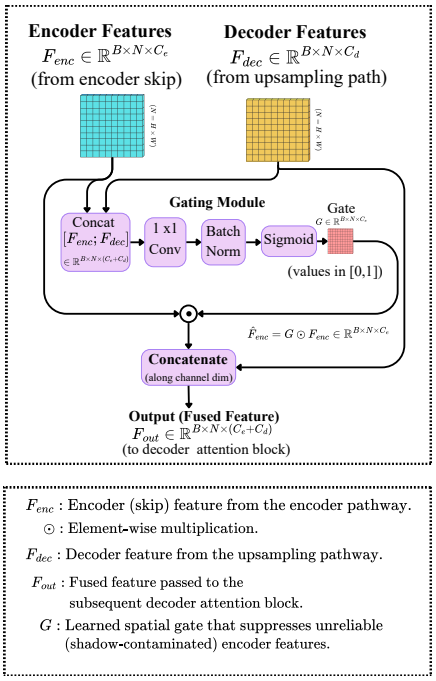


Figure 3. Gated Skip Fusion. The encoder and decoder features are concatenated and passed through a 1×1 convolution, batch normalization and a sigmoid to produce a per-pixel, per-channel gate. The gated encoder features are then concatenated with the decoder features.

track brightness; a causal test with a constant, shuffled or inverted gate at inference is left for future work. With $\mathbf{G}^l = 1$, GSF reduces to plain concatenation, so it can represent the baseline. The three gates add 0.04M parameters.

4.5. Decoupled Multi-Modal Attention (DMMA)

In DenseSR, the priors enter the attention through joint modulation. Consider a window of $N = \text{win_size}^2$ tokens with queries $\mathbf{Q}$, keys $\mathbf{K}$ and values $\mathbf{V}$, head dimension d_h and relative position bias B [26]. DenseSR's Scene-Integrated Module [23] scales the attention logits of every head by two prior maps before the softmax,

$$\mathbf{A} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_h}} \odot M_{\text{sem}} \odot M_{\text{geo}} + B\right), \quad (4)$$

where $\odot$ is element-wise multiplication. For tokens i and j with DINOv3 features $\mathbf{s}_i$, 3D points $\mathbf{p}_i$ and normals $\mathbf{n}_i$, the maps are

$$M^{ij}_{\text{sem}} = \max\left(\hat{\mathbf{s}}_i^\top \hat{\mathbf{s}}_j,\, 0\right), \quad M^{ij}_{\text{geo}} = \exp\left(-\tfrac{1}{2}(d_{ij} + d_{ji})\right), \quad (5)$$

with $\hat{\mathbf{s}}_i = \mathbf{s}_i / \|\mathbf{s}_i\|_2$ and $d_{ij} = |(\mathbf{p}_i - \mathbf{p}_j)^\top \mathbf{n}_i|$, the distance of token j from the tangent plane at token i. Both maps lie in $[0, 1]$: M_{sem} is high for tokens on the same object or material, and M_{geo} is high for co-planar tokens and decreases with their separation in depth.

Because the two maps enter Eq. (4) as a product, the modulated logit vanishes whenever either map does. If the depth estimate is erroneous inside a shadow and M_{geo} becomes small, the semantic information in M_{sem} is suppressed as well, in every head. In the limit $M_{\text{sem}} \odot M_{\text{geo}} \to 0$, the content term vanishes and $\mathbf{A} \to \text{softmax}(B)$, so the attention depends only on the relative positions of the tokens and no longer on their content.

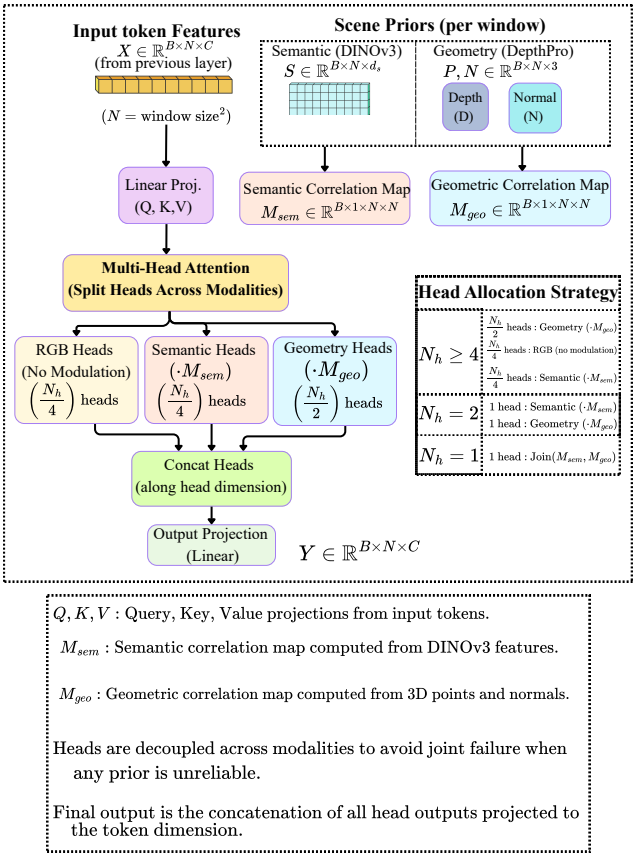


Figure 4. Decoupled Multi-Modal Attention. The attention heads are split into three groups: unmodulated heads, semantic heads modulated by M_{sem}, and geometric heads modulated by M_{geo}. The group outputs are concatenated and combined by the output projection.

DMMA instead partitions the N_h heads of a layer into an unmodulated, a semantic and a geometric group, $g \in \{\text{rgb}, \text{sem}, \text{geo}\}$, and applies at most one prior map to each group (Figure 4):

$$\mathbf{A}_g = \text{softmax}\left(\frac{\mathbf{Q}_g \mathbf{K}_g^\top}{\sqrt{d_h}} \odot M_g + B_g\right), \quad (6)$$

where $\mathbf{Q}_g$, $\mathbf{K}_g$ and B_g collect the queries, keys and position biases of the heads in group g, $M_{\text{rgb}} = \mathbf{1}$, and M_{sem} and M_{geo} are the maps of Eq. (5). For $N_h \geq 4$ the group sizes

are

$$N_{\text{rgb}} = N_{\text{sem}} = \lfloor N_h/4 \rfloor\,, \quad N_{\text{geo}} = N_h - 2\lfloor N_h/4 \rfloor\,. \quad (7)$$

Half of the heads are assigned to the geometric group, since depth and normals provide the most direct evidence of where a shadow boundary lies in 3D. With a single head, DMMA reduces to the joint modulation of Eq. (4). As in DenseSR, prior modulation is applied only in the stages at $H/4$ and $H/8$, and DMMA replaces the joint modulation in exactly these blocks, so the comparison with DenseSR changes only how the priors are combined. These layers have 4, 8 or 16 heads, so the split is always $(1/4, 1/4, 1/2)$; in single runs on the WSRD+ validation set it performed best among the splits compared in the supplementary material. DMMA adds no parameters, since it only changes the assignment of prior maps to heads.

The attended values of all heads are concatenated and mapped back to the token dimension by the shared output projection W_O. Since W_O is linear, the output of the attention block decomposes into one term per group,

$$\mathbf{Y} = \sum_g \mathbf{A}_g \mathbf{V}_g W_O^{(g)}, \quad (8)$$

where $W_O^{(g)}$ denotes the rows of W_O that act on group g. A corrupted map M_g therefore changes only the term of group g: if the depth estimate fails, the geometric heads lose their guidance, while the unmodulated heads still attend by content and the semantic heads by material similarity. Under joint modulation, the same error enters every head through Eq. (4). This separation holds within the attention blocks; depth also enters the network as the fourth input channel and DINOv3 features at the bottleneck. DMMA does not need to know which prior is erroneous; the partition alone prevents one prior from suppressing the guidance of the others. The per-stage configuration is given in the supplementary material.

4.6. Training

We train with the Charbonnier loss [4] between $\hat{I}_f$ and the ground truth I_f,

$$\mathcal{L} = \frac{1}{|\Omega|} \sum_{k \in \Omega} \sqrt{\left(I_f[k] - \hat{I}_f[k]\right)^2 + \epsilon^2}, \quad (9)$$

where Ω indexes all pixels and color channels and $\epsilon = 10^{-3}$. We use AdamW [27] with a learning rate of 2×10^{-4}, weight decay 0.02, 3 epochs of linear warm-up and cosine annealing, and train for 1800 epochs on one NVIDIA A40 GPU (48GB) with batch size 6 and 256×256 patches. The window size is 8 and the MLP ratio is 4. DINOv3 (ViT-L/16) [33] gives 1024-dimensional features and DepthPro [3] gives the depth from which the normals are computed (Sec. 4.2); both models are frozen. ShadowGDNet has 24.78M trainable parameters.

5. Experiments

5.1. Datasets and Protocol

ISTD [37] contains 1870 triplets of shadow image, shadow mask and shadow-free image from 135 scenes, split into 1330 for training and 540 for testing, and ISTD+ [19] uses the same images and split with color-corrected shadow-free images. WSRD+ [35] has a hidden challenge test set and a public validation set of 100 images, the only WSRD+ split with public ground truth. We train one model per benchmark, on the 1330 training pairs for ISTD and ISTD+ and on the 886 NTIRE 2026 training images for WSRD+, with no extra data or augmentation, and evaluate on the ISTD and ISTD+ test sets and the WSRD+ validation set. The network never receives a shadow mask; the ISTD and ISTD+ masks are used only for analysis. All benchmarks share the same hyperparameters, with no test-time augmentation or ensembling. We report the full-reference metrics PSNR and SSIM [39], computed with the official challenge code [36] for WSRD+.

The DMMA head split and the GSF scales were chosen on the WSRD+ validation set and reused unchanged on ISTD and ISTD+. We retrained DenseSR with the same data, hyperparameters, three seeds and checkpoint rule as ShadowGDNet, and its means coincide with the values reported in [23]. For both models we evaluate after every epoch and report the checkpoint with the best PSNR on the evaluation set. Since epoch-to-epoch PSNR fluctuations are comparable to the WSRD+ gain, this selection makes the WSRD+ comparison in particular optimistic; selecting on a held-out split would be cleaner.

5.2. Comparison with Existing Methods

Table 2 compares ShadowGDNet with mask-based methods [6, 12, 53], mask-free methods [8, 16, 46, 47] and its backbone DenseSR [23]. ShadowGDNet has higher PSNR than DenseSR on all three benchmarks and higher SSIM on ISTD+ and WSRD+. Among mask-free methods, it has the best PSNR on ISTD and WSRD+ and the best SSIM on ISTD+ and WSRD+; on ISTD+, the diffusion-based Stable-ShadowDiffusion has a higher PSNR. On WSRD+, it also outperforms the mask-based methods without using a mask. The only metric on which it trails DenseSR is SSIM on ISTD (0.974 vs. 0.976), a difference within the seed standard deviation ($\pm$0.002).

The gain is largest on ISTD (+0.77 dB), followed by ISTD+ (+0.45 dB) and WSRD+ (+0.14 dB). ISTD and ISTD+ share their scenes yet differ by 0.32 dB, so we do not attribute these differences to scene content; the benchmarks also differ in size, resolution and target quality. Within WSRD+, where these factors are fixed, the mean gain grows with shadow area from +0.07 dB in the smallest quartile to +0.23 dB in the largest (per-image gains are heavy-tailed, so

Method	Mask-free	ISTD [37]		ISTD+ [19]		WSRD+ [35]	
		PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
DHAN [6]	No	29.11	0.954	25.66	0.956	22.39	0.796
BMNet [53]	No	28.53	0.952	32.22	0.965	24.75	0.816
ShadowFormer [12]	No	29.90	0.960	31.39	0.946	25.44	0.820
DCShadowNet [16]	Yes	24.02	0.677	25.50	0.694	21.62	0.593
ShadowRefiner [8]	Yes	28.75	0.916	31.03	0.928	26.04	0.827
OmniSR [46]	Yes	30.45	0.964	33.34	0.970	26.07	0.835
StableShadowDiffusion [47]	Yes	—	—	**35.19**	0.970	26.26	0.827
DenseSR [23]	Yes	30.64	**0.976**	33.98	0.974	26.28	0.838
ShadowGDNet (Ours)	Yes	**31.41**	0.974	34.43	**0.985**	**26.42**	**0.865**

Table 2. Results on the ISTD [37] and ISTD+ [19] test sets and the WSRD+ [35] validation set. DenseSR is our retraining (three-seed mean); DCShadowNet and ShadowRefiner are as reported in [8], and the other baselines as reported in [23], which lists mask-based methods without ground-truth masks. HomoFormer [44] is omitted because no results under this protocol are available. Best mask-free result in each column in **bold**.

a few images may drive this trend), and it is larger for scenes with mostly soft boundaries than for scenes with mostly hard ones (+0.18 vs. +0.11 dB), consistent with Sec. 3. Figure 5 shows examples from ISTD+ and ISTD: DCShadowNet leaves residual shadow and boundary artifacts, ShadowRefiner shows color shifts under indirect light, and our results are more uniform across the former shadow region, with sharper boundaries.

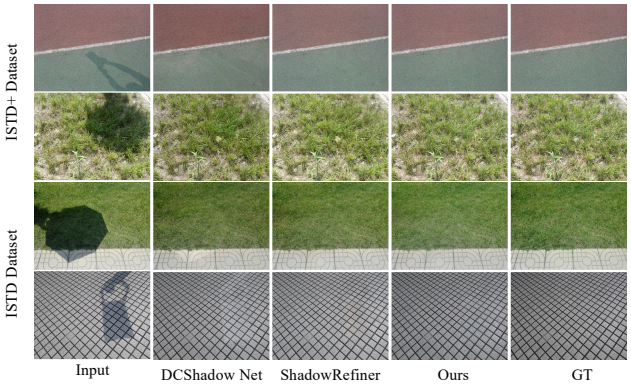


Figure 5. Qualitative comparison. Top two rows: ISTD+ [19]; bottom two rows: ISTD [37]. From left to right: input, DCShadowNet [16], ShadowRefiner [8], ours, ground truth.

5.3. Architecture or Depth Model?

ShadowGDNet differs from DenseSR both in its modules and in its depth model: DenseSR uses the metric-depth model of Depth Anything V2 (DAv2) [48], whereas we use DepthPro [3]; both depth maps are normalized and back-projected in the same way (Sec. 4.2). To separate the two factors, we train both architectures with both depth models. On WSRD+, with the depth model fixed, our architecture is 0.10 dB better with DAv2 (26.38 vs. 26.28) and 0.11 dB better with DepthPro (26.42 vs. 26.31), whereas switching DenseSR from DAv2 to DepthPro adds only 0.03 dB. Both use the same semantic features, so the depth model is the only input that changes. In single runs across the three datasets, with the depth model fixed, the architecture adds 0.10 to 0.72 dB, whereas switching the depth model changes PSNR by at most 0.07 dB, and varying the assumed field of view between 45° and 75° changes either model by at most 0.04 dB (supplementary material). Most of the gain should therefore remain with the computationally lighter DAv2.

5.4. Ablation Study

Configuration	ISTD	ISTD+	WSRD+
Baseline [23]	30.64±.05	33.98±.06	26.28±.03
+ GSF only	30.97±.04	34.17±.05	26.34±.02
+ DMMA only	31.08±.05	34.24±.06	26.36±.03
Ours	**31.41±.04**	**34.43±.05**	**26.42±.02**

Table 3. Component ablation: PSNR (dB), mean ± standard deviation over three seeds. The baseline is DenseSR retrained under our protocol. SSIM is in the supplementary material.

Table 3 adds each module to the baseline separately and then jointly, over three seeds. Each module improves PSNR on its own, with DMMA giving the larger gain on every benchmark, and the full model is best. Since the WSRD+ gain is small, we also test it per image with bootstrap confidence intervals (10,000 resamples) and a two-sided Wilcoxon signed-rank test. The mean gain is 0.14 dB (95% interval $[0.09, 0.19]$, $p = 3.1 \times 10^{-4}$), and PSNR improves on 71 of the 100 images, although the ten most improved images account for 41% of the total gain; on ISTD and ISTD+, both gains are significant with $p < 10^{-6}$. Alternative head splits and GSF scales, compared in single runs

Corruption	DenseSR	Ours	Ratio
Gaussian, σ=0.1	−0.54	−0.18	3.0
Gaussian, σ=0.2	−1.21	−0.42	2.9
Gaussian, σ=0.3	−1.95	−0.71	2.7
Correlated noise, 16 px	−1.44	−0.51	2.8
Boundary displacement, 5 px	−1.02	−0.37	2.8
Missing depth, 10% blocks	−1.63	−0.55	3.0
Depth blur	−0.74	−0.28	2.6
Global scale error, ±20%	−0.58	−0.31	1.9
Noise inside shadow only	−1.38	−0.44	3.1

Table 4. PSNR change (dB) on WSRD+ validation when the geometric priors are corrupted at test time, for the complete models of Table 2. Clean PSNR is 26.28 dB for DenseSR and 26.42 dB for ours. Ratio is the DenseSR loss divided by ours. The structured corruptions have the same mean depth error as σ=0.2.

in the supplementary material, differ by less than the seed variation.

5.5. Robustness to Corrupted Priors

Unlike corruption benchmarks [15, 18], which degrade a model's own input, we degrade the priors at test time, without retraining, and compare the two complete models of Table 2 (DenseSR with DAv2 depth, ShadowGDNet with DepthPro depth), so the test measures the robustness of ShadowGDNet as a whole. The corruptions are applied to the point map $\mathbf{P}$ and the normal map $\mathbf{N}$ after the median normalization of Sec. 4.2, so they are matched in normalized units, not in metric error. The depth input channel is the z-coordinate of $\mathbf{P}$, so the corruption reaches the network through the input as well as through every prior-modulated block.

With zero-mean Gaussian noise of standard deviation σ relative to the normalized depth, our model loses about a third as much PSNR as DenseSR at every noise level (Table 4): the ratio between the two losses stays between 2.7 and 3.0, and the gap between the models grows from 0.14 dB with clean priors to 1.38 dB at $\sigma = 0.3$. Because real depth errors are spatially structured, we also apply the structured corruptions of Table 4, each with the same mean depth error as $\sigma = 0.2$. The ratio remains close to three in all cases except the scale error (1.9), which after normalization multiplies every d_{ij} in M_{geo} by one factor and so changes the sharpness of the map but not the ranking of its scores. Noise restricted to the shadow yields the highest ratio, 3.1. Corrupting the semantic or the geometric prior alone gives a similar picture: DenseSR loses 0.98 and 1.21 dB, and our model 0.35 and 0.42 dB (supplementary material). The ratio is thus similar across corruption types and whether a quarter or half of the heads are affected, so these results do not show that the gain scales with the fraction of heads hit.

5.6. Computational Cost

The two modules add 0.04M parameters and 0.4G multiply-accumulates (MACs) to the 24.74M-parameter, 92.0G-MAC DenseSR backbone. The frozen models add 1.25B parameters and about 7.1T MACs per image, almost all of it in DepthPro, which runs its ViT-L encoders on a 1536^2 image pyramid. With the priors computed once and cached, the trainable network takes 18 ms and 1.4 GB per 256^2 image on one A40 GPU. In the NTIRE 2026 Shadow Removal Challenge, our single-model entry, without ensembling or test-time augmentation, placed seventh on the hidden test set (supplementary material).

5.7. Limitations

The clean-prior gain on WSRD+ (0.14 dB) is smaller than on ISTD and ISTD+, and our single model is below the winning challenge entry, which uses a multi-stage cascade with checkpoint ensembling [2]. Checkpoints were selected on the evaluation sets for both models and the design choices on the WSRD+ validation set, and we have no hidden-test score for DenseSR. The robustness test compares complete models with synthetic corruptions and does not isolate DMMA, and we did not evaluate DenseSR trained with prior dropout, an additive log-bias modulation, a model without priors, or oracle priors from the shadow-free image. The prior analysis measures consistency rather than accuracy, the causal role of the gate is untested, Figure 5 does not include DenseSR, and we have not measured perceptual quality, for example with LPIPS [52].

5.8. Broader Impact

Shadow removal can help detection, tracking and 3D reconstruction, but it is also image editing: shadow geometry is used in forensics to detect manipulated images [17], so removing shadows could make such edits harder to detect. Our outputs are not marked as edited, and the frozen models add compute and energy cost.

6. Conclusion

Depth, normal and semantic priors from pretrained models see the same shadowed image as the restoration network, and our full-reference test shows that they change most inside shadows. ShadowGDNet gates the skip connections and decouples the priors inside attention; it improves DenseSR on all three benchmarks, mainly through the architecture, and loses about a third as much PSNR when the priors are corrupted. As foundation-model priors become standard inputs to restoration networks, their reliability under the degradation being removed should be measured, and full-reference consistency tests such as ours offer a simple way to do so.

References

[1] Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. Estimating and exploiting the aleatoric uncertainty in surface normal estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13137–13146, 2021. 2

[2] Lorenzo Beltrame, Jules Salzinger, Filip Svoboda, Jasmin Lampert, Phillipp Fanta-Jende, Radu Timofte, and Marco Körner. Winner of CVPR2026 NTIRE challenge on image shadow removal: Semantic and geometric guidance for shadow removal via cascaded refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2026. 8

[3] Alexey Bochkovskiy, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun. Depth Pro: Sharp monocular metric depth in less than a second. In *International Conference on Learning Representations (ICLR)*, 2025. 2, 3, 4, 6, 7

[4] Pierre Charbonnier, Laure Blanc-Féraud, Gilles Aubert, and Michel Barlaud. Two deterministic half-quadratic regularization algorithms for computed imaging. In *Proceedings of the 1st IEEE International Conference on Image Processing (ICIP)*, pages 168–172, 1994. 6

[5] Zipei Chen, Chengjiang Long, Ling Zhang, and Chunxia Xiao. CANet: A context-aware network for shadow removal. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4743–4752, 2021. 2

[6] Xiaodong Cun, Chi-Man Pun, and Cheng Shi. Towards ghost-free shadow removal via dual hierarchical aggregation network and shadow matting GAN. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10680–10687, 2020. 6, 7

[7] Samuel Dodge and Lina Karam. Understanding how image quality affects deep neural networks. In *Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6, 2016. 2

[8] Wei Dong, Han Zhou, Yuqiong Tian, Jingke Sun, Xiaohong Liu, Guangtao Zhai, and Jun Chen. ShadowRefiner: Towards mask-free shadow removal via fast Fourier transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 6208–6217, 2024. 2, 6, 7

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. 2

[10] Graham D. Finlayson, Steven D. Hordley, Cheng Lu, and Mark S. Drew. On the removal of shadows from images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(1):59–68, 2006. 2

[11] Lan Fu, Changqing Zhou, Qing Guo, Felix Juefei-Xu, Hongkai Yu, Wei Feng, Yang Liu, and Song Wang. Auto-exposure fusion for single-image shadow removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10571–10580, 2021. 2

[12] Lanqing Guo, Siyu Huang, Ding Liu, Hao Cheng, and Bihan Wen. ShadowFormer: Global context helps shadow removal. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 710–718, 2023. 2, 6, 7

[13] Ruiqi Guo, Qieyun Dai, and Derek Hoiem. Paired regions for shadow detection and removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2956–2967, 2013. 2

[14] Shengfeng He, Bing Peng, Junyu Dong, and Yong Du. Mask-ShadowNet: Toward shadow removal via masked adaptive instance normalization. *IEEE Signal Processing Letters*, 28: 957–961, 2021. 2

[15] Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 2019. 2, 8

[16] Yeying Jin, Aashish Sharma, and Robby T. Tan. DC-ShadowNet: Single-image hard and soft shadow removal using unsupervised domain-classifier guided network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5027–5036, 2021. 2, 6, 7

[17] Eric Kee, James F. O'Brien, and Hany Farid. Exposing photo manipulation with inconsistent shadows. *ACM Transactions on Graphics*, 32(3):28:1–28:12, 2013. 8

[18] Lingdong Kong, Shaoyuan Xie, Hanjiang Hu, Lai Xing Ng, Benoit Cottereau, and Wei Tsang Ooi. RoboDepth: Robust out-of-distribution depth estimation under corruptions. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, pages 21298–21342, 2023. 2, 8

[19] Hieu Le and Dimitris Samaras. Shadow removal via shadow image decomposition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8578–8587, 2019. 2, 3, 6, 7

[20] Chenghua Li, Bo Yang, Zhiqi Wu, Gao Chen, Yihan Yu, and Shengxiao Zhou. Shadow removal based on diffusion, segmentation and super-resolution models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 6045–6054, 2024. 1

[21] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. SwinIR: Image restoration using Swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 1833–1844, 2021. 2

[22] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, 2017. 3

[23] Yu-Fan Lin, Chia-Ming Lee, and Chih-Chung Hsu. DenseSR: Image shadow removal as dense prediction. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM)*, pages 7026–7035, 2025. 2, 3, 4, 5, 6, 7

[24] Wenze Liu, Hao Lu, Hongtao Fu, and Zhiguo Cao. Learning to upsample by learning to sample. In *Proceedings of*

the IEEE/CVF International Conference on Computer Vision (ICCV), pages 6027–6037, 2023. 3

[25] Yuhao Liu, Zhanghan Ke, Ke Xu, Fang Liu, Zhenwei Wang, and Rynson W. H. Lau. Recasting regional lighting for shadow removal. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3810–3818, 2024. 2

[26] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021. 2, 3, 5

[27] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. 6

[28] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y. Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention U-Net: Learning where to look for the pancreas. In *Medical Imaging with Deep Learning (MIDL)*, 2018. 3

[29] Liangqiong Qu, Jiandong Tian, Shengfeng He, Yandong Tang, and Rynson W. H. Lau. DeshadowNet: A multi-context embedding deep network for shadow removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4067–4075, 2017. 2

[30] Wenqi Ren, Lin Ma, Jiawei Zhang, Jinshan Pan, Xiaochun Cao, Wei Liu, and Ming-Hsuan Yang. Gated fusion network for single image dehazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3253–3261, 2018. 3

[31] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241. Springer, 2015. 3

[32] Andres Sanin, Conrad Sanderson, and Brian C. Lovell. Improved shadow removal for robust person tracking in surveillance scenarios. In *Proceedings of the 20th International Conference on Pattern Recognition (ICPR)*, pages 141–144, 2010. 1

[33] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seung Eun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *Transactions on Machine Learning Research*, 2026. 2, 3, 4, 6

[34] Shangquan Sun, Wenqi Ren, Xinwei Gao, Rui Wang, and Xiaochun Cao. Restoring images in adverse weather conditions via histogram transformer. In *Computer Vision – ECCV 2024*, pages 111–129. Springer, 2024. 2

[35] Florin-Alexandru Vasluianu, Tim Seizinger, and Radu Timofte. WSRD: A novel benchmark for high resolution image shadow removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1826–1835, 2023. 1, 2, 6, 7

[36] Florin-Alexandru Vasluianu, Tim Seizinger, Zhuyun Zhou, Zongwei Wu, Cailian Chen, Radu Timofte, Wei Dong, Han Zhou, Yuqiong Tian, Jun Chen, Xueyang Fu, Xin Lu, Yurui Zhu, Xi Wang, Dong Li, Jie Xiao, Yunpeng Zhang, Zheng-Jun Zha, Zhao Zhang, Suiyi Zhao, Bo Wang, Yan Luo, Yanyan Wei, Zhihao Zhao, Long Sun, Tingting Yang, Jinshan Pan, Jiangxin Dong, Jinhui Tang, Bilel Benjdira, Mohammed Nassif, Anis Koubaa, Ahmed Elhayek, Anas M. Ali, Kyotaro Tokoro, Kento Kawai, Kaname Yokoyama, Takuya Seno, Yuki Kondo, Norimichi Ukita, Chenghua Li, Bo Yang, Zhiqi Wu, Gao Chen, Yihan Yu, Sixiang Chen, Kai Zhang, Tian Ye, Wenbin Zou, Yunlong Lin, Zhaohu Xing, Jinbin Bai, Wenhao Chai, Lei Zhu, Ritik Maheshwari, Rakshank Verma, Rahul Tekchandani, Praful Hambarde, Satya Narayan Tazi, Santosh Kumar Vipparthi, Subrahmanyam Murala, Jaeho Lee, Seongwan Kim, S. M. A. Sharif, Nodirkhuja Khujaev, Roman Tsoy, Fan Gao, Weidan Yan, Wenze Shao, Dengyin Zhang, Bin Chen, Siqi Zhang, Yanxin Qian, Yuanbin Chen, Yuanbo Zhou, Tong Tong, Rongfeng Wei, Ruiqi Sun, Yue Liu, Nikhil Akalwadi, Amogh Joshi, Sampada Malagi, Chaitra Desai, Ramesh Ashok Tabib, Uma Mudenagudi, Ali Murtaza, Uswah Khairuddin, Ahmad 'Athif Mohd Faudzi, Adinath Dukre, Vivek Deshmukh, Shruti S. Phutke, Ashutosh Kulkarni, Anil Gonde, Arun Karthik K, Manasa N, Shri Hari Priya, Wei Hao, Xingzhuo Yan, and Minghan Fu. NTIRE 2024 image shadow removal challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 6547–6570, 2024. 6

[37] Jifeng Wang, Xiang Li, and Jian Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1788–1797, 2018. 2, 6, 7

[38] Jiaqi Wang, Kai Chen, Rui Xu, Ziwei Liu, Chen Change Loy, and Dahua Lin. CARAFE: Content-aware ReAssembly of FEatures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3007–3016, 2019. 3

[39] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 3, 6

[40] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general U-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17683–17693, 2022. 2

[41] Silvan Weder, Guillermo Garcia-Hernando, Áron Monszpart, Marc Pollefeys, Gabriel J. Brostow, Michael Firman, and Sara Vicente. Removing objects from neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16528–16538, 2023. 1

[42] Jie Xiao, Xueyang Fu, Feng Wu, and Zheng-Jun Zha.

Stochastic window transformer for image restoration. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 9315–9329, 2022. 2

[43] Jie Xiao, Xueyang Fu, Man Zhou, Hongjian Liu, and Zheng-Jun Zha. Random shuffle transformer for image restoration. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pages 38039–38058, 2023. 2

[44] Jie Xiao, Xueyang Fu, Yurui Zhu, Dong Li, Jie Huang, Kai Zhu, and Zheng-Jun Zha. HomoFormer: Homogenized transformer for image shadow removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25617–25626, 2024. 2, 7

[45] Jie Xiao, Xueyang Fu, Yurui Zhu, and Zheng-Jun Zha. Bayesian window transformer for image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):2431–2444, 2026. 2

[46] Jiamin Xu, Zelong Li, Yuxin Zheng, Chenyu Huang, Renshu Gu, Weiwei Xu, and Gang Xu. OmniSR: Shadow removal under direct and indirect lighting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8887–8895, 2025. 2, 6, 7

[47] Jiamin Xu, Yuxin Zheng, Zelong Li, Chi Wang, Renshu Gu, Weiwei Xu, and Gang Xu. Detail-preserving latent diffusion for stable shadow removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7592–7602, 2025. 6, 7

[48] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything V2. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 21875–21911, 2024. 2, 7

[49] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4471–4480, 2019. 3

[50] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for real image restoration and enhancement. In *Computer Vision – ECCV 2020*, pages 492–511. Springer, 2020. 3

[51] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5728–5739, 2022. 2

[52] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018. 3, 8

[53] Yurui Zhu, Jie Huang, Xueyang Fu, Feng Zhao, Qibin Sun, and Zheng-Jun Zha. Bijective mapping network for shadow removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5627–5636, 2022. 2, 6, 7