PrismNet: Semantic-Aware Image Enhancement via Vision Transformer and Zero-Cost Gating

Abstract

The enhancement of high-fidelity images necessitates precise, spatially-varying adjustments, a task for which extant methodologies frequently compromise between performance and computational efficiency, particularly in the context of high-resolution processing. In this paper, we posit that the foundation of superior enhancement lies in the backbone's capacity for semantic understanding. A systematic investigation of modern architectures has been conducted, revealing the significant advantage of Vision Transformers (ViTs) in providing global, context-aware features that are essential for this task. In light of this key insight, we put forth PrismNet, a novel framework that employs a standard, pre-trained ViT to facilitate semantic-aware enhancement. The core of PrismNet is a zero-cost semantic gating mechanism, which effectively repurposes attention scores from a lightweight parameter head to enable content-aware local control without an auxiliary prediction network. This efficient-by-design approach enables PrismNet to process 4K-resolution images at over 400 frames per second (FPS), while achieving state-of-the-art results that surpass contemporaneous methods on the MIT-Adobe FiveK benchmark. The present study offers a practical, high-performance enhancement model, thereby demonstrating that leveraging the semantic power of pre-trained ViTs is an important and promising direction for the future of image enhancement.

1. Introduction

The widespread use of digital cameras and smartphones has democratized photography. However, the artistic and technical skills required to produce professional-quality photographs remain challenging for most users to master. To address this limitation, deep learning-based image enhancement has emerged as a potent paradigm. In contrast to objective restoration tasks, such as denoising [7, 48] or deblurring [10], high-end photographic retouching is a more nuanced process. A comprehensive understanding of the image's semantic attributes, encompassing its subject, composition, and lighting conditions, is imperative to implement adjustments that are not only technically accurate but also aesthetically pleasing. Consequently, a genuinely practical and state-of-the-art enhancement method must overcome the limitations of generic restoration networks and be informed by a profound, content-aware understanding of the image.

Presently, the prevailing paradigms within the field have placed significant emphasis on the architectural design of the enhancement head. A significant line of work involves the training of fully convolutional networks (FCNs) [2, 5, 6, 14, 21, 28, 45] to directly map input pixels to their enhanced counterparts. While this direct generation approach has the capacity to yield remarkable results, it is confronted by substantial challenges. The computational complexity of FCNs generally increases quadratically with the spatial dimensions, rendering them ill-suited for the ever-increasing resolutions of modern cameras and making real-time applications challenging. Additionally, the inherently local nature of convolutions can impede their capacity to model long-range dependencies, which are imperative for ensuring global color and tonal harmony across the entire scene.

An alternative, more efficient paradigm is to predict a compact set of parameters for a global, differentiable transformation [11]. This parametric approach decouples the feature extraction from the final rendering process, offering remarkable efficiency. However, an examination of existing implementations indicates that this promising paradigm has been predominantly associated with lightweight, from-scratch CNNs, frequently in a quest for maximum speed that often comes at the expense of representational power. Consequently, the full potential of a truly semantic-driven parametric enhancement, one that could perform complex adjustments based on a rich understanding of the scene, has remained largely untapped.

This analysis reveals a common thread across different approaches: a design philosophy that prioritizes innovation in the downstream head while treating the backbone network as a secondary, less critical component. This leads us to a pivotal, albeit contrarian, question: What if the key to the next leap in image enhancement lies not in a more complex or refined head but in a more powerful backbone? Departing from low-level vision conventions, we shift the focus to the front end of the model and investigate using a

large-scale, pretrained Vision Transformer (ViT) [9] as the core feature extractor. We demonstrate that, with its inherent global self-attention mechanism, ViT provides a far superior semantic representation for this task. ViT's ability to model complex, long-range dependencies across the entire image enables it to achieve holistic scene understanding. This, in turn, leads to more stable and robust mapping from image content to enhancement decisions.

Building on this discovery, we propose PrismNet, a framework designed to fully capitalize on the strengths of ViT. PrismNet embodies a new design philosophy: leveraging a powerful, semantic-rich backbone to guide a deliberately simple and efficient enhancement head. The core of our framework is a novel attention-derived semantic gate, which is effectively derived from the internal attention scores of our ParamHead. This mechanism enables precise, content-aware spatial control without the need for an auxiliary prediction network, thus preserving the efficiency of the parametric approach. This efficient-by-design architecture allows PrismNet to achieve state-of-the-art results on the MIT-Adobe FiveK [3] benchmark, while processing 4K images at over 404 FPS.

Our contributions are as follows:

- We provide a systematic analysis that reveals the decisive role of the backbone's semantic representation in image enhancement, and demonstrate the significant superiority of a pre-trained Vision Transformer [9] for this task.
- We propose PrismNet, a new state-of-the-art model featuring a zero-cost semantic gate, which embodies our principle of leveraging a powerful representation to guide a simple transformation.
- We conduct extensive experiments and in-depth analysis, including visualization of internal attention mechanisms, to provide a clear explanation for why ViT's representation is more effective.
- Our work presents a practical and high-performance paradigm, achieving state-of-the-art results on the MIT-Adobe FiveK [3] benchmark while processing 4K images in real-time.

2. Related Work

2.1. Direct Image Generation Methods

The direct generation paradigm treats image enhancement as an image-to-image translation task, often employing encoder-decoder architectures like U-Net [37]. Subsequent works leverage Generative Adversarial Networks (GANs) [12] for both paired (e.g., Pix2Pix [19]) and unpaired (e.g., CycleGAN [52], EnlightenGAN [20]) training regimes. While powerful, these methods are computationally expensive, difficult to train, and can introduce artifacts or fail to maintain long-range consistency due to their local receptive fields.

2.2. Parametric Enhancement Methods

To address the efficiency limitations of direct generation, parametric methods predict a small set of parameters for a global, differentiable transformation. While traditional methods often rely on handcrafted parametric models [17], deep learning approaches learn these parameters from data. Pioneered by HDRNet [11] which predicts coefficients for a bilateral grid, this approach has been extended to parameterizing color curves [34] and 3D Lookup Tables (LUTs) [47], sometimes even without reference images (e.g., Zero-DCE [13]). These methods are exceptionally fast and artifact-free. However, to remain lightweight, they are typically built upon small CNNs, limiting their capacity for the deep semantic understanding needed for complex, content-aware adjustments.

2.3. Sequential Enhancement Methods

Another line of work models image enhancement as an iterative process that mimics a human expert's workflow. For instance, some methods employ deep reinforcement learning (DRL) to learn an optimal sequence of discrete editing actions, such as adjusting brightness or contrast [35]. While this approach offers interpretability, it often relies on a fixed set of discrete actions and can involve a complex, iterative inference process.

2.4. Transformers in Low-Level Vision

Transformers have demonstrated remarkable success in low-level vision. One line of work focuses on designing specialized, from-scratch architectures for objective restoration [25, 42, 46]. Another direction explores task-specific models for enhancement, like the lightweight STAR [51] for real-time applications, and the low-light specific Retinexformer [4]. While pioneering, these approaches either focus on pixel-level reconstruction or rely on bespoke, non-pre-trained backbones, limiting their access to broad semantic priors. The potential of leveraging a standard, large-scale pre-trained Vision Transformer (ViT) [9] for the general and subjective task of photographic enhancement remains largely unexplored. Our work bridges this gap, demonstrating that a pre-trained ViT can serve as a superior semantic backbone within a highly efficient parametric framework.

3. Methodology

3.1. Overall Framework and Design Philosophy

The design of PrismNet is founded on a central principle: the quality of the backbone's semantic representation is the primary determinant of the final enhancement quality. This marks a conceptual shift from conventional methods, which have historically focused on engineering complex enhancement heads (e.g., sophisticated LUT predictors, multi-scale

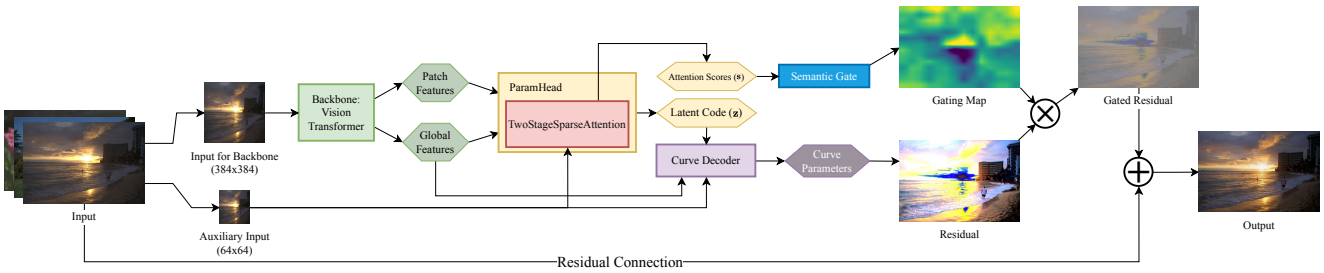


Figure 1. The overall architecture of PrismNet. An input image is processed by a pre-trained ViT backbone to extract a holistic semantic representation. This understanding is then processed by our Semantic Guidance Unit (ParamHead) to generate a latent code and, crucially, a zero-cost semantic gate derived from its internal attention. A lightweight Curve Decoder translates this latent code into global curve parameters. Finally, the adjustment residual computed from these curves is spatially modulated by the semantic gate to produce the final, high-fidelity enhanced image.

generative networks), while treating the feature extraction backbone as a secondary component. Our philosophy inverts this priority. We posit that a powerful, semantic-rich backbone is not just a passive feature extractor, but the very foundation for all subsequent, context-aware enhancement decisions. A profound initial understanding of the image's content—its subject, composition, lighting, and mood—is an indispensable prerequisite for achieving superior, professional-grade results.

This guiding principle organizes our framework into a coherent, three-stage process designed to translate high-level semantic understanding into precise, pixel-level adjustments. The entire pipeline is illustrated in Fig. 1:

1. **Semantic Representation.** The process begins with a holistic perception of the image, moving beyond low-level statistics (like texture and edges) to grasp its compositional and semantic structure. To achieve this, we employ a standard, pre-trained Vision Transformer (ViT) [9] on a resized version of the input. This stage leverages the extensive knowledge of the pre-trained model to extract a multi-level, globally-aware representation that captures the intricate relationships between different parts of the scene.
2. **Semantic-to-Parameter Mapping.** With this rich semantic understanding, the model formulates a targeted enhancement strategy. This translation from perception to action is the role of our lightweight Semantic Guidance Unit, termed the ParamHead. It distills the high-dimensional ViT features into a compact, low-dimensional latent code, which represents the intended enhancement plan. Concurrently, a simple Curve Decoder translates this abstract plan into a minimal set of parametric curves defining a global, differentiable color transformation.
3. **Gated Application of Adjustments.** Finally, the model executes the enhancement with precise, spatially-varying control. A global adjustment, if applied uniformly, can be a blunt instrument, often leading to unnatural results in heterogeneous scenes (e.g., oversaturating a vivid sky while enhancing a foreground subject). To address this, our key innovation is a novel zero-cost semantic gate. Derived directly from the ParamHead's internal attention scores without extra overhead, this gate spatially modulates the global adjustment. This mechanism enables precise local control, strengthening enhancements in salient regions while suppressing them where unneeded.

3.2. Backbone for Semantic Representation

Our approach deliberately departs from the common practice in low-level vision, which favors lightweight, from-scratch CNNs for efficiency. Instead, we leverage the power of a standard `vit_base_patch16_384` model [9], pre-trained on the large-scale ImageNet-21K dataset [8], as our core semantic backbone. For processing, an input image is first resized to a fixed 384×384 resolution before being fed into this model.

This choice is motivated by the unique advantages of a pre-trained ViT for a nuanced task like subjective image enhancement. For instance, low-light image enhancement primarily focuses on recovering information from darkness by adjusting illumination [4, 24, 29, 40, 43, 50], while other methods concentrate on specific colorimetric corrections like white-balance editing [1]. In contrast, general-purpose enhancement requires a holistic understanding of a well-lit scene to perform complex, multi-faceted, content-aware adjustments. Conventional CNNs, often reliant on normalization layers [18], are constrained by local receptive fields and must stack many layers to build a global context. In stark contrast, a ViT's global self-attention mechanism enables it to model long-range dependencies across the entire image from the very first layer. This capability is crucial, allowing the model to reason about complex, non-local interactions, such as considering the background's ambient lighting when adjusting a portrait subject's skin tone, even if spatially distant. We utilize the patch token features from

the final ViT encoder layer, tapping into a vast reservoir of semantic knowledge from a high-level classification task. Our central hypothesis, validated in Sec. 4, is that these rich, global priors—concerning scene composition, object relationships, and mood—provide a more robust and reliable foundation for enhancement than what smaller, task-specific datasets can offer alone.

3.3. Semantic-to-Parameter Mapping

Our enhancement head serves as an efficient bridge, translating the ViT's profound semantic understanding into specific, actionable transformation parameters. It is deliberately lightweight, embodying our philosophy that a powerful representation requires only a concise and effective tool for execution. Our lightweight CNN auxiliary branch follows the design principles of modern residual networks [15].

3.3.1. ParamHead with Two-Stage Sparse Attention

The ViT backbone provides a high-dimensional feature representation, $\mathbf{F} \in \mathbb{R}^{B \times N_p \times C}$, where the number of patch tokens N_p can be large. The ParamHead's primary task is to distill this rich information into a concise latent vector $\mathbf{z} \in \mathbb{R}^D$. To efficiently navigate this space and avoid the prohibitive quadratic complexity of standard self-attention, it employs a highly effective Two-Stage Sparse Attention mechanism, inspired by the efficiency of sparse, query-based sampling in object detection, such as Deformable DETR [53]. This process can be conceptualized as a sequence of coarse-to-fine analysis. We note that this hierarchical strategy, combining coarse-grained global context with fine-grained local details, shares conceptual similarities with concurrent work in natural language processing, such as Native Sparse Attention (NSA) [44], highlighting a convergent evolution in designing efficient attention mechanisms across different domains.

Stage 1: Coarse-grained Attention. First, the mechanism performs a coarse-grained, global analysis to identify regions of high semantic importance. We downsample the features $\mathbf{F}$ to a smaller grid $\mathbf{F}_{low} \in \mathbb{R}^{B \times N_g \times C}$ ($N_g \ll N_p$) via adaptive average pooling. A single, learnable context query vector, $\mathbf{q}_{ctx}$, then performs cross-attention over this coarse map. This acts as a rapid reconnaissance to identify the most salient regions without processing the full-resolution feature map. The attention mechanism is enhanced with Rotary Positional Embeddings (RoPE) [39] to better encode positional information. The raw attention scores (logits) from this stage, $\mathbf{s} \in \mathbb{R}^{B \times 1 \times N_g}$, are computed as:

$$\mathbf{s} = \text{Attention}(\mathbf{q}_{ctx}, \mathbf{F}_{low}) \tag{1}$$

These scores $\mathbf{s}$, which capture a low-resolution map of semantic importance, are preserved for our zero-cost gating mechanism (see Sec. 3.4). Based on the attention weights, we identify the Top-K grid locations with the highest attention, denoted as $\mathcal{I}_{topK}$.

Stage 2: Fine-grained Attention. With the most important regions identified, the mechanism proceeds to a focused, fine-grained analysis. We gather the original, full-resolution patch features $\mathbf{F}$ within these Top-K regions, forming a sparse yet detail-rich set of features $\mathbf{F}_{focus}$. The same context query $\mathbf{q}_{ctx}$ performs a second, more focused cross-attention on this curated feature set $\mathbf{F}_{focus}$. This yields a context summary vector $\mathbf{c}$, efficiently aggregating the most vital spatial information without the cost of a full attention operation on all patch tokens.

Final Latent Code Generation. The final latent code $\mathbf{z}$, encapsulating the complete enhancement strategy, is synthesized by fusing three distinct information streams for a comprehensive and robust basis. The streams are: (1) the global average pooled feature $\mathbf{f}_{\text{gap}}$ from the ViT backbone, providing a holistic summary; (2) the context summary vector $\mathbf{c}$ from the sparse attention, capturing salient local details; and (3) an auxiliary feature vector $\mathbf{f}_{\text{cnn}}$ from a small, parallel CNN on a low-resolution input, offering complementary spatial cues. This lightweight design is inspired by efficient mobile architectures [16]. The fusion is performed by a multi-layer perceptron (MLP):

$$\mathbf{z} = \text{MLP}(\text{concat}[\mathbf{f}_{\text{gap}}, \mathbf{c}, \mathbf{f}_{\text{cnn}}]) \tag{2}$$

3.3.2. Curve Decoder

The Curve Decoder is a lightweight MLP acting as the final translator, converting the abstract enhancement strategy in the latent vector $\mathbf{z}$ into concrete, differentiable transformation parameters. To ensure access to rich, uncompressed global context, its input concatenates the latent code $\mathbf{z}$, the original global feature $\mathbf{f}_{\text{gap}}$ from the ViT, and features from a low-resolution input. It predicts control points for a set of 9 parametric color curves, following the formulation of prior work [38] but simplified to its core components to reduce parametric burden. These curves define a powerful, compact global color adjustment. The adjustment for each output channel $j \in \{R, G, B\}$ is a sum of transformations from all input channels $i \in \{R, G, B\}$, allowing for sophisticated cross-channel corrections:

$$\mathbf{R}_j = \sum_{i \in \{R,G,B\}} \text{ApplyLUT}(\mathbf{I}_{in,i}, \mathbf{p}_{ji}) \tag{3}$$

where $\mathbf{I}_{in,i}$ is the i-th input channel, $\mathbf{p}_{ji}$ are the predicted curve parameters from input channel i to output channel j, and ApplyLUT is the differentiable lookup table operation.

3.4. Zero-Cost Semantic Gating

A key innovation of PrismNet is its ability to perform sophisticated, spatially-varying adjustments without a dedicated, costly spatial prediction network. This is achieved through a concise and efficient mechanism: the "zero-cost semantic gate."

The insight is that the map of semantic importance required for localized control is already implicitly computed during feature extraction. Our zero-cost semantic gate, $\mathbf{M} \in [0,1]^{H \times W}$, is derived directly from the attention scores $\mathbf{s}$ generated during the ParamHead's coarse-grained attention stage (Eq. (1)). These scores, representing what the model "cared about," are repurposed for spatial modulation at a negligible computational cost. The generation is straightforward: the 1D scores vector $\mathbf{s}$ is reshaped into a 2D grid, upsampled to the target image resolution via bilinear interpolation, and passed through a sigmoid function to normalize its values to $[0,1]$:

$$\mathbf{M} = \sigma(\text{Upsample}(\text{Reshape}(\mathbf{s}))). \tag{4}$$

This gate, $\mathbf{M}$, functions as a continuous spatial modulator. The final enhanced image $\mathbf{I}_{out}$ is synthesized by applying the global adjustment residual $\mathbf{R}$ (from the predicted curves), intelligently modulated by this semantic gate:

$$\mathbf{I}_{out} = \text{clamp}(\mathbf{I}_{in} + \mathbf{R} \odot \mathbf{M}, 0, 1), \tag{5}$$

where $\mathbf{I}_{in}$ is the input image and $\odot$ denotes element-wise multiplication. This mechanism transforms a uniform global strategy into a work of localized, precise, context-aware adjustments, amplifying effects in important regions (e.g., a face) while suppressing them in others (e.g., the sky), all guided by the ViT's holistic understanding.

3.5. Implementation Details

Training Strategy. Our model is trained end-to-end using the AdamW optimizer [31] with a weight decay of 1×10^{-2}. To ensure stable fine-tuning of the large pre-trained backbone, we employ a differential learning rate strategy. The pre-trained ViT parameters are fine-tuned with a small learning rate of 1×10^{-5} ($0.1 \times$ LR), while newly initialized components, like the ParamHead and Curve Decoder, use a base learning rate of 1×10^{-4}. The learning rate is dynamically adjusted using a Cosine Annealing scheduler [30]. To prevent gradient explosion, we apply gradient clipping with a maximum L2 norm of 1.0.

Two-Stage Loss Function. To effectively guide the network towards perceptually pleasing results, we adopt a two-stage training curriculum, switching the loss function based on model performance. This guides the model to first learn structurally and perceptually plausible enhancements, then meticulously fine-tune color fidelity.

- **Stage 1: Perceptual Baseline Training.** The model is first trained with a composite loss function $\mathcal{L}_{\text{stage1}}$, designed to establish a strong perceptual baseline:

$$\mathcal{L}_{\text{stage1}} = \lambda_{\text{p}}\mathcal{L}_{\text{LPIPS}} + \lambda_{\text{c}}\mathcal{L}_{\text{color}} + \lambda_{\text{s}}(1 - \text{SSIM}) \tag{6}$$

  where $\mathcal{L}_{\text{LPIPS}}$ [49] is the Learned Perceptual Image Patch Similarity loss based on a pre-trained VGG network. $\mathcal{L}_{\text{color}}$ is an L1 loss in HSV space penalizing deviations in hue, saturation, and value. The SSIM term encourages structural preservation. The weights are empirically set to $\lambda_{\text{p}} = 1.0, \lambda_{\text{c}} = 1.0, \lambda_{\text{s}} = 1.0$.
- **Stage 2: Color Fidelity Fine-tuning.** Once the model's validation PSNR reaches 26.0, we transition to a fine-tuning stage. The loss function is switched to a simple yet effective L1 loss in the perceptually uniform CIELab color space, where Euclidean distances better correspond to human-perceived color differences. This loss, $\mathcal{L}_{\text{stage2}}$, further refines the model's color accuracy:

$$\mathcal{L}_{\text{stage2}} = \mathcal{L}_1(\text{RGB2Lab}(\mathbf{I}_{out}), \text{RGB2Lab}(\mathbf{I}_{gt})) \tag{7}$$

This two-stage curriculum proved crucial for our final model, ensuring both training stability and superior final quality.

4. Experiments

4.1. Experimental Setup

4.1.1. Datasets and Metrics

We conduct all our experiments on the widely-used MIT-Adobe FiveK dataset [3], which is the de-facto standard for evaluating learning-based image enhancement methods. Following the common protocol used in recent image enhancement literature [4, 38, 47], we use the 4,500 images retouched by expert C for training and the remaining 500 images for rigorous testing. We evaluate the performance of all methods using three standard and complementary metrics: Peak Signal-to-Noise Ratio (PSNR) for pixel-level fidelity, Structural Similarity Index (SSIM) for perceptual structural similarity, and the Learned Perceptual Image Patch Similarity (LPIPS) [49] computed with the AlexNet [23] backbone for deep feature-based perceptual quality.

4.1.2. Implementation Details

Our model, PrismNet, is implemented in PyTorch [36]. For the backbone, we use a `vit_base_patch16_384` model with weights pre-trained on ImageNet-21K. The entire network is trained end-to-end using the AdamW optimizer [31] with an initial learning rate of 1×10^{-4} and a weight decay of 1×10^{-2}. The model is trained for 150 epochs with a batch size of 24 on a single NVIDIA RTX 5090 GPU. We employ a two-stage training strategy for stable convergence: the model is first trained using a composite loss of LPIPS

Input | StarEnhancer [38] | Retinexformer [4] | DeepLPF [33] | CURL [34] | **PrismNet (Ours)** | Ground Truth

Figure 2. Qualitative comparison with state-of-the-art methods on a challenging example from the MIT-Adobe FiveK test set. Our method, PrismNet, produces a result that is perceptually closer to the ground truth, especially in restoring the color harmony of the sky and preserving details in the dark flowers, avoiding the color cast or over-smoothing effects seen in other methods.

and an HSV-based color loss, which quickly establishes a good perceptual baseline. After the PSNR reaches a threshold of 26.0, we switch to a CIELab L1 loss for fine-tuning, which further refines color accuracy in a perceptually uniform space.

Table 1. Quantitative comparison with state-of-the-art methods on the MIT-Adobe FiveK dataset (Original Image to Expert C) in the original resolution. Best results are in **bold**, second best are underlined. Some results are replicated from [38, 47].

Method	PSNR ↑	SSIM ↑	LPIPS ↓
DPE [6]	22.15	0.850	–
DeepUPE [41]	23.24	0.893	0.158
HDRNet [11]	24.03	0.919	–
DeepLPF [33]	24.45	0.935	0.089
CURL [34]	24.50	0.931	0.103
Retinexformer [4]	24.93	0.938	0.118
A3DLUT [47]	25.10	0.930	–
StarEnhancer [38]	25.36	0.941	0.079
PrismNet	**26.66**	**0.952**	**0.036**

4.2. Comparison with State-of-the-Art Methods

We compare PrismNet with a comprehensive set of state-of-the-art methods, including direct generation approaches (e.g., DPE [6]), parametric methods (e.g., StarEnhancer [38], HDRNet [11]), and recent Transformer-based restoration models (e.g., Retinexformer [4]). The quantitative results are summarized in Tab. 1. Our method achieves the best performance across all three metrics, demonstrating its overall superiority. Notably, PrismNet significantly outperforms other parametric methods like StarEnhancer, highlighting the profound benefit of replacing a lightweight CNN with a powerful, pre-trained ViT backbone.

Figure 2 provides a visual comparison on challenging images with complex lighting and content. It is evident that PrismNet produces more visually pleasing and faithful results. For instance, in the backlit scene, our method successfully restores the vibrant colors in the sky while preserving details in the dark foreground, avoiding the color artifacts or flat results produced by other methods. This qualitative superiority further validates the effectiveness of our semantic-driven approach.

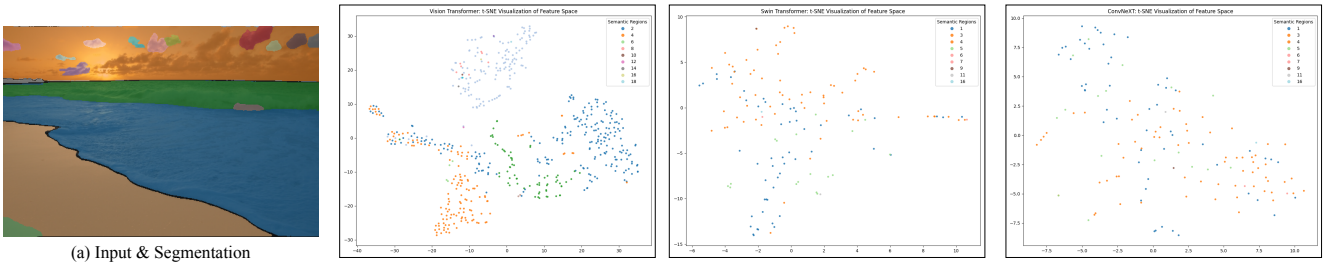
4.3. Analysis and Ablation Studies

To thoroughly validate our design choices and provide deeper insights into our model's behavior, we conduct a series of carefully designed ablation studies.

4.3.1. The Decisive Role of the Backbone

Our central claim is that the backbone's representational power is paramount. To verify this, we replace our ViT-base with other popular architectures of similar scale, including Swin Transformer [26] and ConvNeXt [27], while keeping the rest of our framework unchanged. The results, presented in Tab. 2, show a substantial performance gap between the backbones. The ViT-based model significantly outperforms all others, with a performance gap of over 0.47 dB in PSNR compared to the next best competitor, ConvNeXt. This substantial difference underscores that the choice of backbone is not a minor detail but a primary determinant of performance. Crucially, this performance is contingent on leveraging pre-trained weights; a ViT backbone trained from scratch on this dataset failed to converge (PSNR $\approx$ 18.0 dB), confirming the critical impact of large-scale pre-training.

To understand the fundamental reason why ViT excels, we visualize the internal self-attention maps from the final layer of both ViT and Swin-T in Fig. 4. The visualization clearly shows that ViT's attention is inherently global; a query patch representing a part of an object can attend to all other patches across the entire image, allowing it to form a holistic, object-level understanding of the scene. In stark contrast, Swin-T's attention is strictly confined to



(a) Input & Segmentation (b) t-SNE of ViT Features (c) t-SNE of Swin-T Features (d) t-SNE of ConvNeXt Features

Figure 3. t-SNE visualization of the feature spaces produced by different backbones for the same input image. Each point represents a pixel feature, colored by its semantic region ID derived from an automatic segmentation [22]. (b) The feature space of our ViT-based model exhibits remarkable semantic clustering, with pixels from the same region forming tight, well-separated clusters. (c, d) In stark contrast, the feature spaces of Swin Transformer and ConvNeXt show significant feature entanglement, where different semantic regions are largely inseparable. **Note**: The region IDs and corresponding colors are determined independently for each backbone's visualization and are not directly comparable across subfigures; the key insight lies in the internal clustering structure of each feature space.

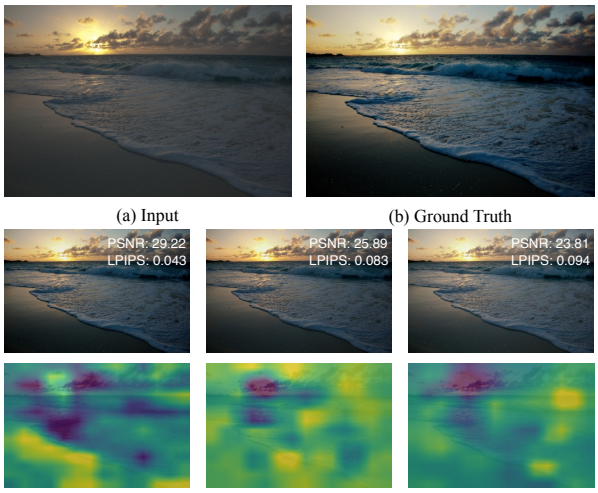
Table 2. Ablation study on backbone architectures. All models are trained with the full PrismNet framework, only varying the backbone. The results highlight the significant performance advantage of using a ViT backbone across all metrics, especially the perceptual LPIPS score.

Backbone	Pre-trained	PSNR ↑	SSIM ↑	LPIPS ↓	Parameters ↓
ConvNeXt-Base [27]	ImageNet-22K	26.19	0.948	0.038	94.27M
Swin-Base [26]	ImageNet-22K	26.14	0.947	0.038	93.59M
ViT-Base (Ours)	**ImageNet-21K**	**26.66**	**0.952**	**0.036**	**92.40M**

pre-defined local windows. This fundamental mechanistic difference explains the quality of the derived semantic gate and the final performance gap. We conclude that only true global attention, as provided by ViT, supplies the necessary long-range context for high-fidelity, semantic-aware enhancement.

Analysis of Feature Space. To provide the most fundamental evidence for ViT's representational superiority, we delve deeper by visualizing the high-dimensional feature spaces of each backbone using t-SNE [32], a powerful non-linear dimensionality reduction technique. We first obtain a semantic segmentation mask for a sample image to serve as ground-truth labels for different semantic regions. We then project the feature vectors of sampled pixels into a 2D space, coloring each point by its semantic class.

The results, shown in Fig. 3, are clear and consistent. The feature space of our ViT-based model exhibits a remarkable degree of semantic clustering; pixels belonging to the same semantic class form tight, well-separated clusters. This demonstrates that ViT learns a feature representation that is inherently organized by the semantic meaning of the image content. In stark contrast, the feature spaces of Swin Transformer and ConvNeXt show significant feature entanglement, where different semantic regions are largely inseparable. This analysis provides strong evidence for our central claim: ViT's global attention mechanism empowers it to learn a feature space that is fundamentally better structured for semantic tasks, which is the root cause of its superior performance in our framework.



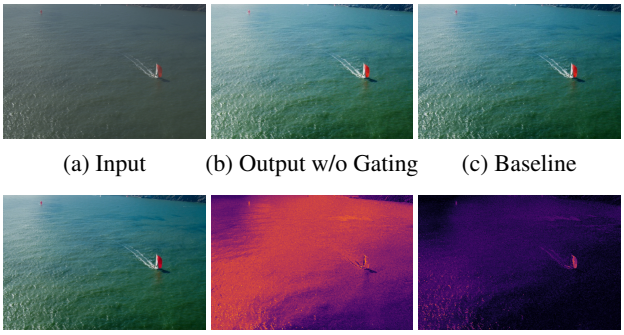
(a) Input (b) Ground Truth

(c) Gating Map from ViT (Ours) (d) Gating Map from Swin-T (e) Gating Map from ConvNeXt

Figure 4. Visualization of the semantic gating map generated using different backbones. The top row shows the input image and the ground truth for reference. The bottom row compares the gating maps from (c) our ViT-based model, (d) a Swin Transformer, and (e) a ConvNeXt. The ViT-based gate (c) exhibits large, coherent regions corresponding to holistic scene structures. In contrast, the gates from Swin-T (d) and ConvNeXt (e) are more fragmented or imprecise. This visually demonstrates ViT's superior ability to form a global scene understanding.

4.3.2. Effectiveness of Semantic Gating

To isolate the contribution of our proposed zero-cost semantic gate, we create a baseline variant of PrismNet where this mechanism is removed. In this ablated model, the ParamHead does not include the TwoStageSparse Attention mod-



(a) Input (b) Output w/o Gating (c) Baseline

(d) Ground Truth (e) Error Map of (b) (f) Error Map of (c)

Figure 5. Qualitative ablation study of our semantic gating mechanism on a challenging subject-background scene. (b) The model without gating applies adjustments globally, causing unnatural brightening and color spill on the background. (c) Our full model with semantic gating correctly confines the enhancement to the subject. The error maps (e, f), which visualize the L1 difference to the ground truth, further confirm that our gating mechanism significantly reduces artifacts in the background region.

ule, and the global adjustment residual is applied uniformly to the entire image.

Table 3. Ablation study on the effectiveness of our zero-cost semantic gating. Removing the gate leads to a consistent drop in performance across all metrics. While U-Net gating achieved similar quality, its significant memory and computational overhead rendered it impractical for processing 24MP images on a single GPU.

Model Variant	**PSNR** ↑	**SSIM** ↑	**LPIPS** ↓
PrismNet w/o Gating (Global Only)	26.43	0.950	0.038
PrismNet with U-Net Gating	26.62	0.951	0.036
PrismNet (Full Model)	**26.66**	**0.952**	**0.036**

Table 3 shows that the addition of our semantic gate brings a clear and consistent performance improvement across all metrics. While the numerical gain may seem modest, the qualitative results in Fig. 5 reveal the gate's true value. The baseline model struggles with scenes requiring careful subject-background separation, often causing adjustments to "leak" into unwanted areas, such as creating a halo effect around a brightened subject. Our full model with semantic gating successfully isolates the adjustments to the intended regions, demonstrating its crucial role in achieving the precise local control necessary for clean, professional results. We also explored an alternative using a dedicated U-Net to predict the gate; while it achieved similar quality, it was computationally prohibitive for high-resolution processing.

5. Conclusion

In this paper, we have argued and experimentally demonstrated that the quality of semantic representation provided by the backbone network is a critical, yet underexplored, factor in image enhancement. We revealed the significant superiority of a pre-trained Vision Transformer [9] in providing the necessary global context for this task. Building on this discovery, we proposed PrismNet, a novel framework that effectively leverages the power of ViT through a simple parametric head and a zero-cost semantic gate derived from attention scores. Our model achieves state-of-the-art performance on standard benchmarks, establishing a new quality ceiling for image enhancement. A key advantage of our parametric, feed-forward design is its high efficiency. On a single NVIDIA RTX 5090 GPU, PrismNet can process 4K-resolution images at a real-time speed of 404 FPS. While methods based on simpler architectures, such as 3D LUTs [47], can achieve even higher throughput, our approach provides a superior balance. It delivers a substantial leap in quality for a computational budget that remains well within the bounds of practical, real-time applications, defining a new, optimal point on the performance-efficiency spectrum. The success of PrismNet advocates for a shift in design focus for enhancement models: from engineering complex downstream networks to leveraging powerful, pre-trained semantic backbones. Future work could explore extending our framework to video enhancement or developing interactive applications based on its robust semantic-to-parameter mapping.

References

[1] Mahmoud Afifi and Michael S. Brown. Deep white-balance editing. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 1394–1403. Computer Vision Foundation / IEEE, 2020. 3

[2] Mahmoud Afifi, Konstantinos G Derpanis, Bjorn Ommer, and Michael S Brown. Learning multi-scale photo exposure correction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9157–9167, 2021. 1

[3] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input / output image pairs. In *The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011*, pages 97–104. IEEE Computer Society, 2011. 2, 5

[4] Yuanhao Cai, Hao Bian, Jing Lin, Haoqian Wang, Radu Timofte, and Yulun Zhang. Retinexformer: One-stage retinex-based transformer for low-light image enhancement. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 12470–12479. IEEE, 2023. 2, 3, 5, 6

[5] Qifeng Chen, Jia Xu, and Vladlen Koltun. Fast image processing with fully-convolutional networks. In *International Conference on Computer Vision (ICCV)*, pages 2497–2506, 2017. 1

[6] Yu-Sheng Chen, Yu-Ching Wang, Man-Hsin Kao, and Yung-Yu Chuang. Deep photo enhancer: Unpaired learning for image enhancement from photographs with gans. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 6306–6314. Computer Vision Foundation / IEEE Computer Society, 2018. 1, 6

[7] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen O. Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Trans. Image Process.*, 16(8):2080–2095, 2007. 1

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society, 2009. 3

[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 *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 2, 3, 8

[10] Robert Fergus, Barun Singh, Aaron Hertzmann, Sam T. Roweis, and William T. Freeman. Removing camera shake from a single photograph. *ACM Trans. Graph.*, 25(3):787–794, 2006. 1

[11] Michaël Gharbi, Jiawen Chen, Jonathan T. Barron, Samuel W. Hasinoff, and Frédo Durand. Deep bilateral learning for real-time image enhancement. *ACM Trans. Graph.*, 36(4):118:1–118:12, 2017. 1, 2, 6

[12] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2672–2680, 2014. 2

[13] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 1777–1786. Computer Vision Foundation / IEEE, 2020. 2

[14] Jingwen He, Yihao Liu, Yu Qiao, and Chao Dong. Conditional sequential modulation for efficient global image retouching. In *European Conference on Computer Vision (ECCV)*, pages 679–695. Springer, 2020. 1

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. 4

[16] 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. *CoRR*, abs/1704.04861, 2017. 4

[17] Yuanming Hu, Hao He, Chenxi Xu, Baoyuan Wang, and Stephen Lin. Exposure: A white-box photo post-processing framework. *ACM Trans. Graph.*, 37(2):26, 2018. 2

[18] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 448–456. JMLR.org, 2015. 3

[19] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5967–5976. IEEE Computer Society, 2017. 2

[20] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. Enlightengan: Deep light enhancement without paired supervision. *IEEE Trans. Image Process.*, 30:2340–2349, 2021. 2

[21] Yuma Kinoshita and Hitoshi Kiya. Convolutional neural networks considering local and global features for image enhancement. In *IEEE International Conference on Image Processing (ICIP)*, pages 2110–2114. IEEE, 2019. 1

[22] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment anything. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 3992–4003. IEEE, 2023. 7

[23] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pages 1106–1114, 2012. 5

[24] Jinxiu Liang, Yong Xu, Yuhui Quan, Jingwen Wang, Haibin Ling, and Hui Ji. Deep bilateral retinex for low-light image enhancement. *arXiv preprint arXiv:2007.02018*, 2020. 3

[25] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *IEEE/CVF International Conference on Computer Vision Workshops, ICCVW 2021, Montreal, QC, Canada, October 11-17, 2021*, pages 1833–1844. IEEE, 2021. 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 *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9992–10002. IEEE, 2021. 6, 7

[27] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the

2020s. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 11966–11976. IEEE, 2022. 6, 7

[28] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3431–3440. IEEE Computer Society, 2015. 1

[29] Kin Gwn Lore, Adedotun Akintayo, and Soumik Sarkar. Llnet: A deep autoencoder approach to natural low-light image enhancement. *Pattern Recognit.*, 61:650–662, 2017. 3

[30] Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with warm restarts. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. 5

[31] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. 5

[32] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (Nov):2579–2605, 2008. 7

[33] Sean Moran, Pierre Marza, Steven McDonagh, Sarah Parisot, and Gregory G. Slabaugh. Deeplpf: Deep local parametric filters for image enhancement. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 12823–12832. Computer Vision Foundation / IEEE, 2020. 6

[34] Sean Moran, Steven McDonagh, and Gregory G. Slabaugh. CURL: neural curve layers for global image enhancement. In *25th International Conference on Pattern Recognition, ICPR 2020, Virtual Event / Milan, Italy, January 10-15, 2021*, pages 9796–9803. IEEE, 2020. 2, 6

[35] Jongchan Park, Joon-Young Lee, Donggeun Yoo, and In So Kweon. Distort-and-recover: Color enhancement using deep reinforcement learning. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 5928–5936. Computer Vision Foundation / IEEE Computer Society, 2018. 2

[36] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8024–8035, 2019. 5

[37] 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 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III*, pages 234–241. Springer, 2015. 2

[38] Yuda Song, Hui Qian, and Xin Du. Starenhancer: Learning real-time and style-aware image enhancement. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 4106–4115. IEEE, 2021. 4, 5, 6

[39] Jianlin Su, Murtadha H. M. Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 4

[40] Li Tao, Chuang Zhu, Guoqing Xiang, Yuan Li, Huizhu Jia, and Xiaodong Xie. LLCNN: A convolutional neural network for low-light image enhancement. In *2017 IEEE Visual Communications and Image Processing, VCIP 2017, St. Petersburg, FL, USA, December 10-13, 2017*, pages 1–4. IEEE, 2017. 3

[41] Ruixing Wang, Qing Zhang, Chi-Wing Fu, Xiaoyong Shen, Wei-Shi Zheng, and Jiaya Jia. Underexposed photo enhancement using deep illumination estimation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6849–6857. Computer Vision Foundation / IEEE, 2019. 6

[42] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 17662–17672. IEEE, 2022. 2

[43] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, page 155. BMVA Press, 2018. 3

[44] Jingyang Yuan, Huazuo Gao, Damai Dai, Junyu Luo, Liang Zhao, Zhengyan Zhang, Zhenda Xie, Yuxing Wei, Lean Wang, Zhiping Xiao, Yuqing Wang, Chong Ruan, Ming Zhang, Wenfeng Liang, and Wangding Zeng. Native sparse attention: Hardware-aligned and natively trainable sparse attention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 23078–23097. Association for Computational Linguistics, 2025. 4

[45] 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 *European Conference on Computer Vision (ECCV)*, pages 492–511. Springer, 2020. 1

[46] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 5718–5729. IEEE, 2022. 2

[47] Hui Zeng, Jianrui Cai, Lida Li, Zisheng Cao, and Lei Zhang. Learning image-adaptive 3d lookup tables for high performance photo enhancement in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020. 2, 5, 6, 8

[48] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep CNN for image denoising. *IEEE Trans. Image Process.*, 26(7):3142–3155, 2017. 1

[49] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 586–595. Computer Vision Foundation / IEEE Computer Society, 2018. 5

[50] Yonghua Zhang, Jiawan Zhang, and Xiaojie Guo. Kindling the darkness: A practical low-light image enhancer. In *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*, pages 1632–1640. ACM, 2019. 3

[51] Zhaoyang Zhang, Yitong Jiang, Jun Jiang, Xiaogang Wang, Ping Luo, and Jinwei Gu. STAR: A structure-aware lightweight transformer for real-time image enhancement. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 4086–4095. IEEE, 2021. 2

[52] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2242–2251. IEEE Computer Society, 2017. 2

[53] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: deformable transformers for end-to-end object detection. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 4