Histogram Assisted Quality Aware Generative Model for Resolution Invariant NIR Image Colorization

Abhinav Attri*, Rajeev Ranjan Dwivedi*, Samiran Das, and Vinod Kumar Kurmi
Indian Institute of Science Education and Research Bhopal, India
{abhinav21, rajeev22, samiran, vinodkk}@iiserb.ac.in

Abstract

We present HAQAGen, a unified generative model for resolution-invariant NIR-to-RGB colorization that balances chromatic realism with structural fidelity. The proposed model introduces (i) a combined loss term aligning the global color statistics through differentiable histogram matching, perceptual image quality measure, and feature-based similarity to preserve texture information, (ii) local hue–saturation priors injected via Spatially Adaptive Denormalization (SPADE) to stabilize chromatic reconstruction, and (iii) texture-aware supervision within a Mamba backbone to preserve fine details. We introduce an adaptive-resolution inference engine that further enables high-resolution translation without sacrificing quality. Our proposed NIR-to-RGB translation model simultaneously enforces global color statistics and local chromatic consistency, while scaling to native resolutions without compromising texture fidelity or generalization. Extensive evaluations on FANVID, OMSIV, VCIP2020, and RGB2NIR using different evaluation metrics demonstrate consistent improvements over state-of-the-art baseline methods. HAQAGen produces images with sharper textures, natural colors, attaining significant gains as per perceptual metrics. These results position HAQAGen as a scalable and effective solution for NIR-to-RGB translation across diverse imaging scenarios.

1. Introduction

Near-Infrared (NIR) imaging unveils a hidden world beyond human perception, capturing important visual information beyond the visible region, recording image information from 780 nm to 1000 nm. This capability makes NIR imaging indispensable in domains such as surveillance, night vision applications [22], where it pierces through darkness; autonomous driving [13, 16], where it enhances visibility in adverse conditions [7]. Compared to visual images, NIR images substantially reduce the scattering of small and micro-scale particles present in smoke and fog, increasing the visibility. Despite the enormous potential, applications of NIR imaging systems are somewhat limited because the human visual system, trained to analyze visual information, cannot comprehend raw infrared images. The NIR-to-RGB translation approach bridges this gap by generating vivid, colorized images aligned with human perception. However, the process is fraught with challenges stemming from the spectral and geometric disparity between the NIR and RGB images, and the perceptual difference between these two modalities [14].

Current NIR-to-RGB translation approaches display several predicaments restricting their practical utility. Existing methods often suffer from textural information loss, geometric aberrations, color distortions, oversmoothing or blurring, and poor generalization. One pervasive issue is texture loss: many techniques generate outputs that lack the fine details present in the NIR input, resulting in blurred or oversmoothed images [5, 12, 19, 21]. Additionally, color distortions further complicate colorization, as the generated RGB images frequently display unnatural hues or inconsistent mappings [14, 24, 25], reducing their overall reliability. Moreover, the majority of existing models are *constrained by fixed input and output sizes*, rendering them inflexible for real-world applications where image dimensions vary widely. Besides, the absence of a sufficient number of diverse, paired NIR-RGB dataset prevent the study of the generalizability of the methods. Compounding these issues, evaluations are typically confined to a single dataset, casting doubt on the generalizability of these methods across diverse scenarios. Finally, computational inefficiency limits their deployment in time-sensitive contexts.

To address these challenges, we propose the *Histogram-Assisted Quality Aware Generative Model* **(HAQAGen)**, a unified translation framework that (i) recovers and preserves fine-grained texture via a texture-aware generation module, (ii) enables the model to produce vivid, natural-coloured images through histogram-based priors, (iii) generalises reliably across multiple datasets, and (iv) supports

*Equal contribution

adaptive-resolution inference for variable input sizes. Our framework unifies these elements into a single pipeline that achieves competitive perceptual quality and texture fidelity.

The remainder of this paper is organized as follows: Section 2 reviews prior work in NIR-to-RGB translation, Section 3 details proposed architecture and objectives, Sections 4 and 5 describe the experimental setup and results, and Section 6 concludes with limitations and future directions.

2. Related Work

NIR to RGB translation has received increasing attention in recent years, driven by its importance in various applications [10], low-light enhancement, remote sensing, and surveillance applications [10]. Early approaches relied on handcrafted features and classical regression models that attempted to directly map NIR pixel intensities to RGB values [39]. These conceptually simple methods lacked robustness to complex scene variations and failed to scale beyond narrow domains, highlighting the need for deep learning solutions. Recent deep learning driven approaches, such as generative and transformer-based architectures emerged as dominant paradigms. We summarize prior works considering three key aspects: spectral translation frameworks, texture preservation mechanisms, and evaluation practices in the subsequent sections:

GAN-based Models The first wave of deep approaches applied adversarial learning to capture the complex mapping between NIR and RGB. Mehri *et al*. [20] employed CycleGANs for unpaired spectral translation, enforcing cycle consistency and adversarial supervision to bridge the modality gap without paired data. [3] utilized a combination of different loss functions in their GAN model. Yan et al. [30] considered multi-scale features in their GAN model, while the work [27] generated three noisy versions of the same NIR scene to allow the GAN model to robustly learn the features. Dou et al. [8] introduced a cycle-GAN model that utilized distinct loss functions to improve robustness. While effective in principle, these models often suffer from unstable training and spectral ambiguity, leading to inconsistent colors.

Transformer-based Models Recently, researchers have turned to transformer-style models to exploit long-range dependencies to uncover the NIR-to-RGB mapping. The prominent *ColorMamba*[37] [37] model augments a state-space transformer backbone with learnable padding tokens, local convolutional modules, and agent-based attention. The model produces sharp boundaries and improved spectral fidelity. In this work, we adopt ColorMamba modules within our backbone but extend them with dual-branch supervision and histogram-based priors. Yang *et al*. [34] introduced a feature embedding strategy to better align statistical and semantic cues across modalities. The framework improves PSNR and structural similarity by embedding features at multiple resolutions. However, generalization across unseen domains remains a challenge for these models.

Texture Preservation in Colorization The prevalent NIR image colorization methods are unable to retain fine-grained texture information since NIR images lead to geometric distortions. Besides, colorized images are easily degraded by oversmoothing. Among the works attempting to resolve this issue, Li *et al*. [17] proposed a bi-stream texture-aware GAN that disentangles global structural cues from local details, fusing them to restore high-frequency components. Building on this, Yang *et al*. [32] introduced an attention-guided network with dedicated modules for semantic reasoning and texture transfer, combined through an adaptive fusion block. Although these advances highlight the necessity of texture retention, existing models are unable to preserve finer texture details and often lack scalability to diverse resolutions.

Evaluation and Generalization Conventional metrics, such as PSNR and SSIM, for quantifying fidelity of the generated RGB images, generally measure the pixelwise similarity, rather than the do not perceptual fidelity. To address this gap, Liu *et al*. [15] developed a deep image quality assessment framework that jointly considers texture, contrast, and color realism. Besides, most models are benchmarked on a single dataset, raising concerns about domain overfitting. Yang *et al*. [18] reported how validating on multiple datasets leads to improved robustness, underscoring the importance of cross-domain generalization.

Although generative models, particularly GANs and transformer models utilizing domain-dependent loss functions, have advanced the state-of-the-art, three fundamental challenges remain. (i) retaining fine-grained texture fidelity on par with high-frequency fusion networks, (ii) generating realistically coloured images for both local and global regions, and (iii) scaling seamlessly to arbitrary resolutions while ensuring cross-dataset generalization. Our proposed framework resolves these predicaments by unifying diverse loss terms, enforcing retention of texture information, perceptual image quality, and histogram alignment, and introducing an efficient, adaptive-resolution engine capable of translating images/patches of varying shapes. To our knowledge, this is the first NIR-to-RGB system that simultaneously enforces global color statistics and local chromatic consistency, while scaling to native resolutions without compromising texture fidelity or generalization.

3. Methodology

We envisage approaches for the translation of a single-channel NIR image $\mathbf{x}_{\text{nir}} \in \mathbb{R}^{H\times W\times 1}$ to a three-channel

RGB image $\hat{\mathbf{y}}_{\text{rgb}} \in \mathbb{R}^{H\times W\times 3}$. Given paired supervision $(\mathbf{x}_{\text{nir}}, \mathbf{y}_{\text{rgb}})$, we learn a mapping $\mathcal{F}_\Theta : \mathbf{x}_{\text{nir}} \mapsto \hat{\mathbf{y}}_{\text{rgb}}$ by minimizing a composite objective that balances (i) photometric and perceptual fidelity and (ii) chromatic realism under the inherent spectral ambiguity of NIR→RGB. We denote the ground-truth HSV (hue/saturation/value) as $\mathbf{y}_{\text{hsv}} = \Psi(\mathbf{y}_{\text{rgb}})$ and the model's auxiliary prediction as $\hat{\mathbf{y}}_{\text{hsv}}$. *Colour-space conventions.* Unless stated otherwise, images are in sRGB and linearly scaled to $[0, 1]$; HSV is computed from sRGB via $\Psi(\cdot)$ and used for auxiliary supervision and SPADE conditioning [23]. Losses that are defined "per channel" default to sRGB channels.

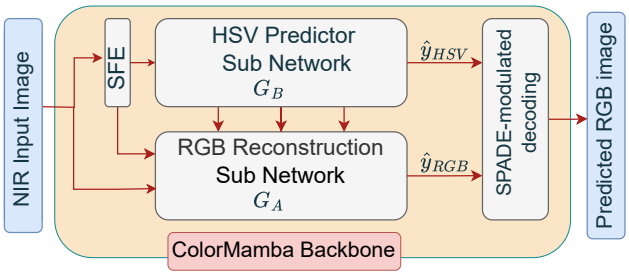

Figure 1. Proposed framework. NIR features feed two branches: an HSV Predictor and an RGB Reconstruction network. HSV guides the RGB decoder via SPADE [23], with dual discriminators and multi-term losses ensuring realism and consistency.

3.1. Backbone and Overall Design

Mamba-based encoder–decoder. We adopt *ColorMamba* [37] as the visual backbone for efficient representation learning while retaining long-range dependencies. Let $\mathcal{E}$ and $\mathcal{D}$ denote the shared encoder and decoder blocks, respectively.

Dual-branch generation with SPADE conditioning [23]. We introduce a dual generator $\mathcal{G} = \{G_A, G_B\}$: G_A is the *RGB branch* that predicts $\hat{\mathbf{y}}_{\text{rgb}}$, and G_B is an *HSV-prior branch* that regresses a dense hue–saturation–value field $\hat{\mathbf{y}}_{\text{hsv}} = G_B(\mathbf{x}_{\text{nir}})$ from the same input. To convey local chromatic priors into G_A, we inject $\hat{\mathbf{y}}_{\text{hsv}}$ into every decoder stage via SPADE-ResNet modulation [28]: for a decoder feature map $\mathbf{F} \in \mathbb{R}^{h\times w\times c}$ we apply

$$\widehat{\mathbf{F}} \;=\; \gamma(\hat{\mathbf{y}}_{\text{hsv}}) \odot \mathbf{F} \;+\; \beta_*(\hat{\mathbf{y}}_{\text{hsv}}), \tag{1}$$

where $\gamma(\cdot), \beta_*(\cdot)$ are lightweight convolution blocks, and $\odot$ denotes Hadamard product. Eq. (1) equips G_A with *region-aware colour cues* while the backbone supplies geometry and texture.

3.2. Learning Objective

Our objective comprises an adversarial tier that attempts to achieve naturalness in both colour spaces and a reconstruction tier that aligns texture, semantics, and global colour statistics.

Adversarial tier. Two PatchGAN discriminators $(D_{\text{RGB}}, D_{\text{HSV}})$ operate on $\hat{\mathbf{y}}_{\text{rgb}}$ and $\hat{\mathbf{y}}_{\text{hsv}}$ respectively, enforcing complementary constraints on luminance/chrominance. We use the *hinge* adversarial loss with 70×70 receptive fields, spectral normalization on D, and a 1:1 G:D update ratio. For $c\in\{\text{RGB}, \text{HSV}\}$,

$$\begin{aligned}\mathcal{L}_{\text{GAN}}^{G,c} &= -\,\mathbb{E}\big[D_c(\hat{\mathbf{y}}_c)\big], \\ \mathcal{L}_{\text{GAN}}^{D,c} &= \mathbb{E}\big[\max(0, 1 - D_c(\mathbf{y}_c))\big] + \mathbb{E}\big[\max(0, 1 + D_c(\hat{\mathbf{y}}_c))\big],\end{aligned} \tag{2}$$

and $\mathcal{L}_{\text{GAN}}^{G} = \sum_c \mathcal{L}_{\text{GAN}}^{G,c}$, $\mathcal{L}_{\text{GAN}}^{D} = \sum_c \mathcal{L}_{\text{GAN}}^{D,c}$. Supervising RGB and HSV with distinct critics makes hue failures detectable even when luminance appears plausible.

HAQAGen reconstruction tier. We regularize with a multi-purpose feature- and statistics-aware loss

$$\begin{aligned}\mathcal{L}_{\text{rec}}(\hat{\mathbf{y}}, \mathbf{y}) = &\underbrace{\alpha\,\|f(\hat{\mathbf{y}}) - f(\mathbf{y})\|_2^2 + \gamma\big[1 - \cos(f(\hat{\mathbf{y}}), f(\mathbf{y}))\big]}_{\text{task-specific texture basis (frozen autoencoder)}} \\ &+ \underbrace{\beta\,\|\,\text{CDF}(\hat{\mathbf{y}}) - \text{CDF}(\mathbf{y})\|_1}_{\text{global colour prior (differentiable CDF)}} \\ &+ \quad \underbrace{\delta\,\|g(\hat{\mathbf{y}}) - g(\mathbf{y})\|_2^2}_{\text{perceptual mid-level semantics (VGG-19)}}\end{aligned} \tag{3}$$

with $(\alpha, \beta, \gamma, \delta) = (1.0, 1.5, 1.0, 0.2)$. Here $f(\cdot)$ is a frozen four-layer autoencoder capturing task-specific textural features, and $g(\cdot)$ extracts `relu4_2` activations from VGG-19. The autoencoder terms stabilize high-frequency detail; the VGG term anchors semantic structure; the CDF term combats colour drift.

Differentiable histogram loss. Following [2], we compute a soft histogram $\mathbf{h} \in \mathbb{R}^B$ for each *sRGB* channel with temperature τ and bin centers $\{c_b\}_{b=1}^B$: $h_b = \frac{1}{N}\sum_{i=1}^N k_\tau(\hat{y}_i - c_b)$, where k_τ is a smooth kernel (e.g., triangular or logistic); the CDF is $\mathbf{H}$ with $H_b = \sum_{j\le b} h_j$. We set B=64 and τ=0.02 by default (see sensitivity in Sec. 5). We attempt to penalize the mismatch between output and target CDFs channel-wise and average across channels using ℓ_1 norm. The loss term yields stable, smooth gradients that align global chromatic statistics without distorting the local structure.

Full objective:

$$\begin{aligned}\mathcal{L}_G = \lambda_{\text{adv}}\,\mathcal{L}_{\text{GAN}}^{G} &+ \lambda_{\text{mse}}[\text{MSE}(\hat{\mathbf{y}}_{\text{rgb}}, \mathbf{y}_{\text{rgb}}) + \text{MSE}(\hat{\mathbf{y}}_{\text{hsv}}, \mathbf{y}_{\text{hsv}})] \\ &+ \lambda_{\text{feat}}[\mathcal{L}_{\text{rec}}(\hat{\mathbf{y}}_{\text{rgb}}, \mathbf{y}_{\text{rgb}}) + \mathcal{L}_{\text{rec}}(\hat{\mathbf{y}}_{\text{hsv}}, \mathbf{y}_{\text{hsv}})]\,, \\ \mathcal{L}_D = \mathcal{L}_{\text{GAN}}^{D}.&\end{aligned} \tag{4}$$

Texture-Aware Feature Enhancement Rather than relying solely on pixel losses, we capture *high-level, intermediate* representations that correlate with human sensitivity to

Algorithm 1 Dynamic Patching – Sliding-Window Inference

Require: NIR image $\mathbf{I}$; model M; patch size P; overlap O
Ensure: RGB image $\mathbf{O}$

$S \leftarrow P - O$ ▷ stride
Pad $\mathbf{I}$ to cover multiples of S
Compute grid $\{(y_i, x_i)\}_{i=1}^{N}$ of patch origins
Build feather mask $\mathbf{M} \in \mathbb{R}^{P \times P}$
Initialize accumulators $\mathbf{O}_{\text{pad}}, \mathbf{W}_{\text{pad}} \leftarrow 0$
for $i = 1 \dots N$ **do**
    Extract patch $\mathbf{p}_i \leftarrow \mathbf{I}[y_i : y_i + P,\ x_i : x_i + P]$
    Predict $\hat{\mathbf{y}}_i \leftarrow M(\mathbf{p}_i)$
    Add $\hat{\mathbf{y}}_i \odot \mathbf{M}$ into $\mathbf{O}_{\text{pad}}$
    Add $\mathbf{M}$ into $\mathbf{W}_{\text{pad}}$
end for
Normalize $\mathbf{O}_{\text{pad}} \leftarrow \mathbf{O}_{\text{pad}} \oslash \mathbf{W}_{\text{pad}}$
Crop to original size and **return** $\mathbf{O}$

edges and micro-texture. The feature similarity loss $\mathcal{L}_{\text{rec}}$ divides $(\hat{\mathbf{y}}, \mathbf{y})$ includes two complementary terms: (i) a frozen autoencoder $f(\cdot)$ that captures task-specific fine structure via ℓ_2 and cosine similarity; and (ii) a VGG-19 encoder $g(\cdot)$ capturing mid-level semantics via ℓ_2. All feature losses operate on 256×256 patches during training for stability, and over full-resolution outputs at inference for fidelity. *Autoencoder pretraining.* The texture autoencoder is trained *once* on the union of training splits (no test images) using an ℓ_2 reconstruction objective on 256×256 patches; after convergence, it is frozen and reused across all experiments to avoid dataset-specific leakage. Inputs are scaled to $[0, 1]$; Autoencoder feature vectors are instance-normalized before computing Eq. (3).

Global - Local Colour Guidance Since the NIR images generally lack chromatic cues, purely local criteria are insufficient. To resolve this issue, our model melds: (i) a *global* differentiable CDF loss (Eq. 3) that *aligns global colour statistics*, and (ii) *local* HSV priors injected through SPADE [23] (Eq. 1) so that decoder features are modulated by spatially varying hue/saturation hints. This combination ensures identical NIR intensities correspond to different materials (e.g., foliage vs. rock), yielding globally realistic and locally coherent colourization.

3.3. Adaptive-Resolution Inference

Since NIR images are generally high-resolution, naïve resizing to 256×256 introduces irreversible blur and texture loss. We therefore adopt a resolution-agnostic pipeline containing three components:

(i) Patch-based training. We optimize on overlapping 256×256 cropped patches to learn translation locally while regularizing with global terms (CDF/perceptual).

(ii) Sliding-window inference. At test time, we tile the image into overlapping patches of size $P = 256$ with stride $S \in \{222, 240\}$ (overlap $O = P-S$), process each patch independently with (G_A, G_B), and recompose. We additionally report a sensitivity sweep over S and a simple seam-energy diagnostic (gradient variance across patch borders) in Sec. 5.

(iii) Feather blending. Let $\mathbf{p}_i$ be the i-th output patch and $\mathbf{M} \in \mathbb{R}^{P\times P}$ a separable 2-D Hanning mask. We accumulate $\mathbf{O}_{\text{pad}} \mathrel{+}= \mathbf{p}_i \odot \mathbf{M}$ and the weights $\mathbf{W}_{\text{pad}} \mathrel{+}= \mathbf{M}$, then normalize $\mathbf{O} = \text{crop}\big(\mathbf{O}_{\text{pad}} \oslash \mathbf{W}_{\text{pad}}\big)$, where $\oslash$ is element-wise division. This eliminates seam artifacts and preserves edge continuity.

Content-aware downscaling (mitigating uniform-patch bias). Large scenes often over-sample local regions that do not contain detailed texture information (e.g., sky), creating bias in training. We therefore isotropically clamp the long side to ≤ 512 during training-time cropping (preserving aspect ratio), improving semantic diversity of patches and stabilizing optimization.

Network Heads and Discriminators HSV prior branch (G_B). A compact depthwise-separable CNN (≈4M params) regresses $\hat{\mathbf{y}}_{\text{hsv}}$. **Texture encoder $f(\cdot)$.** A frozen, lightweight 4-layer autoencoder (no skips) provides the texture basis in Eq. (3). **Dual critics.** D_{RGB} and D_{HSV} are 70×70 PatchGANs sharing all weights except the first conv layer to respect channel semantics. All images are treated as sRGB in $[0, 1]$; HSV targets are obtained via $\Psi(\cdot)$ from sRGB.

4. Experimental Setup & Implementation Details

Datasets We evaluate HAQAGEN on four public benchmarks spanning faces, urban/outdoor scenes, and mixed environments: FANVID [9], OMSIV [26], VCIP2020 [33], and RGB2NIR [4]. FANVID contains 5,144 paired VIS–NIR images (700–800 nm) at 2048×1536, emphasizing facial imagery and dynamic scenes. OMSIV contains 532 NIR–RGB pairs at 580×320 covering varied outdoor settings. VCIP2020 comprises 400 pairs at 256×256 across indoor/outdoor scenes. RGB2NIR includes 477 TIFF image pairs with variable resolution (up to 1024×768) over nine categories (countryside, field, forest, indoor, mountain, old buildings, streets, urban, water). Table 1 summarizes statistics and splits. Unless specified, we use official splits; otherwise, we adopt 80/10/10 train/val/test without scene overlap.

Training protocol. Unless stated otherwise, we train for 50 epochs on random 256^2 crops with AdamW (β_1=0.5,

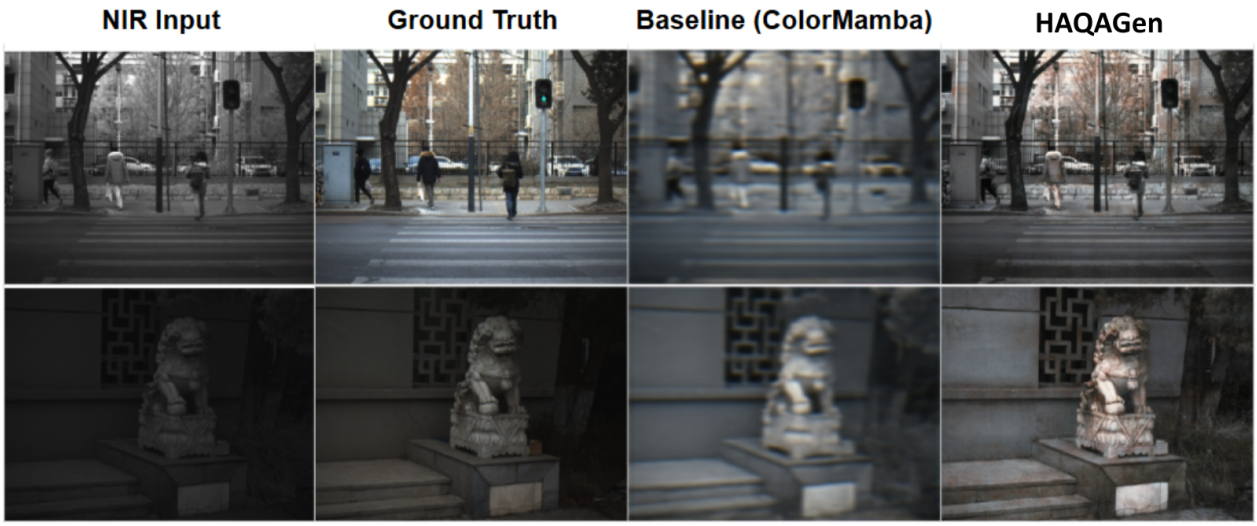

Figure 2. Comparison of FANVID dataset: (1) NIR input, (2) ground-truth RGB, (3) prediction with resizing (blurred), (4) prediction with adaptive resolution (sharper texture, better color).

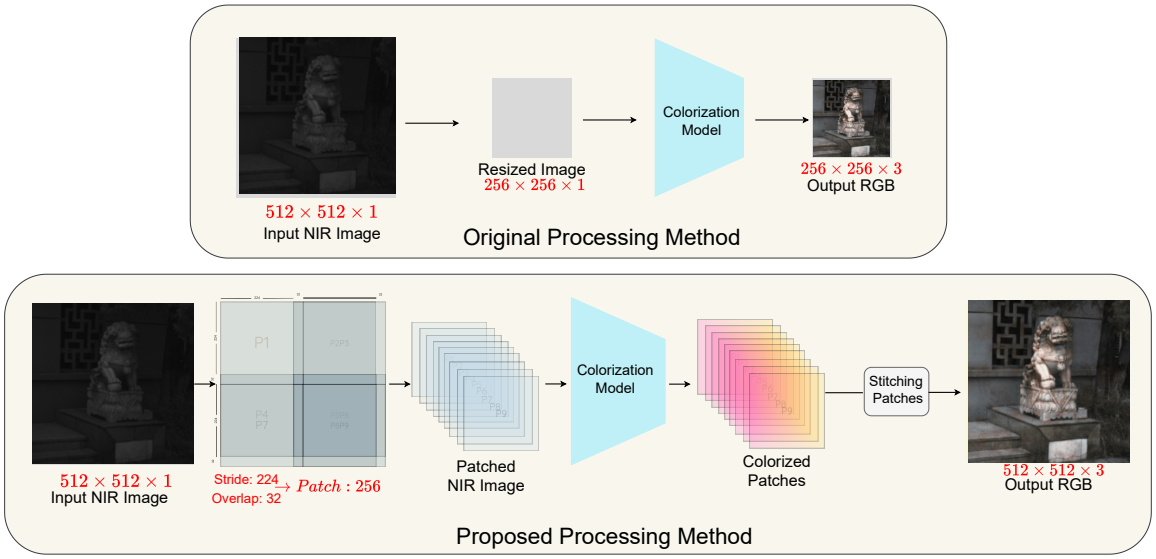

Figure 3. Adaptive patching: stride-based tiling, patch-wise colorization, and feathered stitching for seamless RGB output.

β_2=0.999, weight decay 10^{-4}) and cosine learning rate decay ($1 \times 10^{-4} \rightarrow 1 \times 10^{-6}$). Training is performed in mixed precision (AMP, fp16) with a global batch size of 16 across four RTX 4090 GPUs. The composite objective weights are set to $\{\lambda_{\text{MSE}}, \lambda_{\text{feat}}, \lambda_{\text{adv}}\}$= 15:15:1, balancing distortion, feature, and adversarial terms (see Sec. 3). All hyperparameters match the implementation described in the manuscript, and normalizations and colour-space conversions follow Sec. 4.

4.1. Preprocessing and Augmentation

We convert RGB images to *linear* sRGB, then perform min-max normalization to convert both modalities to $[0, 1]$. We also apply histogram equalization to the NIR channel to reduce illumination bias prior to training. Ground-truth RGB is converted to HSV (for colour-consistency checks), while predicted RGB is mapped to sRGB→LAB for perceptual losses. Unless stated otherwise, quantitative metrics are computed on linear sRGB.

Data augmentation. We employ random horizontal flips (p=0.5), 90° rotations, and HSV-saturation jitter (±10%) on the *reference* RGB only. For high-resolution datasets

Dataset	Type	#Pairs	Train / Val / Test	Modal Res.	Bit depth	Year
VCIP2020	indoor/outdoor	400	320 / 40 / 40	256×256	8	2020
FANVID	faces & urban	5144	4100 / 514 / 530	2048×1536	8	2024
OMSIV	outdoor	532	426 / 53 / 53	580×320	8	2017
RGB2NIR	mixed scenes	477	382 / 48 / 47	var. ($\leq 1024 \times 768$)	16	2011

Table 1. Dataset statistics and splits. Resolution reports the modal native size; "var." indicates multiple aspect ratios.

(FANVID, RGB2NIR), we additionally sample random 384×384 crops to encourage scale robustness prior to 256×256 patch formation.

Inference at Arbitrary Resolution We adopt sliding-window inference with feather blending (Sec. 3.3) to avoid loss of detailed information due to naïve resizing. We consider patch size P=256; stride $S \in \{222, 240\}$ (overlap 16–34 px); Hanning feather masks for seamless stitching; reflective padding for small borders. The approach enables resolution-agnostic testing with preserved textures and clean seams. Representative qualitative examples displayed in Figs. 4 and 5 underlines that HAQAGen not only preserves the detailed information, but also matches the color, and geometric information.

Evaluation Protocol and Metrics. We evaluate our model using four complementary metrics. Peak Signal-to-Noise Ratio (PSNR) [11] measures pixelwise fidelity as $\mathrm{PSNR} = 10 \cdot \log_{10}\left(\frac{MAX_I^2}{\mathrm{MSE}}\right)$ with $\mathrm{MSE} = \frac{1}{N}\sum_i (x_i - y_i)^2$, where MAX_I is the intensity range. Structural Similarity Index (SSIM) [11] captures luminance, contrast, and structural consistency via $\mathrm{SSIM}(x, y) = \frac{(2\mu_x\mu_y+C_1)(2\sigma_{xy}+C_2)}{(\mu_x^2+\mu_y^2+C_1)(\sigma_x^2+\sigma_y^2+C_2)}$, with μ, σ^2 denoting local means and variances, and σ_{xy} the covariance. Angular Error (AE) [1] quantifies chromatic accuracy as $\mathrm{AE}(p, g) = \cos^{-1}\left(\frac{p\cdot g}{\|p\|\|g\|}\right)$, measuring hue differences while remaining invariant to intensity scaling. Finally, Learned Perceptual Image Patch Similarity (LPIPS) [38] estimates perceptual distance by comparing deep features, $\mathrm{LPIPS}(x, y) = \sum_l \frac{1}{H_l W_l}\sum_{h,w} \| w_l \odot (\phi_l(x)_{hw} - \phi_l(y)_{hw}) \|_2^2$, where ϕ_l are pretrained features and w_l are learned channel weights. Together, PSNR and SSIM assess fidelity and structure, AE evaluates chromatic alignment, and LPIPS measures perceptual quality, holistically covering most perspectives holistically.

Baselines & Reproduction We benchmark against three categories of baselines: (i) *GAN-based* methods (e.g., CycleGAN [20]) trained with paired or unpaired NIR–RGB supervision; (ii) *Transformer/ state-space models* such as ColorMamba [37], which we adopt as our backbone; and (iii) *Texture-aware/attention models* (e.g., MPFNet [34], AttentionGAN [32]) that explicitly model texture or semantic priors.

For fairness, all baselines are retrained (when open-source code is available) using our unified preprocessing pipeline and training schedule, with input resizing handled consistently: 256^2 for VCIP2020/OMSIV, adaptive sliding-window for FANVID/RGB2NIR. When official pre-trained weights are used, we re-evaluate them under our metrics (PSNR, SSIM, AE, LPIPS) to ensure comparability. This harmonized protocol guarantees that reported gains stem from model design rather than differences in data processing or evaluation.

5. Results and Discussion

Table 2 benchmarks HAQAGen against twelve SOTA methods on VCIP2020. Our model achieves the best PSNR (24.96 dB) and the lowest LPIPS (0.18), while matching the top SSIM (0.71). Although AE is marginally higher than ColorMamba (2.96 vs. 2.81), visual inspection in Fig. 4 indicates that this trade-off correlates with richer chroma and sharper textures. Across the broader set of baselines, HAQAGen reduces AE by at least 23.3% (vs. SST) and LPIPS by 34.6% (vs. NIR-GNN), indicating strong perceptual fidelity.

Table 2. Quantitative results on VCIP2020. Best in **bold**.

Methods	PSNR(↑)	SSIM(↑)	AE(↓)	LPIPS(↓)
SST [30]	14.26	0.57	5.61	0.361
NIR-GNN [29]	17.50	0.60	5.22	0.384
MFF [30]	17.39	0.61	4.69	0.318
ATCGAN [31]	19.59	0.59	4.33	0.295
Restormer [35]	19.43	0.54	4.41	0.267
DRSformer [6]	20.18	0.56	4.22	0.254
MPFNet [34]	22.14	0.63	3.68	0.253
CoColor [33]	23.54	0.69	**2.68**	0.233
MCFNet [36]	20.34	0.61	3.79	0.208
ColorMamba [37]	24.56	**0.71**	2.81	0.212
HAQAGen	**24.96**	**0.71**	2.96	**0.18**

Cross-Dataset Generalization & Adaptive Resolution We study generalization across FANVID, OMSIV, VCIP2020, and RGB2NIR using fixed-size vs. adaptive sliding-window inference. Fig. 5 illustrates that patch-wise inference better preserves texture and tonal continuity on high-resolution imagery. Quantitatively (Table 3), adaptive inference delivers consistent LPIPS and AE gains on FANVID/OMSIV/RGB2NIR. On VCIP2020, where the target resolution matches the training crop, global resizing slightly

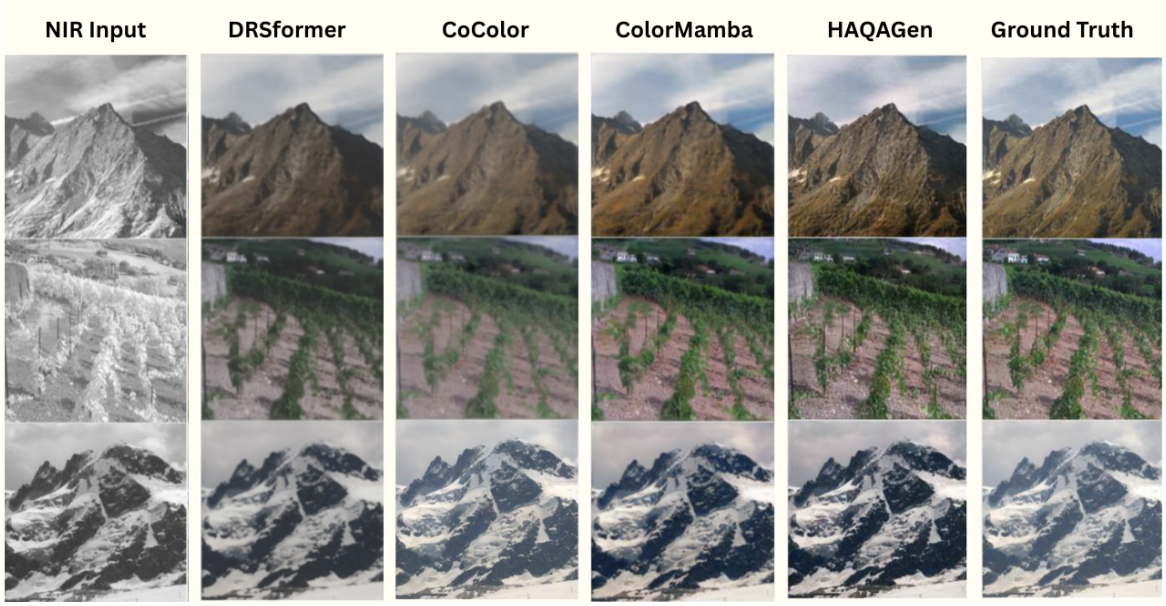

Figure 4. Qualitative comparison on the VCIP2020 dataset [33]: (1) NIR input, (2) DRSformer [6], (3) CoColor [33], (4) ColorMamba [37], (5) our proposed HAQAGen, and (6) ground-truth RGB. HAQAGen achieves sharper textures, more natural chromatic distributions, and better structural fidelity compared to prior baselines.

favours PSNR (consistent with reduced blending overhead), yet HAQAGEN still achieves the best LPIPS.

Table 3. Cross-dataset comparison of HAQAGEN vs. ColorMamba. Best per metric in **bold**.

Dataset	Method	PSNR ↑	SSIM ↑	AE ↓	LPIPS ↓
FANVID[9]	ColorMamba	17.63	0.65	26.79	0.64
FANVID[9]	HAQAGEN	**18.4**	**0.724**	**4.65**	**0.52**
OMSIV [26]	ColorMamba	**17.61**	0.58	25.87	0.52
OMSIV[26]	HAQAGEN	16.67	**0.61**	**6.90**	**0.37**
VCIP2020 [33]	ColorMamba	24.56	0.71	**2.81**	0.21
VCIP2020[33]	HAQAGEN	**24.96**	**0.71**	2.96	**0.18**
RGB2NIR [4]	ColorMamba	**17.22**	0.58	29.30	0.61
RGB2NIR [4]	HAQAGEN	15.97	**0.60**	**7.41**	**0.38**

Qualitative Analysis Fig. 4 contrasts ColorMamba with HAQAGEN (using $\mathcal{L}_{\text{rec}}$). We consistently observe:

1. **Texture Fidelity**: Fine details (foliage, contours, fabric) are better preserved, with reduced oversmoothing relative to adversarial-only baselines.
2. **Chromatic Realism**: The CDF prior curbs tinting and enforces natural tonal distributions across materials.
3. **Edge Consistency**: Boundaries at depth changes remain aligned after colourization, suggesting SPADE-conditioned decoding improves local hue assignment. Similar behaviour is visible at larger scales.

Ablation Studies Table 4 evaluates reconstruction-loss variants on VCIP2020. Replacing the baseline objective (MSE+Cosine+SSIM) with *VGG Perceptual* alone reduces PSNR by 1.12 dB and nearly doubles AE, indicating that perceptual loss without statistics/texture guidance is insufficient. *Histogram-only* narrows AE to 3.66 but lacks sharpness (SSIM 0.68). Our composite $\mathcal{L}_{\text{rec}}$ provides the best PSNR while balancing AE and SSIM, confirming the complementarity of texture features (AE, cosine) and global statistics (CDF). The extended ablation in the second table corroborates that removing either the CDF term or the AE-based texture supervision degrades colour or structure, respectively.

Loss Variant	**PSNR↑**	**SSIM↑**	**AE↓**
MSE + Cosine (ColorMamba)	24.56	0.71	2.81
+ VGG perceptual	23.63	0.70	4.32
+ Histogram only	23.81	0.68	3.66
+ Texture (f) only	24.12	0.69	3.01
Full L_{rec} (ours)	**24.96**	**0.71**	**2.96**

Table 4. Ablation on reconstruction losses (VCIP2020). Composite L_{rec} balances fidelity, color, and structure.

Table 5. Ablation on the HSV-SPADE branch. Removing HSV-SPADE conditioning degrades AE and SSIM, confirming that spatial hue priors improve local chromatic consistency.

Variant	**PSNR ↑**	**SSIM ↑**	**AE ↓**
Without HSV-SPADE branch	24.21	0.69	3.52
With HSV-SPADE (Ours)	**24.96**	**0.71**	**2.96**

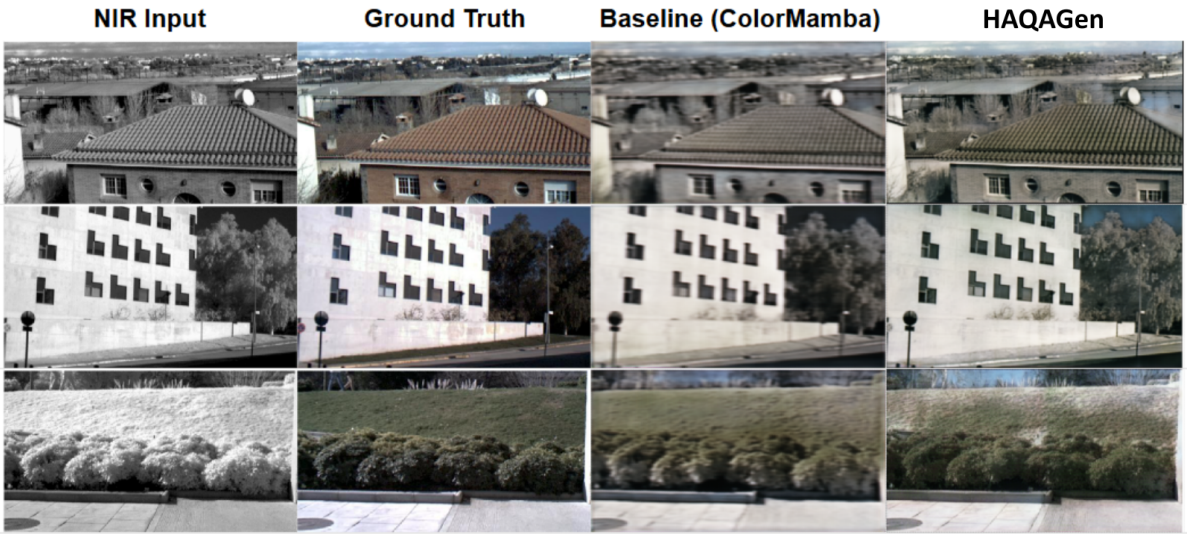

Figure 5. OMSIV [26]: Col. 1 NIR; Col. 2 GT; Col. 3 ColorMamba (resized); Col. 4 HAQAGEN (adaptive). Sliding-window inference preserves texture and tone continuity in high-resolution settings, outperforming global resizing.

Qualitative Gallery Resizing predictions to the original resolution introduces blur and geometric distortion, especially on FANVID and OMSIV. Adaptive patching avoids this by predicting at native scale, preserving edges and micro-texture.

Beyond quantitative metrics and ablation results, it is important to emphasize the broader implications of HAQAGen's performance. The improvements in perceptual quality (LPIPS), chromatic fidelity (AE), and structural preservation (SSIM) are not merely incremental gains but represent a step toward bridging the gap between synthetic translation and real-world usability. The qualitative comparisons (Figures 4–2) consistently show that HAQAGen avoids the common pitfalls of over-smoothing and spectral ambiguity that plague prior methods. Instead, it generates outputs that maintain edge sharpness and textural richness, qualities that directly impact downstream tasks such as detection and recognition. The robustness across datasets of varying resolution and content diversity further highlights the scalability of the framework. Collectively, these findings validate HAQAGen not only as a strong benchmark model for NIR-to-RGB translation but also as a practical tool for real-world deployment where both perceptual realism and structural fidelity are paramount.

6. Conclusion

In this paper, we introduced **HAQAGen**, a unified histogram-assisted framework that advances the frontier of NIR-to-RGB spectral translation. By jointly leveraging global colour statistics, HSV-based chromatic priors, and texture-aware feature supervision within a Mamba backbone, HAQAGen resolves the trade-off between *chromatic realism* and *textural fidelity*. Extensive experiments across four diverse benchmarks demonstrated noticeable improvement: quantitatively, HAQAGen achieves gains of up to **1.63 dB** in PSNR and **15.09%** improvement in LPIPS over state-of-the-art methods; qualitatively, it produces outputs with vivid colours, sharp structural boundaries, and reliable preservation of scene detail across varying scales and environments. Moreover, the adaptive-resolution inference engine ensures scalability to high-resolution imagery, enabling real-time deployment on commodity hardware without sacrificing quality.

Beyond numerical performance, our analyses highlight HAQAGen's practical impact. Its robustness across disparate datasets and strong compatibility with downstream tasks (e.g., object detection) indicate that NIR-to-RGB translation can evolve from a purely generative challenge to a foundation for actionable perception in adverse visual conditions.

Looking forward, several directions hold promise: (i) exploring *self-supervised colour priors* to reduce reliance on paired RGB supervision, (ii) distilling the dual-branch architecture into ultra-lightweight variants tailored for edge devices, and (iii) joint optimisation with higher-level tasks such as segmentation, tracking, and low-light enhancement to enable end-to-end NIR-aware vision systems.

We believe HAQAGen establishes a strong step toward practical and scalable NIR-to-RGB colourisation, laying the groundwork for next-generation perception in autonomous systems, security, remote sensing, and other human-centric applications where visibility is mission-critical.

References

[1] Dimitrios Androutsos, Konstantinos N Plataniotis, and Anastasios N Venetsanopoulos. A novel vector-based approach to color image retrieval using a vector angular-based distance measure. *Computer Vision and Image Understanding*, 75(1-2):46–58, 1999. 6

[2] Mor Avi-Aharon, Assaf Arbelle, and Tammy Riklin Raviv. Differentiable histogram loss functions for intensity-based image-to-image translation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):11642–11653, 2023. 3

[3] Kancharagunta Kishan Babu and Shiv Ram Dubey. Pcsgan: Perceptual cyclic-synthesized generative adversarial networks for thermal and nir to visible image transformation. *Neurocomputing*, 413:41–50, 2020. 2

[4] M. Brown and S. Süsstrunk. Multispectral SIFT for scene category recognition. In *Computer Vision and Pattern Recognition (CVPR11)*, pages 177–184, Colorado Springs, 2011. 4, 7

[5] Qidong Chen, Xiao Liu, Lili Du, Bo Song, Xiaobing Sun, and Zhengyu Shen. Nir-to-rgb image colorization based on a conditional gan. *Appl. Opt.*, 64(11): 2968–2978, 2025. 1

[6] Xiang Chen, Hao Li, Mingqiang Li, and Jinshan Pan. Learning a sparse transformer network for effective image deraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5896–5905, 2023. DRSformer. 6, 7

[7] Lark Kwon Choi, Jaehee You, and Alan Conrad Bovik. Referenceless prediction of perceptual fog density and perceptual image defogging. *IEEE Transactions on Image Processing*, 24(11):3888–3901, 2015. 1

[8] Hao Dou, Chen Chen, Xiyuan Hu, and Silong Peng. Asymmetric cyclegan for unpaired nir-to-rgb face image translation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1757–1761. IEEE, 2019. 2

[9] Yunyi Gao, Lin Gu, Qiankun Liu, and Ying Fu. Object-aware nir-to-visible translation. In *European Conference on Computer Vision*, pages 93–109. Springer, 2024. 4, 7

[10] Duncan L Hickman. Colour fusion of rgb and nir imagery for surveillance applications. In *Electro-Optical and Infrared Systems: Technology and Applications XVII*, pages 105–124. SPIE, 2020. 2

[11] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010. 6

[12] Cheolkon Jung, Qihui Han, Kailong Zhou, and Yuanquan Xu. Multispectral fusion of rgb and nir images using weighted least squares and convolution neural networks. *IEEE Open Journal of Signal Processing*, 2:559–570, 2021. 1

[13] Chongyi Li, Chunle Guo, Linghao Han, Jun Jiang, Ming-Ming Cheng, Jinwei Gu, and Chen Change Loy. Low-light image and video enhancement using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9396–9416, 2021. 1

[14] Wei Liang, Derui Ding, and Guoliang Wei. An improved dualgan for near-infrared image colorization. *Infrared Physics & Technology*, 116:103764, 2021. 1

[15] Bohua Liu, Jianli Ding, Jie Zou, Jinjie Wang, and Shuai Huang. Ldanet: A lightweight dynamic addition network for rural road extraction from remote sensing images. *Remote Sensing*, 15(7), 2023. 2

[16] Jiaying Liu, Dejia Xu, Wenhan Yang, Minhao Fan, and Haofeng Huang. Benchmarking low-light image enhancement and beyond. *International Journal of Computer Vision*, 129:1153–1184, 2021. 1

[17] Weirong Liu, Chengrui Cao, Jie Liu, Chenwen Ren, Yulin Wei, and Honglin Guo. Fine-grained image inpainting with scale-enhanced generative adversarial network. *Pattern Recognition Letters*, 143:81–87, 2021. 2

[18] Chuang Ma, Shaokai Zhao, Yu Pei, Liang Xie, Erwei Yin, and Ye Yan. A multi-prior fusion network for video-based micro-expression recognition. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2025. 2

[19] Xiaoyu Ma, Wei Huang, Rui Huang, and Xuefeng Liu. Near-infrared image colorization using asymmetric codec and pixel-level fusion. *Applied Sciences*, 12(19), 2022. 1

[20] Armin Mehri and Angel D Sappa. Colorizing near infrared images through a cyclic adversarial approach of unpaired samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2, 6

[21] Lin Mei and Cheolkon Jung. Deep fusion of rgb and nir paired images using convolutional neural networks. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 6802–6803, 2021. 1

[22] Muyao Niu, Zhihang Zhong, and Yinqiang Zheng. Nir-assisted video enhancement via unpaired 24-hour data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10778–10788, 2023. 1

[23] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2332–2341, 2019. 3, 4

[24] Chao Qu, Shuo Zhu, Yuhang Wang, Zongze Wu, Xiaoyu Chen, Edmund Y Lam, and Jing Han. Near-infrared image deblurring and event denoising with synergistic neuromorphic imaging. *arXiv preprint arXiv:2503.01193*, 2025. 1

[25] Chang-Hwan Son and Xiao-Ping Zhang. Near-infrared fusion via color regularization for haze and color distortion removals. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(11):3111–3126, 2017. 1

[26] X. Soria, A. D. Sappa, and A. Akbarinia. Multispectral single-sensor rgb-nir imaging: New challenges and opportunities. In *2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6. IEEE, 2017. 4, 7, 8

[27] Patricia L Suárez, Angel D Sappa, and Boris X Vintimilla. Infrared image colorization based on a triplet dcgan architecture. In *Proceedings of the IEEE Conference on Com-*

puter Vision and Pattern Recognition Workshops, pages 18–23, 2017. 2

[28] Tian Sun and Cheolkon Jung. Nir image colorization using spade generator and grayscale approximated self-reconstruction. In *2020 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, pages 463–466, 2020. 3

[29] Diego Valsesia, Giulia Fracastoro, and Enrico Magli. Nir image colorization with graph-convolutional neural networks. In *IEEE VCIP*, pages 451–454, 2020. NIR-GNN. 6

[30] Longbin Yan, Xiuheng Wang, Min Zhao, Shumin Liu, and Jie Chen. A multi-model fusion framework for nir-to-rgb translation. In *IEEE VCIP*, pages 459–462, 2020. MFF (same primary paper as SST, but referencing MFF approach). 2, 6

[31] Xingxing Yang and Zhenghua Chen. Learning from paired and unpaired data: Alternately trained CycleGAN for near infrared image colorization. In *2020 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, pages 467–470. IEEE, 2020. CycleGAN for NIR colorization. 6

[32] Xingxing Yang, Jie Chen, Zaifeng Yang, and Zhenghua Chen. Attention-guided nir image colorization via adaptive fusion of semantic and texture clues. *arXiv preprint arXiv:2107.09237*, 2021. 2, 6

[33] Xingxing Yang, Jie Chen, and Zaifeng Yang. Cooperative colorization: Exploring latent cross-domain priors for nir image spectrum translation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 2409–2417, 2023. CoColor. 4, 6, 7

[34] Xingxing Yang, Jie Chen, and Zaifeng Yang. Multi-scale progressive feature embedding for accurate nir-to-rgb spectral domain translation. In *2023 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, pages 1–5, 2023. MPFNet. 2, 6

[35] 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*, pages 5728–5739, 2022. Restormer. 6

[36] Huiyu Zhai, Mo Chen, Xingxing Yang, and Gusheng Kang. Multi-scale hsv color feature embedding for high-fidelity nir-to-rgb spectrum translation. arXiv preprint arXiv:2404.16685, 2024. MCFNet. 6

[37] Huiyu Zhai, Guang Jin, Xingxing Yang, and Guosheng Kang. Colormamba: Towards high-quality nir-to-rgb spectral translation with mamba. *arXiv preprint arXiv:2408.08087*, 2024. 2, 3, 6, 7

[38] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6

[39] Wenwen Zhang, Liyanaarachchi Chamara Kasun, Qi Jie Wang, Yuanjin Zheng, and Zhiping Lin. A review of machine learning for near-infrared spectroscopy. *Sensors*, 22 (24):9764, 2022. 2