Hist2Style: Histogram-Guided Stylization with Bilateral Grids

Dekel Galor[1,2,†] Adam Pikielny[1] Zhoutong Zhang[1] Ke Wang[1]
Laura Waller[2] Jiawen Chen[1] Ilya Chugunov[1]

[1]Adobe Nextcam [2]University of California, Berkeley

Abstract

Photorealistic style transfer aims to match the color and tone of an input image to that of a style target while preserving the content and details of the original scene. Although existing large image models can facilitate these kinds of appearance edits, their high computational demands, potential for hallucinations, and limited user control make them unsuitable for high-resolution, real-time workflows. We introduce Hist2Style, a bilateral-grid formulation for fast, edge-aware stylization that preserves visual fidelity by constraining operations to locally affine transforms in bilateral space. Our model distills a large image editing model into a lightweight network by training on a large supervised corpus generated with language and vision-language models, targeting spatially varying color edits. The network conditions on a histogram-based embedding of the style target to provide an interpretable interface for adjusting the output style by modifying the target color distribution. Overall, Hist2Style maintains content structure by construction, avoids hallucinations, and supports real-time, high-resolution photorealistic stylization with interactive user-controllable color and tone adjustments. Our project page is available at dgalor.github.io/hist2style/.

1. Introduction

Color and tone define much of an image's mood, coherence, and narrative [12], and photographers and filmmakers have long relied on color grading to evoke emotion and maintain visual consistency across scenes [12, 20, 55]. Yet this process remains painstakingly manual, requiring individual adjustment of lighting, contrast, and palette across images and video clips [20]. Automating such color grading while preserving realism motivates the study of style transfer: the ability to borrow the visual *look* of one image and apply it to another without altering its content [17, 24, 32].

While the goal of style transfer is conceptually simple, what constitutes *style* and *content* varies considerably across methods. In this work, we adopt the photorealistic convention: style is defined by color and tone, content by structure and edges [32, 45]. Many existing approaches instead rely on deep neural features that, while expressive, are difficult to interpret or control [24], leaving users with little flexibility to achieve their intended look [32].

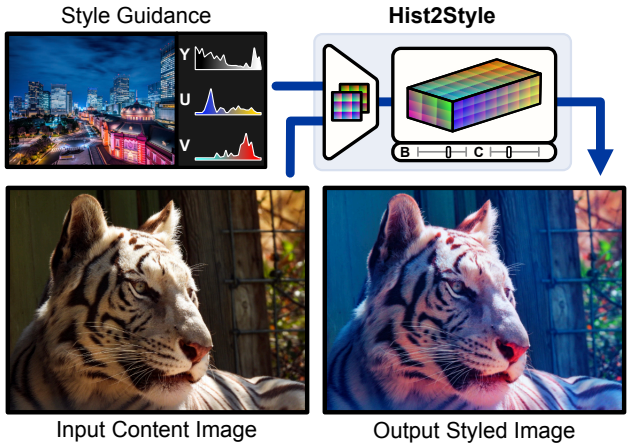

Figure 1. Our method preserves content and detail by mapping nearby pixels of similar color to similar outputs via a lightweight network conditioned on a histogram-based style embedding. This enables fast, high-resolution stylization, and the histogram itself is directly editable, giving users intuitive control over color and tone.

Recently, large image editing models have transformed how users approach stylization [21, 47, 52]. These models let users specify visual intentions through image or text prompts, promising far greater flexibility than traditional style transfer methods [21]. In principle, their expressive power could render specialized photorealistic stylization algorithms obsolete, but in practice this versatility introduces key limitations. *Performance:* To support diverse editing tasks, these models sacrifice efficiency for generality, demanding high computation, large memory, and long inference times [59]. *Hallucination:* Edits can introduce identity drift, structural distortions, and other artifacts that break photorealism [33]. *Controllability:* Expressing subtle color

[†]Work partially done during an internship at Adobe.

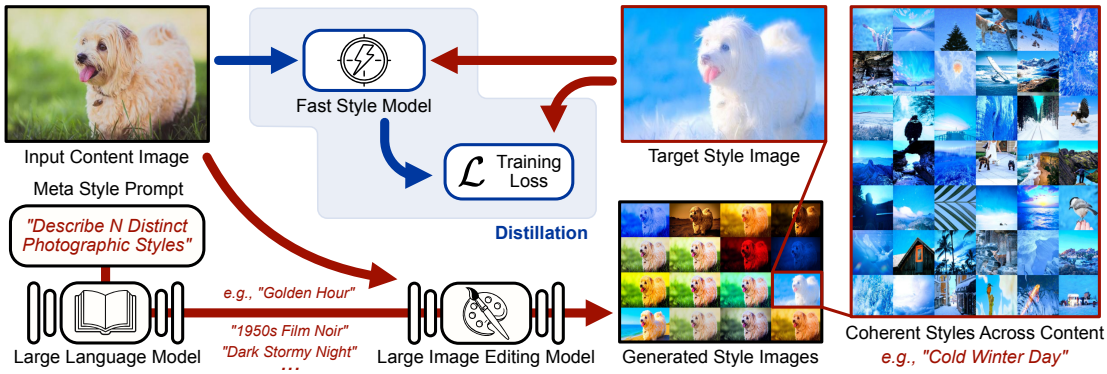

Figure 2. **Selective distillation** to condense the capabilities of a large image editing model into a lightweight, specialized network for photorealistic stylization. We procedurally generate style editing prompts with a large language model. The prompts are used to instruct an editing model to edit standard images into different style variations, which are coherent across images. Our model is then trained to mimic the stylization of the large editing model with a regression loss (*red:* precomputed, *blue:* live training).

and tone through prompts introduces ambiguities that hinder stylistic precision [8].

We address these limitations by selectively distilling a large image editing model into an efficient sub-model specialized for photorealistic stylization. To improve *performance*, the distillation compresses a foundation model into a lightweight task-specific network. To suppress *hallucinations*, we constrain edits to locally affine transformations in bilateral space. To enhance *controllability*, we introduce a histogram-based style embedding that enables intuitive, interactive manipulation of color and tone. Named Hist2Style, our model foregoes content generation to focus strictly on photorealism: precisely adjusting color and tone without distorting the image structure.

Our key contributions are:

1. A photorealistic, histogram-guided stylization network robust to hallucinations and scalable to high resolutions.
2. An interactive system for real-time image editing with artistically expressive, histogram-based control.
3. A stylization quality assessment metric for automated and reproducible evaluation of stylization models that correlates highly with human preference.

2. Related Works

The field of image style transfer has evolved through several distinct paradigms, each defining *style* and *content* differently and balancing trade-offs between semantic richness, photorealism, and user control [24, 59].

Classical Color Transfer Early methods of image stylization focused on transferring the global color distribution, adjusting a photograph's overall color statistics to match a reference [39–43, 45]. Approaches from the early 2000s such as Reinhard *et al*. [45] matched mean and variance of color channels between images, while Pitié *et al*. [39–43] introduced more advanced methods such as *iterative distribution transfer* (IDT) [42] to align full histograms. These techniques could be categorized as *photorealistic* since they preserve the original spatial structure [39–43, 45]. However, as these color mappings are calculated from and applied to global color statistics, they lack spatial or semantic awareness [19]. This property often leads to visible artifacts. For example, [39] reports "graininess" when source and target images differ greatly and requires postprocessing to suppress. Our work shares the spirit of these classical color transfer methods, but extends them with semantic and spatial awareness.

Neural Style Transfer A significant change in the field was marked by the rise of neural networks, which redefined style transfer [24]. Gatys *et al*. introduced the seminal method for artistic style transfer, defining style not as a global color profile but as the second-order statistics (Gram matrices) of deep features from a VGG network [17, 48]. This unlocked unprecedented stylistic richness but was inherently "artistic" and non-photorealistic [17, 32]. Subsequent works like Universal Style Transfer (WCT) accelerated this process to interactive speeds, but remained focused on artistic application, whereas our objective is photorealistic stylization [22, 29]. Achieving photorealistic style transfer became a major focus around 2017 [32, 59]. Luan *et al*. introduced Deep Photo Style Transfer, which added a photorealism regularizer and used semantic segmentation to constrain Gatys' style loss [17, 32]. This avoided the "painting" distortions and kept stylized results plausible, but the method required slow per-image optimization [32]. Li *et al*. proposed PhotoWCT, which applied a similar constraint but as a postprocessing step for WCT [29, 30]. Yoo *et al*. alleviated the need for postprocessing with WCT^2, a wavelet-based stylization network that preserves image structures [58]. Afifi *et al*. applied histogram style conditioning and guided upsampling to re-

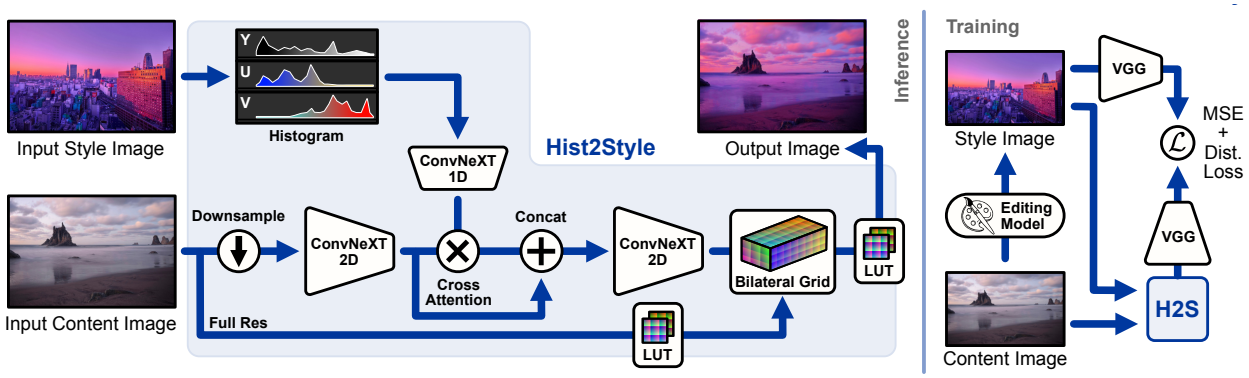

Figure 3. **Model architecture (left)**. Our model separately embeds a downsampled content image and style color histogram via ConvNeXt blocks. The two are then fused with cross attention, and fed through the output head to produce a spatially adaptive color transform known as the bilateral grid. The bilateral grid is applied to the content image and fed through a learned per-channel nonlinearity (LUT) to produce a stylized image. **Training (right)** is done by first creating synthetic ground-truth stylization pairs using a large image-editing model. Then the model is fed a content image, the histogram of the ground-truth image, and trained to regress to the ground-truth with a perceptual loss.

duce artifacts with ReHistoGAN [13]. Another significant breakthrough came from Xia *et al*. [57], which reframed the soft photorealism regularization as learning to predict the parameters of local affine color transforms known as bilateral grids [18, 57]. Subsequent works continued to improve efficiency and fidelity. PhotoWCT^2 achieved more efficient inference while maintaining the quality of PhotoWCT [11]. D-LUT employed a diffusion process for super-efficient color distribution transfer, but constrained to global adjustments [28]. SA-LUT then introduced spatially varying LUTs via quadrilinear interpolation [19]. These methods all share a strategy of constraining the transform space to prevent the free-form distortions of artistic style transfer, thereby producing outputs that could pass as real photographs [59]. Such constraints appear in various forms, including wavelet filters [11, 58], local color maps [19, 32, 57], and global color maps [10, 25, 28]. We opt for using the bilateral grid, a local color map used in [57], since it is efficient, scalable, and provides a balance between expressivity and photorealistic constraints [3, 5, 6, 18, 53, 57].

Foundation Models for Image Editing A new paradigm has since emerged with the rise of foundational image editing models [8, 27, 33, 47, 52, 56]. Scaling trends in diffusion and transformer architectures have enabled general-purpose systems capable of performing a wide range of edits [8, 33], from recoloring [52] and relighting [52] to object replacement [52], all within a single unified model [52]. Open-source models such as Flux Kontext [27] and Qwen Image Edit [56] exemplify this shift toward universal editing [52]. These systems achieve remarkable flexibility, reframing visual manipulation as a form of conditional generation rather than task-specific transformation [47]. Yet this generality introduces new challenges: high computational demands [59], unpredictable hallucinations [33], and limited user control over precise outcomes [8].

In this work, we leverage the semantic priors and expressivity of image editing models while avoiding the hallucination [33] and performance [59] issues by sacrificing generality (adding photorealistic constraints) and distilling into a lightweight model. Additionally, we tackle the controllability [8] issue by building an interactive mechanism for matching results to user intent.

3. Hist2Style

We train a lightweight neural network that maps a content image and a style embedding to a stylized image as illustrated in Fig. 3. We define the style embedding as the marginal (per channel) histogram of the style image, providing interpretability and controllability (discussed further in Sec. 3.3 and Sec. 3.6).

3.1. Selective Distillation of Image Editing Models

Our network is trained to distill a subset of the capabilities of a large image editing model [27]. To this end, we first prompt a large language model [49] to generate a diverse set of names and descriptions of photorealistic styles such as “Golden Hour”, “1950s Film Noir”, etc. Then we take a standard photography dataset [50] and prompt the image editing model [27] with the LLM-generated descriptions to create many stylized versions of each image in the dataset. As discussed in Sec. 1, large image editing models struggle with hallucinations [33] and controllability issues [8], which affect some images generated with the teacher model. Other than automatically filtering data (described in supplementary material), we propose the use of bilateral grids to bake photorealistic constraints into the model itself [18, 57].

3.2. Bilateral Grids for Photorealistic Stylization

The bilateral grid is a parameterization of image-to-image functions that explicitly encodes the edges of 2D images via a guidance dimension. By storing affine transformations at each grid cell, the bilateral grid can compactly represent locally affine edge-preserving functions, a desirable property for photorealistic stylization [32, 57]. The bilateral grid decouples the resolution of the transform from that of the image: computing the transform is done by *slicing* into the grid using multilinear interpolation followed by a per-pixel operation. This lets our method scale to arbitrarily high resolutions.

In this work, we adopt the strategy from prior methods and define the guidance dimension of the bilateral grid as a learned function of RGB $g : \mathbb{R}^3 \to \mathbb{R}$ [18, 57]. This effectively acts as a luminance dimension that prevents smoothing across edges. For content image I_c with shape $(H, W, 3)$, the model predicts a bilateral grid of size (G_g, G_h, G_w), corresponding to the grid resolution in the guidance dimension, height, and width, respectively. We use a grid size of $(8, 16, 16)$, which proved effective in [57]. For each grid entry, the model predicts an affine transform, which corresponds to $3 \times 4 = 12$ scalar values. Additionally, the model predicts a premultiplied α value per grid entry as a measure of model uncertainty. The result of slicing the bilateral grid via trilinear interpolation is a per-pixel affine transform, which is applied to the content image to produce the stylized output.

3.3. Histogram-Guided Style Transfer

Stylization is often encoded via features of pretrained image models [22, 32, 57] or language prompt embeddings [52]. However, these representations lack two critical properties: *interpretability*, as disentangling style from content within a neural network is non-trivial [44]; and *controllability*, since matching network embeddings to a user's precise intent remains challenging [8, 44].

To address the limitations in interpretability and control, we represent style using the marginal color histogram of the style image. This choice is motivated by the histogram's foundational role in traditional photo editing [4]. Unlike classic methods, however, our goal is not pure distribution transfer, which can introduce artifacts by disregarding spatial structure [42]. Instead, we leverage synthetic training data, and jointly optimize spatial and distribution objectives so that the model learns both.

We use MSE as a spatial loss between the output and ground truth images. For our distribution loss, we estimate the squared one-dimensional Wasserstein-2 metric by sorting pixels across space, and computing the MSE [38] (see Algorithm 1). We find that applying these losses in perceptual space (VGG [48]) yields better generalization than applying them in image color space, an observation we attribute to imperfections in the synthetic dataset.

Algorithm 1 Marginal distribution matching loss.

Input: flattened image features $X, Y \in \mathbb{R}^{HW \times C}$
Output: 1D Wasserstein loss L
for each channel $c \in \{1, ..., C\}$ **do**
  $X_c = X[:, c], Y_c = Y[:, c] \in \mathbb{R}^{HW}$
  Sort: $x_c \leftarrow \text{sort}(X_c), \quad y_c \leftarrow \text{sort}(Y_c)$
  Compute squared distance: $d_c = \|x_c - y_c\|^2$
end for
Return: mean loss $L = \frac{1}{C} \sum_{c=1}^{C} d_c$

3.4. Model Architecture

As illustrated in Fig. 3, our model follows a dual-branch design that processes the content image and the style embedding separately, then fuses them to predict a spatially adaptive color transform represented as a bilateral grid [18, 57].

The content branch is a convolutional encoder with ConvNeXt-style blocks introduced by Liu *et al*. [31]. Each block includes large-kernel depthwise convolutions (7×7) and inverted bottleneck layers with GELU activations and LayerNorm. The content encoder downsamples the image spatially to match the intended bilateral grid size, 16×16 in our experiments. In parallel, the style branch encodes the source and target color histogram. We treat the 1D histograms as a sequence, one per color channel, and apply an analogous ConvNeXt-structured 1D CNN [31]. This yields a global style token representing the style editing operation.

To inject the global style editing information into the local content features, we employ a cross-attention module [51]. We then concatenate the content image features with the results of this module, and feed the combined features into a small convolutional output head that predicts an affine bilateral grid [57].

In addition to the affine coefficients of the grid, the model predicts an uncertainty channel α with a softplus nonlinearity for stability. Next, the affine coefficients get multiplied by α in-place, while the α channel itself remains untouched. The content image is used to *slice* into the grid using trilinear interpolation, resulting in per-pixel affine transform coefficients which are divided by the final *sliced* α values. Before and after applying the affine transforms, we apply an additional per-channel nonlinearity, parameterized as smooth monotonic functions.

3.5. Implementation

We implement all models in PyTorch [36] with PyTorch Lightning [7, 16]. Optimization is performed with Adam [26] using a learning rate of 3×10^{-4}, $(\beta_1, \beta_2) = (0.9, 0.99)$, together with a linear warm-up schedule of one epoch. The model has 1.5M parameters, and was trained

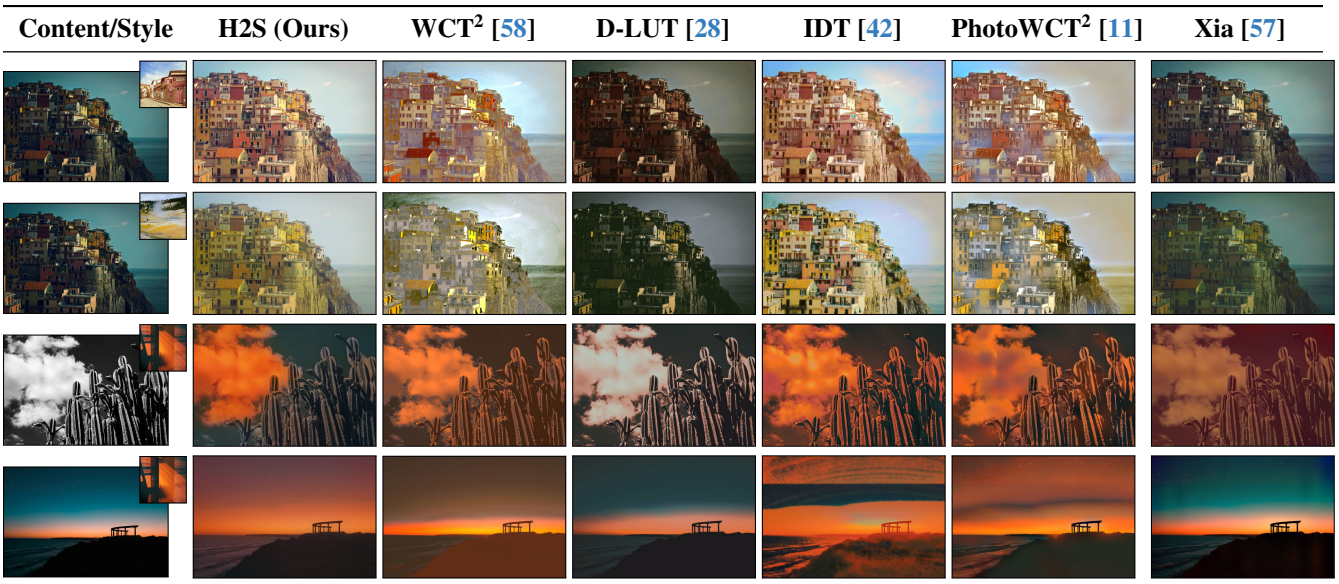

Figure 4. Qualitative comparisons with baseline methods on images from the user-study evaluation set (Tab. 1), illustrating how Hist2Style produces compelling, spatially adaptive stylizations that avoid high-frequency edge and color artifacts (best viewed zoomed in).

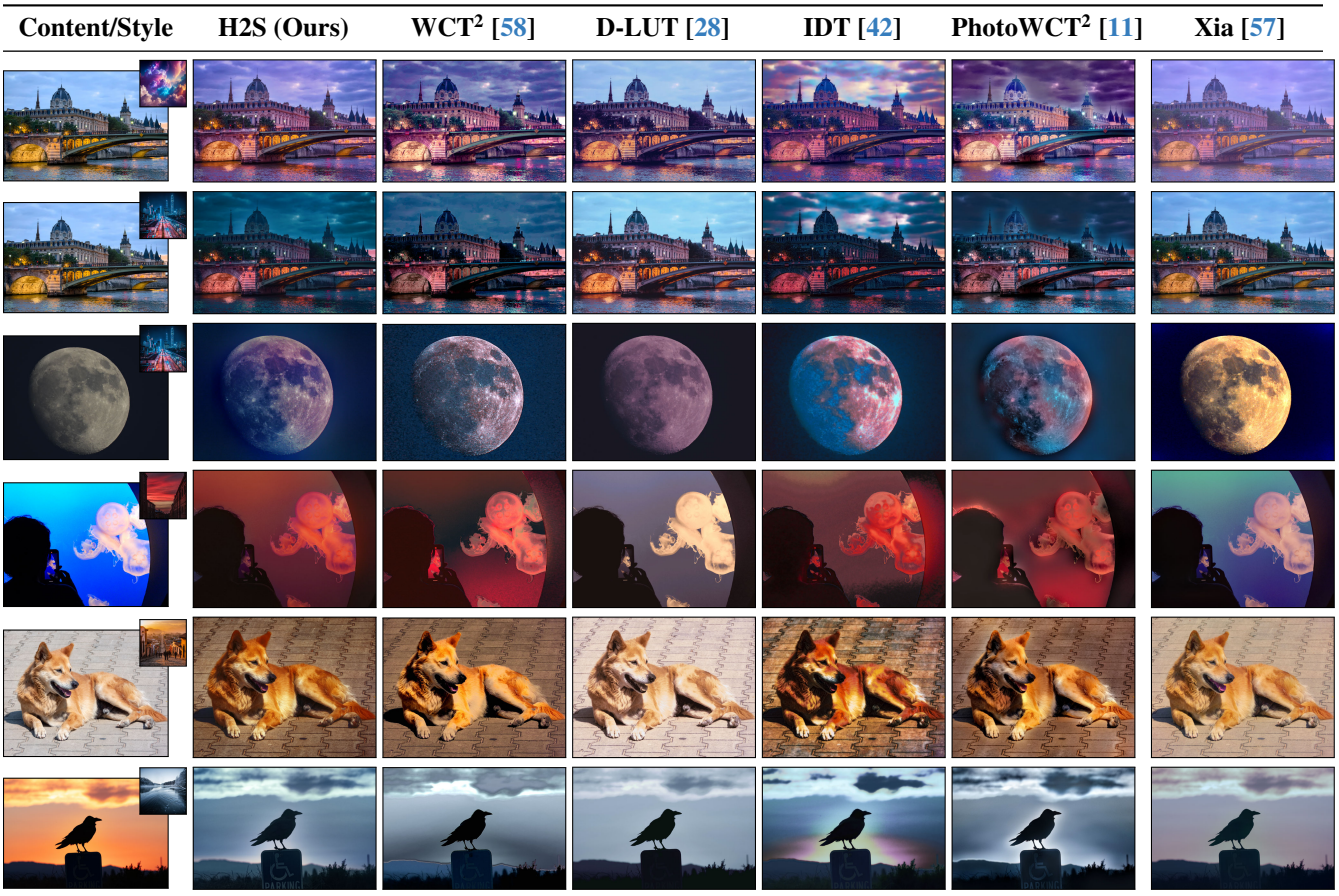

Figure 5. Qualitative comparisons with baseline methods on an independently collected, non-public dataset, highlighting Hist2Style's performance across a wide range of style–content pairs, including daytime, evening, vibrant, and monochrome scenes.

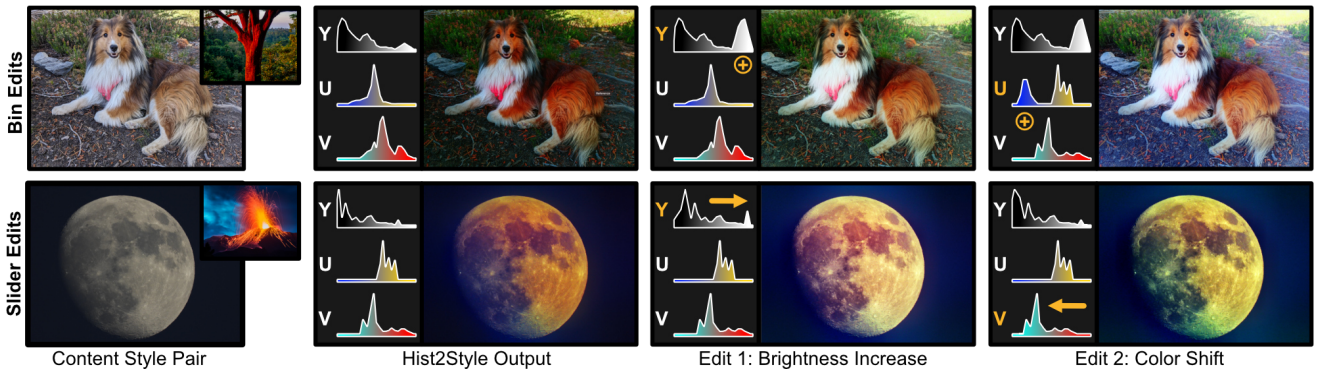

Figure 6. **User control**. We can interactively control the stylization process by directly editing the guiding histogram. This *global* user intent is fed to Hist2Style, which applies *local* changes that are adaptive to the image being edited. Users can edit the histograms directly by dragging the curve or use more familiar operations via custom sliders (see Sec. 3.6 for more details).

for 1127 epochs (each of 22.5K images) with batches of 64 images. Experiments are conducted on a single A100 GPU.

During training, for each content image, the dataloader randomly samples two style variants, one used as the input content image, and the other as simultaneously both the style image and ground truth image. Standard data augmentations such as random horizontal flips and resized crops are applied to improve generalization.

3.6. Interactive Histogram Manipulation

Histogram-conditioned stylization admits a natural interactive interface: rather than adjusting the output directly, the user manipulates the input histogram fed to the network, which translates **global** intent into **locally** coherent edits guided by priors learned during training. This mirrors traditional photography workflows such as sliders and tone curves [4], while operating entirely in histogram space to give users interpretable, precise control over color and tone. We recommend watching the accompanying supplementary video, available on our project page.

We implement the following sliders for the marginal histogram in Y'CbCr color space:

- Exposure slider $E \in [0, 1]$ applies a multiplicative factor on the luminance channel.
- Contrast slider $C \in [0, \infty]$ interpolates between a delta function at the peak of the luminance ($C = 0$), and the original luminance histogram ($C = 1$).
- U-shift slider $U \in [-1, 1]$ horizontally shifts the histogram mass of the Cb channel.
- V-shift slider $V \in [-1, 1]$ horizontally shifts the histogram mass of the Cr channel.
- Smoothing slider $S \in [0, 1]$ applies a smoothing function.

Additionally, we include an amount slider $A \in [-\infty, \infty]$ applied to the output of the model, which interpolates between the identity transformation ($A = 0$) and the model's predicted transformation ($A = 1$). Apart from slider control, we developed an interface for "direct histogram manipulation", allowing users to manually sculpt the histogram by dragging it. This mechanism can selectively add mass to, or subtract it from, a certain luma or chroma bin, while the model ensures the transformation obeys photorealistic image statistics, illustrated in the top row of Fig. 6.

4. Results

The data used for training is synthetically generated from the Unsplash Lite dataset consisting of 25K high-quality images [50]. We generated an average of 6 7 synthetic style variations per image, resized to 256×256px for training.

For evaluation, we first randomly selected 200 content photos from the Unsplash dataset [50] which we did not use for training, and then manually curated 136 natural images, removing ones with artificial or edited content. We also curated 19 style images from the Unsplash website [1] that were not in the training set.

We compare our algorithm's automated stylization to the state of the art in photorealistic style transfer via a user study. Then we compare to standard performance metrics such as runtime, memory, cycle consistency, and color matching score. Lastly, we propose an automated VLM-based Stylization Quality Assessment (SQA) metric which we find to be highly aligned with user preference.

Beyond automation, histogram guidance allows for fine user control and creative expression, as shown in Fig. 6. Further ablations on model components and training are shown in Fig. 8. Qualitative comparisons are presented in Fig. 4 and Fig. 5. Additional results and details for all of the above are provided in the supplementary material.

User Study

Leveraging the evaluation dataset described above, we designed a two-choice, anonymous user study to assess preference among stylized output. All comparisons used the **default** output of our model *without* user control, ensuring

Table 1. We conducted a **user study** comparing Hist2Style (H2S) to prior work. The table reports the percentage of trials that H2S won, tied, or lost against each baseline method. Our method outperforms the baselines in this study, winning over 61% of the trials against any single method. Note that SA-LUT was trained on log-encoded images which may contribute to its lower performance on RGB according to authors.

Method	H2S Win %	Tie %	Lose %
SA-LUT	**82.57%**	3.56%	13.86%
WCT^2	**72.58%**	5.85%	21.57%
Xia *et al.*	**73.24%**	4.10%	22.66%
IDT	**73.75%**	3.01%	23.25%
D-LUT	**70.59%**	3.04%	26.37%
PhotoWCT^2	**61.62%**	6.26%	32.12%

Table 2. **Runtime** (s) is reported across image resolutions and models. Hist2Style and Xia *et al.* are the fastest models when evaluated on novel content-style pairs. D-LUT requires an amortized cost for new style images, but is faster to apply if style is known in advance.

Method	256^2	512^2	1024^2	2048^2	4096^2
Hist2Style	**0.001**	**0.003**	*0.009*	*0.04*	*0.1*
Xia *et al.*	*0.003*	**0.003**	**0.004**	**0.008**	**0.03**
D-LUT	100	100	100	100	100
SA-LUT	0.2	0.2	0.2	0.2	0.2
PhotoWCT^2	0.3	0.3	0.3	0.4	1
IDT	0.1	0.2	0.3	0.4	0.9
WCT^2	0.04	0.07	0.1	0.4	OOM
ReHistoGAN	0.01	0.08	0.47	2.22	8.83

a fair comparison of automated stylization quality. In each trial, participants were shown a random content-style pair along with two stylized outputs, Hist2Style and a randomly selected baseline presented in random order. See Fig. 4 for example images from the study. Participants were instructed to choose "the image that best preserves the structure and details of the content image, while matching the color palette, tone, and overall aesthetic of the style image without introducing artifacts". They could also mark a trial as a tie when no clear preference existed.

We collected 3,000 valid trials from 31 experts in photography. As summarized in Tab. 1, Hist2Style achieves the highest overall preference, winning more than 61% of all recorded trials. Against PhotoWCT^2, the next strongest method, over 6% of comparisons resulted in ties; after accounting for these, PhotoWCT^2 attains a win rate below 33%. In addition to perceptual quality, Hist2Style provides an order-of-magnitude improvement in both runtime and peak memory usage relative to PhotoWCT^2, as discussed in Sec. 4. To facilitate comparison with other metrics, we define the User Score for each baseline method as its win rate against Hist2Style plus half the tie rate (H2S Lose % + $\frac{1}{2}$Tie %) as used in Fig. 7 and Tab. 3.

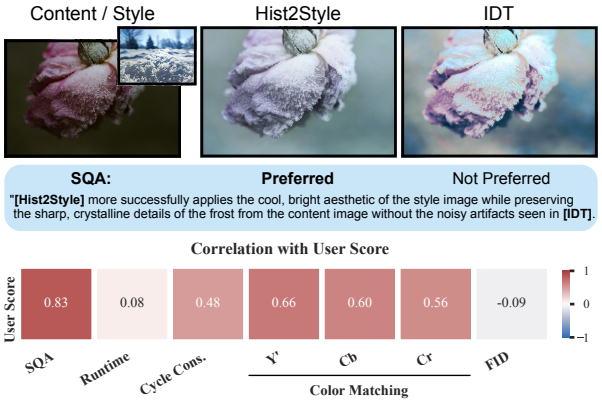

Figure 7. **Stylization Quality Assessment** (SQA) metric. We introduce SQA, a new metric for assessing stylization quality which is aligned to human preference, as shown in the correlation plot. For clarity, metrics were oriented so that higher is better.

Table 3. Quantitative comparison of style transfer methods. Lower is better (↓) for cycle consistency, color matching, and FID [9, 35], while higher is better (↑) for User Score and SQA (N=7000).

Method	User Score (↑)	SQA (↑)	Cycle Cons. (MSE ↓)	Color Matching (W_2^2 ↓)			FID (↓)
				Y'	Cb	Cr	
Hist2Style	**50.00**	**50.00**	*401.92*	387.30	49.85	80.53	71.44
PhotoWCT^2	*35.25*	*36.70*	1012.90	**224.86**	**23.14**	**52.41**	112.07
Xia *et al.*	24.71	24.12	646.40	2510.83	123.06	175.31	**50.46**
D-LUT	27.89	25.30	430.35	3850.52	201.46	307.20	87.65
SA-LUT	15.64	28.20	1613.97	4194.92	215.25	280.91	71.72
IDT	24.75	25.70	**215.57**	2692.25	169.78	203.85	92.32
WCT^2	24.50	16.10	990.43	*308.57*	*29.06*	*55.09*	97.96
ReHistoGAN	-	33.90	583.97	3000.75	136.84	207.18	*66.84*

Metrics

Stylization Quality Assessment Image stylization is inherently subjective, making quantitative evaluation challenging. While user studies remain the most reliable approach, they are difficult to replicate and their results can be ambiguous to interpret. To address this, we propose a stylization quality assessment (SQA) metric based on a vision-language model (VLM). Following recent work leveraging VLMs for image quality assessment [14, 15], we prompt a large VLM [49] with the same question posed to photographers in our user study. As shown in Fig. 7, SQA is highly correlated with the user score. For direct comparison with other metrics, see Tab. 3.

Cycle Consistency Inspired by [60], we assess methods' *cycle consistency* by stylizing an image from style A to B and back to A, and reporting the MSE relative to the original in Tab. 3. IDT attains the lowest cycle-consistency error as it maps colors globally back to their original distribution. Hist2Style follows, reinforcing our claims of spatial consistency, as local affine transformations in one bilateral grid can be effectively reversed by those of a subsequent grid.

Color Matching To assess how well each method reproduces the target color statistics, we evaluate the distribution-matching score defined in Algorithm 1 on the predicted stylizations. The results are summarized in Tab. 3. PhotoWCT^2 and WCT^2 achieve the lowest scores, corresponding to a high stylization strength. Hist2Style follows, demonstrating a strong compromise between reliable matching of global color statistics and higher perceptual quality, as supported by the user score and SQA in Tab. 3.

Runtime, Resolution, and Memory

Runtime and memory usage are key considerations for practical deployment of image-editing models, particularly on mobile or edge devices where compute and thermal budgets are limited. We therefore evaluate runtime and memory consumption across a range of input resolutions, with results summarized in Tab. 2.

Hist2Style and Xia *et al*. are the two fastest methods. Hist2Style is faster at lower resolutions, and both methods run in 0.003 s at 512×512. At higher resolutions, Xia *et al*. scales more favorably, processing 16 MP images in 0.03 s compared to 0.1 s for Hist2Style. All remaining baselines are several times slower. Although D-LUT incurs an amortized initialization cost of 100 s per style image, its runtime is the lowest (0.001 s at 16 MP) once this cost is paid. Hist2Style also remains lightweight in terms of memory usage, requiring only 1 GB to process 4 MP images, with additional comparisons provided in the supplementary material.

Ablations and Extensions

For further model ablations see the supplementary material.

Robust Training. Our training paradigm synthesizes ground truth images that are absent in test time. While our dataset's coherent styles allow us to mimic test-time conditions, the increased ambiguity causes such models to struggle with accurate color matching (see Fig. 8).

Color Space Loss. We find that using a perceptual loss improves fidelity (see Fig. 8). We attribute this to imperfect pixel alignment in our artificially generated data.

5. Discussion and Future Work

This work presents Hist2Style, a fast photorealistic stylization model that distills spatially varying color edits from a large image editing model into a lightweight network. By constraining the transform to local affine operations in bilateral space and conditioning on a histogram-based style embedding, Hist2Style preserves content and fine details while supporting expressive, spatially adaptive color changes and a simple interactive editing framework. In future work, we hope our proposed model can be further tailored to specific image editing tasks such as:

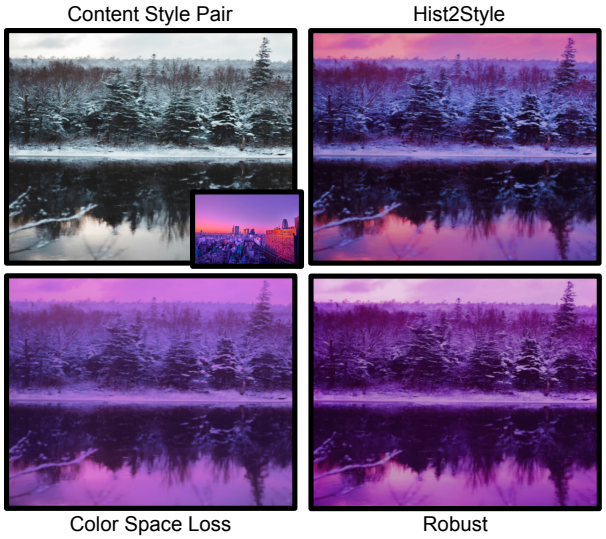

Figure 8. **Ablation studies and extensions.** We explored different architectural and training configurations. "Color Space Loss" refers to the output from a model where the loss was applied to raw color space during training, instead of perceptual space. Compared to this model, Hist2Style expresses a richer color transfer and better maintains image contrast. "Robust" refers to the output from a model trained not with the true histogram of the target image, but with that of another image in the same style. Compared to the robust model, Hist2Style displays similar contrast but improved color richness and accuracy. See the supplementary materials for additional experiments and analysis.

Multi-Layer Stylization. Although bilateral grids are spatially varying, a user may want further spatial control, such as assigning different styles to different objects in the scene. This could be explored in future work by co-optimizing a segmentation model to produce masks and interpolating the affine transforms across multiple selected styles and masks.

Nondestructive Content Editing. Although we focus on stylization without altering content, for some styles it may be necessary to add effects such as film grain, speckle, or bokeh. A potential direction for future work could combine Hist2Style's nondestructive color adjustments with optional effect layers to reproduce these looks while preserving photorealism.

Video and 3D Assets. While Hist2Style is designed for image stylization, the approach could potentially be extended to other domains such as video [28, 55, 58] and 3D representations [53]. Future work would include enforcing temporal consistency for long video sequences and integrating the method with radiance-field-based scene representations [53].

Acknowledgments

We thank Marc Levoy, Florian Kainz, Yifei Fan, Kevin Zhang, Ruiming Cao, Lars Jebe, Ethan Weber, Justin Yu, the Adobe Nextcam team, and WallerLab for valuable discussions and feedback. We thank the creators of software packages [2, 23, 34, 37, 46, 54]. Dekel Galor is supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1752814, and by the Center for Innovation in Vision and Optics. Laura Waller is a Chan Zuckerberg Biohub SF investigator, and partially funded by U.S. Air Force Office of Scientific Research award no. FA955-22-1-0521.

References

[1] Unsplash. https://unsplash.com. Accessed: 2025-11-13. 6

[2] Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. Gradio: Hassle-free sharing and testing of ml models in the wild. *arXiv preprint arXiv:1906.02569*, 2019. 9

[3] Andrew Adams, Jongmin Baek, and Abe Davis. Fast high-dimensional filtering using the permutohedral lattice. *Eurographics*, 2010. 3

[4] Adobe Systems Incorporated. Adobe photoshop lightroom, 2007. 4, 6

[5] Jonathan T Barron and Ben Poole. The fast bilateral solver. *ECCV*, 2016. 3

[6] Jonathan T Barron, Andrew Adams, YiChang Shih, and Carlos Hernández. Fast bilateral-space stereo for synthetic defocus. *CVPR*, 2015. 3

[7] Lukas Biewald. Experiment tracking with weights and biases, 2020. 4

[8] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3, 4

[9] Chaofeng Chen and Jiadi Mo. IQA-PyTorch: Pytorch toolbox for image quality assessment. [Online]. Available: https://github.com/chaofengc/IQA-PyTorch, 2022. 7

[10] Yaosen Chen, Han Yang, Yuexin Yang, Yuegen Liu, Wei Wang, Xuming Wen, and Chaoping Xie. Nlut: Neural-based 3d lookup tables for video photorealistic style transfer, 2023. 3

[11] Tai-Yin Chiu and Danna Gurari. Photowct^2: Compact autoencoder for photorealistic style transfer resulting from blockwise training and skip connections of high-frequency residuals, 2021. 3, 5

[12] Andrew Elliot and Markus Maier. Color psychology: Effects of perceiving color on psychological functioning in humans. *Annual review of psychology*, 65, 2013. 1

[13] Afifi et al. Histogan: Controlling colors of gan-generated and real images via color histograms. In *CVPR*, 2021. 3

[14] Chen et al. Toward generalized image quality assessment: Relaxing the perfect reference quality assumption, 2025. 7

[15] Li et al. Image quality assessment: From human to machine preference, 2025. 7

[16] William Falcon and The PyTorch Lightning team. PyTorch Lightning, 2019. 4

[17] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. A neural algorithm of artistic style. *CoRR*, abs/1508.06576, 2015. 1, 2

[18] Michaël Gharbi, Jiawen Chen, Jonathan T. Barron, Samuel W. Hasinoff, and Frédo Durand. Deep bilateral learning for real-time image enhancement. *ACM Transactions on Graphics*, 36(4):1–12, 2017. 3, 4

[19] Zerui Gong, Zhonghua Wu, Qingyi Tao, Qinyue Li, and Chen Change Loy. Sa-lut: Spatial adaptive 4d look-up table for photorealistic style transfer, 2025. 2, 3

[20] Charles Haine. *Color Grading 101: getting started color grading for editors, cinematographers, directors, and aspiring colorists*. Routledge, 2019. 1

[21] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 1

[22] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 2, 4

[23] John D Hunter. Matplotlib: A 2d graphics environment. *Computing in science & engineering*, 9(03):90–95, 2007. 9

[24] Yongcheng Jing, Yezhou Yang, Zunlei Feng, Jingwen Ye, and Mingli Song. Neural style transfer: A review. *CoRR*, abs/1705.04058, 2017. 1, 2

[25] Zhanghan Ke, Yuhao Liu, Lei Zhu, Nanxuan Zhao, and Rynson W.H. Lau. Neural preset for color style transfer. In *Computer Vision and Pattern Recognition Conference (CVPR)*, 2023. 3

[26] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[27] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. 3

[28] Mujing Li, Guanjie Wang, Xingguang Zhang, Qifeng Liao, and Chenxi Xiao. D-LUT: Photorealistic Style Transfer via Diffusion Process . In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 9206–9214, Los Alamitos, CA, USA, 2025. IEEE Computer Society. 3, 5, 8

[29] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. *CoRR*, abs/1705.08086, 2017. 2

[30] Yijun Li, Ming-Yu Liu, Xueting Li, Ming-Hsuan Yang, and Jan Kautz. A closed-form solution to photorealistic image stylization. *CoRR*, abs/1802.06474, 2018. 2

[31] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *CoRR*, abs/2201.03545, 2022. 4

[32] Fujun Luan, Sylvain Paris, Eli Shechtman, and Kavita Bala. Deep photo style transfer. *arXiv preprint arXiv:1703.07511*, 2017. 1, 2, 3, 4

[33] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models, 2022. 1, 3

[34] The pandas development team. pandas-dev/pandas: Pandas. 2020. 9

[35] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *CVPR*, 2022. 7

[36] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019. 4

[37] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011. 9

[38] Gabriel Peyré and Marco Cuturi. Computational optimal transport, 2020. 4

[39] P. F. Pitié and A. C. Kokaram. The linear monge-kantorovitch linear colour mapping for example-based colour transfer. In *Proceedings of the 4th IEE European Conference on Visual Media Production (CVMP)*, London, United Kingdom, 2007. 2

[40] P. F. Pitié, A. C. Kokaram, and R. Dahyot. N-dimensional probability density function transfer and its application to colour transfer. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Beijing, China, 2005.

[41] P. F. Pitié, A. C. Kokaram, and R. Dahyot. Towards automated colour grading. In *Proceedings of the 2nd IEE European Conference on Visual Media Production (CVMP)*, London, United Kingdom, 2005.

[42] P. F. Pitié, A. C. Kokaram, and R. Dahyot. Automated colour grading using colour distribution transfer. *Computer Vision and Image Understanding*, 2007. 2, 4, 5

[43] P. F. Pitié, A. C. Kokaram, and R. Dahyot. Enhancement of digital photographs using color transfer techniques. In *Single-Sensor Imaging*, pages 295–321. CRC Press, 2008. 2

[44] Julien Porquet, Sitong Wang, and Lydia B. Chilton. Copying style, extracting value: Illustrators' perception of AI style transfer and its impact on creative labor, 2025. 4

[45] E. Reinhard, M. Ashikhmin, B. Gooch, and P. Shirley. Color transfer between images. *IEEE Computer Graphics and Applications*, 21(5):34–41, 2001. 1, 2

[46] Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010. 9

[47] Xincheng Shuai, Henghui Ding, Xingjun Ma, Rongcheng Tu, Yu-Gang Jiang, and Dacheng Tao. A survey of multimodal-guided image editing with text-to-image diffusion models, 2024. 1, 3

[48] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2, 4

[49] Gemini Team, Rohan Anil, Sebastian Borgeaud, et al. Gemini: A family of highly capable multimodal models, 2025. 3, 7

[50] Unsplash. Unsplash lite dataset photos. GitHub repository. 3, 6

[51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. 4

[52] Jia Wang, Jie Hu, Xiaoqi Ma, Hanghang Ma, Xiaoming Wei, and Enhua Wu. Image editing with diffusion models: A survey, 2025. 1, 3, 4

[53] Yuehao Wang, Chaoyi Wang, Bingchen Gong, and Tianfan Xue. Bilateral guided radiance field processing, 2024. 3, 8

[54] Michael L Waskom. Seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021. 9

[55] Cheng-Yu Wei, N. Dimitrova, and Shih-Fu Chang. Color-mood analysis of films based on syntactic and psychological models. In *2004 IEEE International Conference on Multimedia and Expo (ICME) (IEEE Cat. No.04TH8763)*, pages 831–834 Vol.2, 2004. 1, 8

[56] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. 3

[57] Xide Xia, Meng Zhang, Tianfan Xue, Zheng Sun, Hui Fang, Brian Kulis, and Jiawen Chen. Joint bilateral learning for real-time universal photorealistic style transfer. *CoRR*, abs/2004.10955, 2020. 3, 4, 5

[58] Jaejun Yoo, Youngjung Uh, Sanghyuk Chun, Byeongkyu Kang, and Jung-Woo Ha. Photorealistic style transfer via wavelet transforms. In *International Conference on Computer Vision (ICCV)*, 2019. 2, 3, 5, 8

[59] Tianshan Zhang and Hao Tang. Style transfer: A decade survey, 2025. 1, 2, 3

[60] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *CoRR*, abs/1703.10593, 2017. 7