CLIPtone: Unsupervised Learning for Text-based Image Tone Adjustment

Hyeongmin Lee[1,*]   Kyoungkook Kang[2,*]   Jungseul Ok[1,2]   Sunghyun Cho[1,2]

POSTECH [1]GSAI & [2]CSE

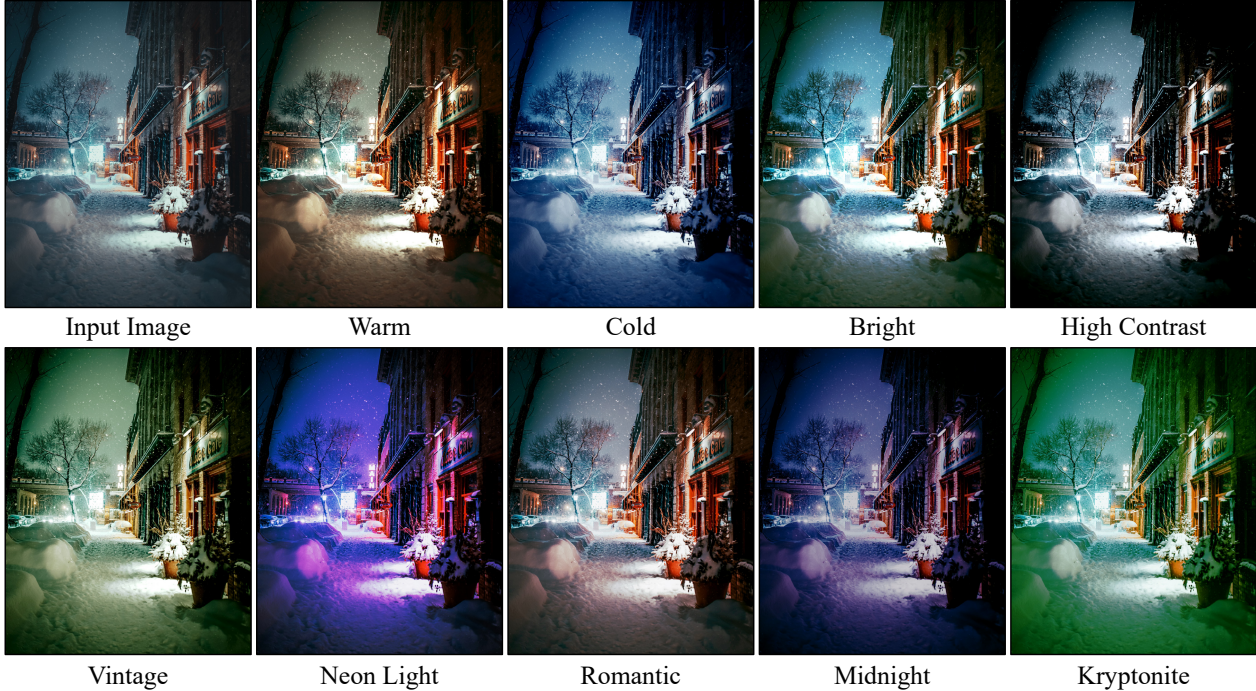{hmin970922, kkang831, jungseul, s.cho}@postech.ac.kr

Figure 1. We present CLIPtone, a text-based image tone adjustment framework trained in an unsupervised manner. With its superior understanding of natural languages, CLIPtone is capable of performing successful adjustments across a range of text descriptions, including those previously deemed challenging.

Abstract

Recent image tone adjustment (or enhancement) approaches have predominantly adopted supervised learning for learning human-centric perceptual assessment. However, these approaches are constrained by intrinsic challenges of supervised learning. Primarily, the requirement for expertly-curated or retouched images escalates the data acquisition expenses. Moreover, their coverage of target styles is confined to stylistic variants inferred from the training data. To surmount the above challenges, we propose an unsupervised learning-based approach for text-based image tone adjustment, CLIPtone, that extends an existing image enhancement method to accommodate natural language descriptions. Specifically, we design a hyper-network to adaptively modulate the pretrained parameters of a backbone model based on a text description. To assess whether an adjusted image aligns with its text description without a ground-truth image, we utilize CLIP, which is trained on a vast set of language-image pairs and thus encompasses the knowledge of human perception. The major advantages of our approach are threefold: (i) minimal data collection expenses, (ii) support for a range of adjustments, and (iii) the ability to handle novel text descriptions unseen in training. The efficacy of the proposed method is demonstrated through comprehensive experiments including a user study.

1. Introduction

Image tone adjustment aims for the alteration of the tonal properties of an image, including brightness, contrast, and color balance. It is also termed image tone enhancement, as it has primarily been exploited to enhance the image aesthetics. It is essential for various applications ranging from photography to medical imaging, and the importance of ef-

---

* Equal contribution

fective tone adjustment techniques has grown significantly.

Recent learning-based approaches [15, 17, 35, 47] predominantly rely on pairwise datasets and supervised learning to learn perceptual adjustments aligned with human perception. Representative datasets like MIT-Adobe 5K [8] and PPR10K [31], which comprise pairs of an image and its expert-retouched version, have emerged, and several methods [17, 25, 36, 50] have been proposed to learn the mappings between such pairs of images. However, such supervised learning-based methodologies bear several limitations. First, they demand cost-intensive datasets stemming from expert retouching. Second, they are confined to limited amounts of adjustments that are inferred from the datasets. Lastly, they allow for only automatic image enhancement and lack the capability for diverse adjustments with any guidance, as the datasets are composed of paired images without specific guidance annotations.

To reduce the burden of collecting paired images, a few studies [11, 19, 22, 26, 52] suggest weakly-supervised learning-based approaches to learn the mapping from the unpaired images. They predominantly leverage the adversarial learning of GANs [16] to capture the tonal properties of a target image set. Although they do use unpaired data, there is still a requirement to manually define distinct sets for both source and target images. In addition, they also share the other limitations of the supervised learning-based approaches.

Meanwhile, a handful of works [23, 42] introduce text-based approaches to enable controllable adjustments utilizing text descriptions. To this end, they construct datasets consisting of triplets (i.e., a source image, a target image, and a text description on the adjustment) and train neural networks on them. However, due to the reliance on a limited range of descriptions in the datasets, they cannot accommodate natural language descriptions beyond them.

In this paper, we propose a novel text-based image tone adjustment framework, CLIPtone, that is trained in an unsupervised manner requiring only source images and any tone-related text descriptions without the need for paired images. Our core insight is that the assessment of perceptual adjustments can be achieved with the recent language-image representation model, CLIP [38], as it is trained on a vast amount of language-image pairs and thus encapsulates the human perception. Grounded on this insight, we extend an existing image enhancement model to enable varied adjustments based on text descriptions. Specifically, we design a hyper-network to adaptively modulate the pretrained parameters of a backbone model based on a text description. In order for our network to learn tone adjustment in an unsupervised manner, we further introduce a CLIP directional loss and regularization losses.

CLIPtone enjoys several unique benefits stemming from introducing CLIP as criterion of human perception. It necessitates only arbitrary images and tone-related text descriptions for its training, which can be collected with minimal costs. It also supports vast amounts of adjustments previously deemed challenging with text descriptions, as shown in Fig. 1, thanks to CLIP's comprehensive understanding of natural language. Lastly, it is capable of handling novel text descriptions unseen in training.

Recently, in line with remarkable capabilities of contemporary generative models, numerous studies [7, 12, 14, 34] have been proposed dictating the attributes of the generated images with text descriptions. Among them, our work shares similarities with text-based image colorization and text-based image editing approaches, in the aspect of modifying input images based on text descriptions. Compared to these methods, CLIPtone stands out for its remarkably fewer parameters and shorter inference time.

Our contributions are summarized as follows:
- We present, for the first time, an unsupervised image tone adjustment framework CLIPtone, that captures human perception on tonal properties utilizing CLIP, enabling a range of adjustments and zero-shot prediction.
- We design a hyper-network to effectively modulate the pretrained parameters of an existing image enhancement network and suggest training strategies tailored for unsupervised learning of CLIPtone.
- Through extensive experiments, we demonstrate that our approach is lighter, faster, and more effective compared to existing methods.

2. Related Work

Image Tone Adjustment Methods. To improve the aesthetic quality of images, a range of learning-based automatic tone adjustment (or enhancement) methods have been developed. They have explored optimal neural architectures to directly estimate dense pixel values [21, 35, 36] or the parameters of traditional physical models such as color transform matrices [9, 15, 49], and Look-Up Tables (LUTs) [17, 25, 32, 47, 50–53]. Among these, methods exploiting 3D LUTs [32, 47, 50–53] have garnered attention for their superior capacity and efficiency. Since Zeng *et al*. [52] have proposed adaptive 3D LUTs to input images, Wang *et al*. [47] proposed spatially-aware 3D LUTs for local adjustment, and Yang *et al*. [50] increased the capacities of 3D LUTs through non-uniform sampling.

For neural networks to learn adjustments aligned with human perception, supervised learning-based approaches [15, 17, 35, 47] have predominantly been attempted. They rely on paired datasets such as MIT-Adobe 5K [8] and PPR10K [31], which consist of source images and their expert-retouched counterparts. To alleviate the collection costs associated with expert retouching, weakly-supervised learning-based approaches [11, 19, 26, 52] have been introduced. They use unpaired data comprised of separate source and target sets of images. These methods primarily leverage the adversarial loss from GANs [16] to capture the tonal properties of target sets, and the cycle-consistency loss [55] for preserving the structures of source images.

A few methods [23, 42] employ text descriptions as conditional inputs to enable diverse and controllable adjustment beyond automatic adjustment. They construct train-
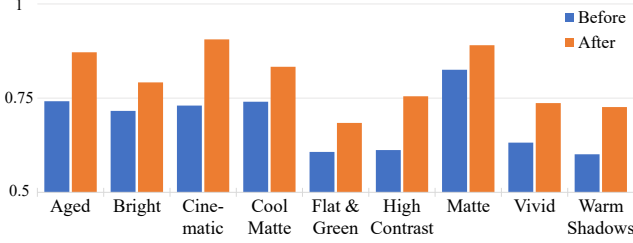
Figure 2. We apply tone adjustment filters from Adobe Lightroom Classic [1] to 500 images from the MIT-Adobe 5K dataset [8], and calculate the relative similarities between the images and filter names in the CLIP space. For all filters, the filtered images have higher similarity scores than the source images, implying that CLIP can assess the tonal properties of images in a manner aligning with human perception.

ing datasets consisting of triplets: a source image, a target image, and a text description on the adjustment and train a network on these datasets. Specifically, Shi *et al.* [42] annotated requests to each pair of the MIT-Adobe 5K dataset [8], and Jiang *et al.* [23] further proposed dataset augmentation strategies for imbalanced annotations.

Despite the aforementioned efforts, existing tone adjustment methods have limitations including high data collection costs (even for curating target sets), confined adjustments to stylistic variations of training datasets. In contrast, CLIPtone is independent of these issues thanks to leveraging CLIP as a perceptual criterion.

Text-driven Generative Models. Recently, cutting-edge generative models have found their uses in image-to-image translation tasks such as image colorization [10, 20] and editing [5, 7, 12]. Early approaches use GANs [16] trained on specific domains to manipulate real images through GAN inversion techniques [3, 24, 40, 45], changing their latent vectors [6, 37, 48] or adapting generator networks [14, 30] to different domains according to text descriptions. More recently, with advances in auto-regressive models [13] and diffusion models [18, 43, 44], generative models trained on large amounts of text-image pairs have been showing superior performance in image synthesis [41]. Several studies have proposed to finetune these models to manipulate images based on text descriptions [7, 10, 54] benefiting from their superior generative prior. While these image manipulation methods seem to have similarities with our methods, they struggle to maintain the structures of original images and require excessive time for manipulation.

3. CLIP as a Perceptual Criterion

As mentioned in Sec. 1, we aim to train a text-based tone adjustment network without paired images by leveraging CLIP [38]. This is based on a hypothesis that CLIP inherently encapsulates information about human perception learned from a vast amount of image-language pairs so that it can be employed for assessing tone adjustment in a manner consistent with human perception.

To verify our hypothesis, we conduct an experiment to examine whether CLIP can correctly assess images whose tones are already known. To this end, we adopt tone adjustment filters manually designed by experts. Specifically, we randomly sample 500 images from the MIT-Adobe 5K dataset [8], and apply nine filters such as "aged", "bright", "cinematic" from Adobe Lightroom Classic [1], and construct a dataset consisting of source and filtered image pairs for each filter. In each dataset, the source images have natural tones while the filtered images have specific tones produced by each filter. We then compute their similarities to their corresponding filter names in the CLIP embedding space. Specifically, adopting the idea of CLIP-IQA [46], we compute relative similarities defined as:

$$S(I, T) = \text{softmax}(\text{sim}(I, T), \text{sim}(I, T')) \qquad (1)$$

where I and T are an image and a filter name, respectively. $\text{sim}(I, T)$ is the cosine similarity between the CLIP embeddings of I and T. T' is a text description "normal photo". To obtain CLIP embeddings, we use the CLIP-RN50 model in this experiment.

Fig. 2 illustrates the result where the filtered images have higher similarity scores than the source images for all filters. This result indicates that CLIP holds the potential to serve as a differentiable perceptual criterion for tone adjustment evaluation. Additional results employing other CLIP models are included in the supplementary material.

4. Network Architecture of CLIPtone

The network architecture of CLIPtone comprises a tone adjustment network and a text adapter as shown in Fig. 3. The tone adjustment network, for which we adopt an existing image enhancement network, takes an input image and alters the tonal properties of the image, while the text adapter takes a target text description as input and modulates the parameters of the tone adjustment network for text-based tone adjustment. Our design philosophy can be regarded as modulating a backbone network using a hyper-network, a strategy whose efficacy has been demonstrated in various prior studies [4, 27, 33, 39]. Here, the selection of the backbone network and defining an efficient and effective modulation strategy are crucial, and we describe our choices in the following.

Tone Adjustment Network. For the tone adjustment network, we adopt the image enhancement network of Yang *et al.* [50], which is based on 3D LUTs. Here, we take a brief look at Yang *et al.*'s method for completeness. Their method takes an input image and extracts a feature map using a convolutional neural network (CNN). Then, from the feature map, a weight predictor network estimates the weights w of learnable basis LUTs. The basis 3D LUTs are then linearly combined using the weights w to form a fused 3D LUT. Meanwhile, from the feature map extracted by the CNN, an AdaInt module estimates sampling coordinates, which determine the source color values of the fused 3D LUT. Specifically, an AdaInt module estimates sampling coordinates $x_c(i)$ where $c \in \{r, g, b\}$ and i denote the color axis
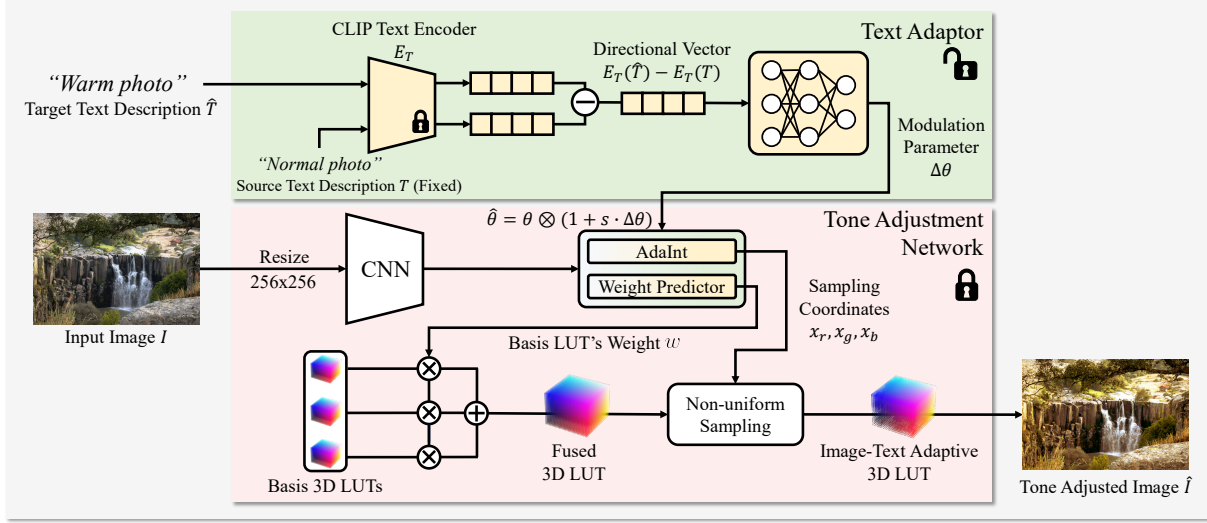
Figure 3. CLIPtone consists of a text adapter and a tone adjustment network. From a target text description, the text adapter calculates a directional vector within the CLIP embedding space from the source to target text descriptions and estimates the modulation parameter $\Delta\theta$ for the AdaInt module and the weight predictor of the tone adjustment network. The modulated tone adjustment network adaptively constructs an image-text adaptive 3D LUT through fusing basis 3D LUTs and non-uniform sampling, ultimately adjusting the color values of an input image.

and an index along each color axis. For example, $x_r(1) = 0.2$, $x_g(2) = 0.3$, and $x_b(3) = 0.4$ mean that the color value $(0.2, 0.3, 0.4)$ of an input image will be converted to the color value of the fused 3D LUT at $(1, 2, 3)$. Then, the non-uniform sampling step resamples the fused 3D LUT according to the estimated sampling coordinates, and obtains a final image-adaptive 3D LUT. Finally, the color values of the input image undergo trilinear-interpolated look-up operations to produce a color-enhanced image. We refer the readers to [50] for more details.

We build our approach on top of the method of Yang *et al.* [50] because it adjusts only the tonal properties of an input image using LUTs without modifying the contents, and provides an efficient and effective architecture leveraging basis LUTs. In our approach, we use the basis LUTs from the Yang *et al.*'s model, which are pre-trained for the image enhancement task, without modification. We modulate only the parameters of the weight predictor and the AdaInt module with our text adapter to construct an *image-text adaptive 3D LUT* as shown in Fig. 3. These two modules are the most crucial ones as they determine how to combine the basis 3D LUTs, how to resample a fused 3D LUT, and ultimately how to adjust an image. All the other parameters including the basis LUTs are kept frozen. This selective modulation enables the efficient and effective learning of our text adapter.

Text Adapter. The text adapter modulates the parameters for the AdaInt and weight predictor modules according to a target text description $\hat{T}$. To this end, the text adapter first embeds the input target text description $\hat{T}$ into the CLIP embedding space using a CLIP text encoder E_T. The text adapter also uses a source text description T, which is fixed to "normal photo", to represent the tone of the input image.

To embed T and $\hat{T}$, we adopt the CLIP-RN50 text encoder. Then, it computes a directional vector by computing the difference between the embeddings of the target and source text descriptions, i.e., $E_T(\hat{T}) - E_T(T)$, and feeds the directional vector to a two-layered MLP to compute modulation parameter $\Delta\theta$. Finally, similar to the previous hypernetwork-based approach [4], the text adapter modulates the parameters of the AdaInt and weight predictor modules as:

$$\hat{\theta} = \theta \otimes (1 + s \cdot \Delta\theta) \qquad (2)$$

where θ is the original parameter of the AdaInt module and the weight predictor, and $\hat{\theta}$ is their modulated parameter. s is a scaling factor that controls the degree of adjustment, and $\otimes$ denotes Hadamard product.

Regarding the source description, we have explored different schemes: text-dependent, image-dependent, and neutral schemes. The text-dependent scheme derives source descriptions from input texts, *e.g.*, antonyms of input texts, while the image-dependent scheme utilizes the CLIP image embedding of an input image instead of the CLIP text embedding of a source text description. The neutral scheme utilizes a fixed text description representing neutrality, such as "normal photo".

Among these options, we found that the neutral scheme works the best due to the following reasons. (i) The image-dependent scheme tends to cause unstable training due to the continuous variation in the source prompt. (ii) The text-dependent scheme is impractical as it requires finding antonyms for the input text, which can be challenging. (iii) The neutral scheme fits our scenario the best, where an image may already be colorful, but a user may still want to edit the image with the keyword "colorful" for a more enhanced

result. Additionally, we have evaluated different neutral descriptions, *e.g.*, "ordinary photo" and "photo", and empirically found that "normal photo" yields the best results.

5. Unsupervised Learning for CLIPtone.

Training Strategy. We train CLIPtone on two datasets: one for text descriptions and the other for source images. We note that those are neither paired nor labeled, i.e., unsupervised learning. Specifically, for text descriptions, we adopt the Color Names Dataset [2] encompassing 30,138 color names with duplicates to learn adjustments to various tones. Before we feed the descriptions to the CLIP text encoder, we append "photo" to each description (e.g., a color name "red" is transformed into "red photo"). For source images, we use 4,500 source images from the MIT-Adobe 5K dataset [8]. We augment the images with random-crop and horizontal-flip, and random adjustments of bright and saturation, following previous work [52]. For initialization of the text adapter, the weights are sampled from a uniform distribution $\mathcal{U}(0, 0.01)$, and the biases are set to zero. For training, we use the Adam optimizer [28] with the learning rate $1e-3$ and the coefficients $\beta_1 = 0.9$ and $\beta_2 = 0.999$.

Training Loss. Given a randomly sampled pair of a target description $\hat{T}$ and a source image I, we train CLIPtone using a loss $\mathcal{L}$ defined as:

$$\mathcal{L} = \lambda_{\text{content}}\mathcal{L}_{\text{content}} + \lambda_{\text{CLIP}}\mathcal{L}_{\text{CLIP}} + \lambda_{\text{LUT}}\mathcal{L}_{\text{LUT}} , \quad (3)$$

where $\mathcal{L}_{\text{content}}$, $\mathcal{L}_{\text{CLIP}}$, and $\mathcal{L}_{\text{LUT}}$ represent a content-preserving loss, a CLIP directional loss, and an LUT regularization loss, respectively. The parameters λ_{content}, λ_{CLIP}, and λ_{LUT} are used to balance these loss terms, and we set each of them to 1.

The content-preserving loss $\mathcal{L}_{\text{content}}$ prevents the tone-adjusted output $\hat{I}$ from excessively deviating from the source image I, and is defined as the mean-squared error (MSE) between I and $\hat{I}$. The CLIP directional loss $\mathcal{L}_{\text{CLIP}}$ aligns the difference from I to $\hat{I}$ with that from T to $\hat{T}$. To this end, we employ the CLIP directional loss of StyleGAN-NADA [14], which is defined as:

$$\mathcal{L}_{\text{CLIP}}(\Delta I, \Delta T) = 1 - \frac{\Delta I \cdot \Delta T}{\|\Delta I\|\|\Delta T\|} \quad (4)$$

where $\Delta I := E_I(\hat{I}) - E_I(I)$ and $\Delta T := E_T(\hat{T}) - E_T(T)$.

Lastly, we devise the LUT regularization loss $\mathcal{L}_{\text{LUT}} = \lambda_{\text{weight}}\mathcal{L}_{\text{weight}} + \lambda_{\text{interval}}\mathcal{L}_{\text{interval}}$ to ensure that the image-text adaptive 3D LUT is locally smooth, thereby enabling natural and stable adjustments. We set λ_{weight} and $\lambda_{\text{interval}}$ to $1e-4$ and 0.5, respectively. The regularization is given in two-fold: $\mathcal{L}_{\text{weight}}$ and $\mathcal{L}_{\text{interval}}$ for the weight predictor module and for the AdaInt module, respectively. We use L2 regularization $\mathcal{L}_{\text{weight}} := \|w\|^2$ on the weights w of the basis LUTs to discourage individual weights from becoming disproportionately large compared to the others.

$\mathcal{L}_{\text{interval}}$ is a sampling interval loss for regularizing the adaptive sampling coordinates from the AdaInt module, which is defined as:

$$\mathcal{L}_{\text{interval}} = \sum_{l,i,j,k,c} \left| \frac{L_c^{(l)}(i+1,j,k) - L_c^{(l)}(i,j,k)}{(x_r(i+1) - x_r(i))^\alpha} \right|^2$$
$$+ \left| \frac{L_c^{(l)}(i,j+1,k) - L_c^{(l)}(i,j,k)}{(x_g(j+1) - x_g(j))^\alpha} \right|^2$$
$$+ \left| \frac{L_c^{(l)}(i,j,k+1) - L_c^{(l)}(i,j,k)}{(x_b(k+1) - x_b(k))^\alpha} \right|^2, \quad (5)$$

where $l \in \{1, \cdots, L\}$ is an index to a basis LUT, and L is the number of the basis LUTs, which is 3 in our framework following Yang *et al.*'s method [50]. $c \in \{r, g, b\}$ is an index to a color channel. $L_c^{(l)}(i, j, k)$ is the color value of the c color channel of the l-th basis 3D LUT at (i, j, k). $x_c(i)$ is the i-th sampling coordinate for the c color channel estimated by the AdaInt module. α is a hyper-parameter to control the regularization strength. Larger α more strongly penalizes narrow intervals between consecutive sampling coordinates, leading to more evenly distributed sampling coordinates, so that disrupt color changes can be avoided. Conversely, smaller α gives a weaker penalty to narrow intervals, allowing the sampling coordinates to better reflect the target text description. In our experiments, we set α to 0.7. $\mathcal{L}_{\text{interval}}$ ensures that the distances in the estimated sampling coordinates, i.e., the sampling intervals, are not too small especially when neighboring LUT entries have large color difference, thereby preventing abrupt color changes in tone-adjusted images.

6. Experiments

6.1. Comparative Evaluation

Baselines. As CLIPtone is the first unsupervised learning-based approach for text-based image tone adjustment, we compare it with other state-of-the-art text-based image manipulation methods: T2ONet [42], IP2P [7], and CLIPstyler [29]. T2ONet [42] is a request-based image enhancement model, which is trained on text-based requests and corresponding before-and-after images in a supervised manner. IP2P [7] is also a request-based image editing model that fine-tunes a pretrained text-based image generative model to accommodate a source image and a text-based request. CLIPstyler [29] is a style transfer model, which is trained to increase the directional CLIP similarity for a single target text description. As these models use different forms for target descriptions, we alter input text descriptions as follows: A text description "krypton" is transformed to "krypton photo" for CLIPtone, "Make a krypton photo." for T2ONet [42] and IP2P [7], and "krypton" for CLIPstyler [29]. For the baseline models, we use the authors' official implementations.

Qualitative Comparisons. Fig. 4 shows a qualitative comparison. In the figure, T2ONet [42] barely changes the in-
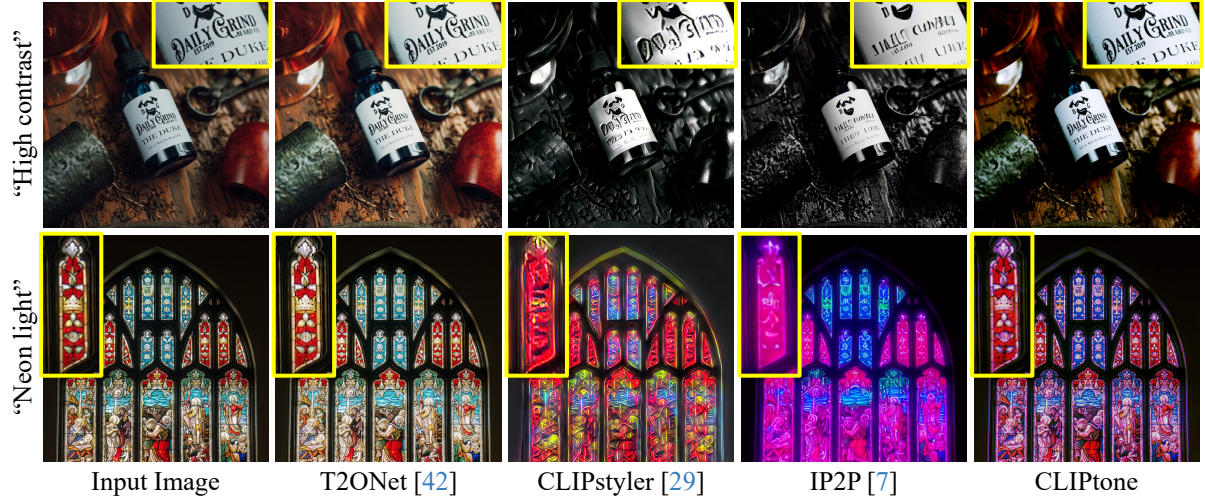
Figure 4. Qualitative comparisons against baselines for modifications of input images according to given text descriptions. T2ONet [42] induces only subtle changes and fails to perform appropriate adjustments. While CLIPstyler [29] and IP2P [7] make appropriate adjustments aligned with the text descriptions, they fail to preserve the image contents. In contrast, CLIPtone successfully makes appropriate adjustments while preserving the contents of the input images.

Method	Grayscale SSIM ↑	CLIP image similarity ↑	CLIP text-image directional similarity ↑
T2ONet [42]	**0.936**	**0.994**	0.021
CLIPstyler [29]	0.484	0.669	**0.095**
IP2P [7]	0.765	0.932	0.073
CLIPtone	<u>0.858</u>	<u>0.975</u>	<u>0.083</u>

Table 1. Quantitative comparison of different methods on 500 images from the MIT-Adobe 5K dataset [8] and manually collected 50 tone-related text descriptions. CLIPtone achieves overall higher scores in both Grayscale SSIM and CLIP image similarity, which evaluate structural preservation, as well as in CLIP text-image directional similarity, which assesses the alignment with text descriptions, compared to existing methods.

put images and fails to perform appropriate adjustment. Although CLIPstyler [29] and IP2P [7] succeed in making adjustment to fit the descriptions, they fail to preserve the structures of the input images. CLIPtone outperforms all the other methods, achieving both appropriate adjustments reflecting the target descriptions and successful preservation of the structural information in the input images.

Quantitative Comparisons. For quantitative evaluation, we construct a test set consisting of 500 source images from MIT-Adobe 5K [8], which are not used for training CLIPtone, and 50 manually collected text descriptions related to tonal properties such as "warm", "cold", and "pastel tone". We provide all the descriptions used in our evaluation in the supplementary material.

To assess the adjustments by each method, we focus on two primary factors: the preservation of image structures, and the alignment with text descriptions. To assess the structure preservation, we use grayscale SSIM and CLIP image similarity. Grayscale SSIM computes the SSIM index between two images after grayscale conversion, while CLIP image similarity measures the cosine similarity between

two images in the CLIP embedding space to assess the similarity between their contents.

However, it is important to note that both metrics are limited in assessing the structure preservation since assessing the structure preservation between two images is not a trivial task, which requires isolating the image structure from tones and colors. Nevertheless, both metrics are less affected by tones and colors than other metrics such as PSNR since grayscale SSIM discards the colors and also includes contrast normalization, and CLIP image similarity focuses on content similarity rather than tones and colors.

To assess alignment with text descriptions, we measure the CLIP text-image directional similarity. Specifically, given a pair of an input image and a target text description, we compute the directional vector between the CLIP embeddings of the input image and its tone-adjusted image, and the directional vector between the CLIP embeddings of the source and target text descriptions. Then, we compute the cosine similarity between the directional vectors. Meanwhile, as T2ONet [42] supports only predefined descriptions for adjustments, we evaluate the performance of T2ONet only on 21 compatible descriptions out of 50.

In Tab. 1, we report the quantitative comparison. The CLIP image similarity and CLIP text-image directional similarity scores in the table are computed using the CLIP ViT-B/16 model, which is not used for training any methods for a fair comparison. Other results with other CLIP models are included in the supplementary material. In the table, T2ONet [42], which induces subtle modifications in adjustment process, yields high scores in Grayscale SSIM and CLIP image similarity. Yet, it markedly underperforms in CLIP text-image directional similarity than the other methods. Since CLIPstyler [29] is trained solely on target descriptions, it records the highest CLIP text-image directional similarity score. However, it struggles to preserve

Methods	T2ONet [42]	CLIPstyler [29]	IP2P [7]	CLIPtone
CLIP image similarity	**0.956**	0.704	0.891	<u>0.948</u>
CLIP directional similarity	<u>0.359</u>	0.297	0.310	**0.467**

Table 2. Quantitative comparison using pseudo-ground-truth images created by tone adjustment filters of Adobe Lightroom Classic [1].

Method	T2ONet [42]	CLIPstyler [29]	IP2P [7]	Ours
Param (M)	<u>21.1</u>	**6.7**	1016.9	111.4
Time (ms)	45.1	<u>26.6</u>	7177.4	**10.3**

Table 3. Comparison of the number of parameters (M) and inference times (ms) for processing a 512×512-sized image using an NVIDIA GeForce RTX 3090 GPU.

Method	Grayscale SSIM ↑	CLIP Image Similarity ↑	CLIP Text-Image Directional Similarity ↑
CLIPtone-0 (trained w/o target keyword)	**0.859**	**0.968**	<u>0.087</u>
CLIPtone (trained w/ target keyword)	<u>0.858</u>	<u>0.964</u>	**0.089**

Table 4. Quantitative comparison of CLIPtone-0 trained without target keywords and CLIPtone trained with target keywords. CLIPtone-0 performs nearly as well as CLIPtone in all the metrics.

image structures, leading to the lowest scores in Grayscale SSIM and CLIP image similarity. CLIPtone achieves high scores in all three metrics, demonstrating that it outperforms existing methods when considering both primary factors.

Comparison using Pseudo-Ground-Truth Images. We also compare the baseline methods and ours using pseudo-ground-truth images, which are generated by applying nine tone adjustment filters of Adobe Lightroom Classic [1] to our test set as done in Sec. 3. Fig. 5 shows a qualitative comparison. As the figure shows, CLIPtone excels not only in preserving image structures but also in adjusting tones, producing results more similar to the results of the filters designed by experts. For quantitative analysis, we measure the structural similarity between tone-adjusted images and their corresponding pseudo-ground-truth images using CLIP image similarity. We also measure the similarity between the adjustments made to a tone-adjusted image and to its corresponding pseudo-ground-truth image for a given input image using CLIP directional similarity. Tab. 2 shows that CLIPtone achieves the highest CLIP directional similarity score and the second highest CLIP image similarity score, indicating that CLIPtone performs tone adjustment similar to the filters manually designed by experts.

User Study. Given that image tone adjustment is fundamentally a human-centric task, we conducted a user study to evaluate the baseline methods and ours. For this study, we recruited 20 participants from our institution. Each participant was presented 30 examples consisting of an input image and randomly arranged resulting images from each method, and was asked to score these on a scale from 1 (poor) to 5 (excellent) based on the following criteria: (a) How well the structure of the input image is preserved. (b) How well the modification suits the text description. (c) The

aesthetic quality of the result image. The complete questionnaires are included in the supplementary material. Fig. 6 illustrates the average of the results of the user study. Against all the other methods, CLIPtone is preferred in all three aspects, demonstrating that CLIPtone outperforms existing methods from the perspective of human perception.

Model Complexity. As shown in Tab. 3, CLIPtone has more parameters than T2ONet [42] and CLIPstyler [29], each designed to address single and restricted text descriptions, respectively. Nevertheless, CLIPtone exhibits notably shorter inference times than the others thanks to its efficient backbone network and our hyper-network-based approach.

6.2. Deeper Analysis of CLIPtone

In this section, we conduct a detailed analysis on CLIPtone, including the validation of the sampling interval loss and a series of experiments that demonstrate the unique and practical advantages of CLIPtone.

Sampling Interval Loss. Fig. 7 shows that the absence of $\mathcal{L}_{\text{interval}}$ causes the model to over-adjust, significantly diminishing the aesthetic quality of the output image. This happens because the estimated sampling coordinates from the AdaInt module lead to the construction of an image-text adaptive 3D LUTs with abrupt color transitions, focusing on matching the target text descriptions. In contrast, CLIPtone with the sampling interval loss constructs an image-text adaptive 3D LUT with more gradual color transitions, ensuring that the result image is not only aligned with the target text descriptions but also visually appealing.

Zero-shot Prediction. Thanks to leveraging CLIP that is pretrained on a vast amount of image-text pairs, CLIPtone can also make suitable adjustments for text descriptions unseen during training. Here we verify this zero-shot prediction capability of CLIPtone. To this end, we construct a test set using 50 target keywords used in the quantitative evaluation. We then exclude any text descriptions that include the keywords from our training set. As a result, we exclude 2,890 text descriptions, and obtain a subset of 27,248 text descriptions. Then, we train CLIPtone on the subset. We refer to this version of CLIPtone as 'CLIPtone-0'. Finally, we compare CLIPtone-0 and CLIPtone on the test set of the sampled target keywords. Tab. 4 and Fig. 8 present the quantitative and qualitative comparisons. These comparisons reveal that CLIPtone-0 performs nearly as well as CLIPtone, suggesting that CLIPtone is capable of handling novel text descriptions not included in its training set. A more comprehensive collection of zero-shot prediction results is included in the supplementary material.

Scaling the Strength of Adjustment. CLIPtone also offers controllable adjustment by changing the scaling factor s in Eq. (2), which scales the weight offsets of the backbone network estimated by the text adapter. Fig. 9 illustrates an example that shows smooth transition of the tone according to the scaling factor s.

Figure 5. Qualitative comparison against pseudo-ground-truth images created by tone adjustment filters of Adobe Lightroom Classic [1]. Compared to the baselines, CLIPtone performs more similar adjustment to the Adobe filters while preserving the original image structure.
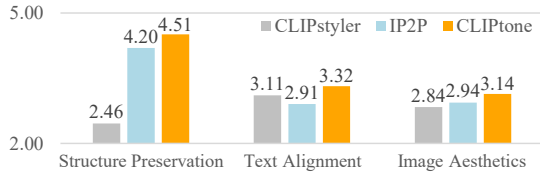

Figure 6. Summary of the user study with 20 participants evaluating 30 examples. Each graph represents the average score of each of the compared methods for each of the following questions: (a) How well the structure of the input image is preserved. (b) How well the modification suits the text description. (c) The aesthetic quality of the result image.
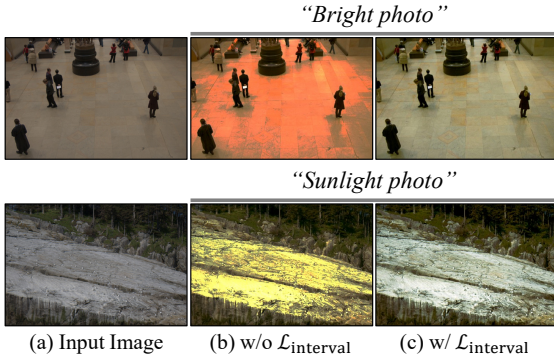

Figure 7. The effect of the sampling interval loss $\mathcal{L}_{\text{interval}}$. Without the loss, CLIPtone may perform excessive adjustments solely to align with the text descriptions, compromising aesthetic quality.

7. Conclusion

In this paper, we propose CLIPtone, the first unsupervised learning-based approach for text-based image tone adjustment. CLIPtone effectively extends a 3D LUT-based image enhancement model to accommodate a range of adjustments guided by text descriptions. Thanks to utilizing CLIP for perceptual supervision, CLIPtone requires no paired datasets, and also supports zero-shot prediction for unseen text descriptions. Through comprehensive experiments, we have validated the effectiveness and efficiency of CLIPtone.

Limitations. CLIPtone has some limitations. Firstly, the inherent bias in a pretrained CLIP model may lead CLIP-
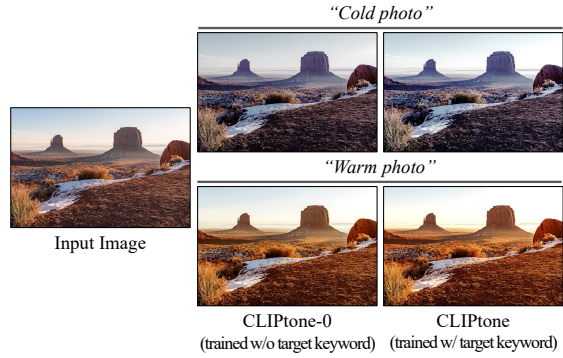

Figure 8. Qualitative comparison of CLIPtone-0 trained without target keywords and CLIPtone trained with target keywords. In both cases, both models make appropriate and similar adjustments implying that CLIPtone is capable of robustly adjusting even to novel descriptions not encountered during training.

Figure 9. Controlling the strength of adjustment by adjusting the modulation scaling factor $s \in \{0, 1, 2\}$. A result with the default scaling factor $s = 1$ is shown in the middle.

tone to perform adjustments that do not align with human perception. We include an analysis in the supplementary material. Secondly, CLIPtone relies global tone adjustment, thus cannot perform tone adjustment that requires local tone adjustment. Extending CLIPtone to support local adjustment could be a promising research avenue. Thirdly, as CLIPtone depends solely on the directional vectors from text descriptions, it cannot support adjustments for more diverse and specific styles, particularly those difficult to express with text alone. Leveraging a CLIP image encoder could be an interesting future direction.

References

[1] URL https://www.adobe.com/au/products/photoshop-lightroom-classic.html. 3, 7, 8

[2] URL https://github.com/meodai/color-names. 5

[3] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Restyle: A residual-based stylegan encoder via iterative refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6711–6720, 2021. 3

[4] Yuval Alaluf, Omer Tov, Ron Mokady, Rinon Gal, and Amit Bermano. Hyperstyle: Stylegan inversion with hypernetworks for real image editing. In *Proceedings of the IEEE/CVF conference on computer Vision and pattern recognition*, pages 18511–18521, 2022. 3, 4

[5] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022. 3

[6] David Bau, Alex Andonian, Audrey Cui, YeonHwan Park, Ali Jahanian, Aude Oliva, and Antonio Torralba. Paint by word. *arXiv preprint arXiv:2103.10951*, 2021. 3

[7] 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*, pages 18392–18402, 2023. 2, 3, 5, 6, 7

[8] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input/output image pairs. In *CVPR 2011*, pages 97–104. IEEE, 2011. 2, 3, 5, 6

[9] Yoav Chai, Raja Giryes, and Lior Wolf. Supervised and unsupervised learning of parameterized color enhancement. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 992–1000, 2020. 2

[10] Zheng Chang, Shuchen Weng, Peixuan Zhang, Yu Li, Si Li, and Boxin Shi. L-cad: Language-based colorization with any-level descriptions using diffusion priors. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 3

[11] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6306–6314, 2018. 2

[12] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. Vqgan-clip: Open domain image generation and editing with natural language guidance. In *European Conference on Computer Vision*, pages 88–105. Springer, 2022. 2, 3

[13] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 3

[14] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022. 2, 3, 5

[15] 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 (TOG)*, 36(4):1–12, 2017. 2

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2, 3

[17] Jingwen He, Yihao Liu, Yu Qiao, and Chao Dong. Conditional sequential modulation for efficient global image retouching. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 679–695. Springer, 2020. 2

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[19] Yuanming Hu, Hao He, Chenxi Xu, Baoyuan Wang, and Stephen Lin. Exposure: A white-box photo post-processing framework. *ACM Transactions on Graphics (TOG)*, 37(2): 1–17, 2018. 2

[20] Zhitong Huang, Nanxuan Zhao, and Jing Liao. Unicolor: A unified framework for multi-modal colorization with transformer. *ACM Transactions on Graphics (TOG)*, 41(6):1–16, 2022. 3

[21] Andrey Ignatov, Nikolay Kobyshev, Radu Timofte, Kenneth Vanhoey, and Luc Van Gool. Dslr-quality photos on mobile devices with deep convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 3277–3285, 2017. 2

[22] Andrey Ignatov, Nikolay Kobyshev, Radu Timofte, Kenneth Vanhoey, and Luc Van Gool. Wespe: weakly supervised photo enhancer for digital cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 691–700, 2018. 2

[23] Wentao Jiang, Ning Xu, Jiayun Wang, Chen Gao, Jing Shi, Zhe Lin, and Si Liu. Language-guided global image editing via cross-modal cyclic mechanism. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2115–2124, 2021. 2, 3

[24] Kyoungkook Kang, Seongtae Kim, and Sunghyun Cho. Gan inversion for out-of-range images with geometric transformations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13941–13949, 2021. 3

[25] Hanul Kim, Su-Min Choi, Chang-Su Kim, and Yeong Jun Koh. Representative color transform for image enhancement. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4459–4468, 2021. 2

[26] Han-Ul Kim, Young Jun Koh, and Chang-Su Kim. Global and local enhancement networks for paired and unpaired image enhancement. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pages 339–354. Springer, 2020. 2

[27] Seongtae Kim, Kyoungkook Kang, Geonung Kim, Seung-Hwan Baek, and Sunghyun Cho. Dynagan: Dynamic few-shot adaptation of gans to multiple domains. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–8, 2022. 3

[28] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[29] Gihyun Kwon and Jong Chul Ye. Clipstyler: Image style transfer with a single text condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18062–18071, 2022. 5, 6, 7

[30] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip HS Torr. Manigan: Text-guided image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7880–7889, 2020. 3

[31] Jie Liang, Hui Zeng, Miaomiao Cui, Xuansong Xie, and Lei Zhang. Ppr10k: A large-scale portrait photo retouching dataset with human-region mask and group-level consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 653–661, 2021. 2

[32] Chengxu Liu, Huan Yang, Jianlong Fu, and Xueming Qian. 4d lut: learnable context-aware 4d lookup table for image enhancement. *IEEE Transactions on Image Processing*, 32: 4742–4756, 2023. 2

[33] Rabeeh Karimi Mahabadi, Sebastian Ruder, Mostafa Dehghani, and James Henderson. Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks. *arXiv preprint arXiv:2106.04489*, 2021. 3

[34] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 2

[35] Sean Moran, Pierre Marza, Steven McDonagh, Sarah Parisot, and Gregory Slabaugh. Deeplpf: Deep local parametric filters for image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12826–12835, 2020. 2

[36] Sean Moran, Steven McDonagh, and Gregory Slabaugh. Curl: Neural curve layers for global image enhancement. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 9796–9803. IEEE, 2021. 2

[37] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021. 3

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3

[39] Neale Ratzlaff and Li Fuxin. Hypergan: A generative model for diverse, performant neural networks. In *International Conference on Machine Learning*, pages 5361–5369. PMLR, 2019. 3

[40] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2287–2296, 2021. 3

[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3

[42] Jing Shi, Ning Xu, Yihang Xu, Trung Bui, Franck Dernoncourt, and Chenliang Xu. Learning by planning: Language-guided global image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13590–13599, 2021. 2, 3, 5, 6, 7

[43] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 3

[44] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 2020. 3

[45] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4): 1–14, 2021. 3

[46] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2555–2563, 2023. 3

[47] Tao Wang, Yong Li, Jingyang Peng, Yipeng Ma, Xian Wang, Fenglong Song, and Youliang Yan. Real-time image enhancer via learnable spatial-aware 3d lookup tables. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2471–2480, 2021. 2

[48] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2256–2265, 2021. 3

[49] Zhicheng Yan, Hao Zhang, Baoyuan Wang, Sylvain Paris, and Yizhou Yu. Automatic photo adjustment using deep neural networks. *ACM Transactions on Graphics (TOG)*, 35(2): 1–15, 2016. 2

[50] Canqian Yang, Meiguang Jin, Xu Jia, Yi Xu, and Ying Chen. Adaint: Learning adaptive intervals for 3d lookup tables on real-time image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17522–17531, 2022. 2, 3, 4, 5

[51] Canqian Yang, Meiguang Jin, Yi Xu, Rui Zhang, Ying Chen, and Huaida Liu. Seplut: Separable image-adaptive lookup tables for real-time image enhancement. In *European Conference on Computer Vision*, pages 201–217. Springer, 2022.

[52] 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*, 44(4):2058–2073, 2020. 2, 5

[53] Fengyi Zhang, Hui Zeng, Tianjun Zhang, and Lin Zhang. Clut-net: Learning adaptively compressed representations of 3dluts for lightweight image enhancement. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6493–6501, 2022. 2

[54] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 3

[55] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 2