Image Harmonization with Transformer

Zonghui Guo[1] Dongsheng Guo[1] Haiyong Zheng[1,*] Zhaorui Gu[1] Bing Zheng[1,2] Junyu Dong[3]
[1]Underwater Vision Lab (http://ouc.ai), Ocean University of China
[2]Sanya Oceanographic Institution, Ocean University of China
[3]College of Computer Science and Technology, Ocean University of China

Abstract

Image harmonization, aiming to make composite images look more realistic, is an important and challenging task. The composite, synthesized by combining foreground from one image with background from another image, inevitably suffers from the issue of inharmonious appearance caused by distinct imaging conditions, i.e., lights. Current solutions mainly adopt an encoder-decoder architecture with convolutional neural network (CNN) to capture the context of composite images, trying to understand what it looks like in the surrounding background near the foreground. In this work, we seek to solve image harmonization with Transformer, by leveraging its powerful ability of modeling long-range context dependencies, for adjusting foreground light to make it compatible with background light while keeping structure and semantics unchanged. We present the design of our harmonization Transformer frameworks without and with disentanglement, as well as comprehensive experiments and ablation study, demonstrating the power of Transformer and investigating the Transformer for vision. Our method achieves state-of-the-art performance on both image harmonization and image inpainting/enhancement, indicating its superiority. Our code and models are available at https://github.com/zhenglab/HarmonyTransformer.

1. Introduction

Combining regions of different photographs into a realistic composite is a fundamental problem in many vision and graphics applications, such as image compositing, mosaicing, editing, and scene completion [33]. However, the composite, synthesized by combining the foreground from one image with the background from another image, inevitably suffers from the issue of inharmonious appearance between foreground and background caused by distinct imaging conditions (*e.g.*, day and night, sunny and cloudy, outdoors and indoors). Therefore, making the composite look more realistic, namely image harmonization, is an important and challenging task [33, 35, 10, 9].

Figure 1. We create two composite images about pig that flies (top) and tiger at party (bottom), also show harmonization comparison between state-of-the-art DoveNet [9] and our method.

Image harmonization aims to adjust the foreground to make it compatible with the background on the appearance. Essentially, the appearance of a natural image depends on various factors in the scene, such as illumination, material, and shape [41, 1]. For a composite image, the foreground and the background are considered semantically harmonious, although sometimes it might be impractical or unreasonable (*e.g.*, pig that flies and tiger at party in Figure 1). Thus, the inharmony of composite images is mainly caused by the distinct lights in the different scenes between foreground and background while imaging, for instance, a tiger captured in the wild under natural light as foreground and the party captured in a hall under artificial lighting as background, yielding inharmonious color appearance because an object appears coloured due to the way it interacts with light. Hence, adjusting the foreground color to make it compatible with the background color, while keeping the structure and semantics unchanged, are crucial and essential for harmonizing composite images.

Traditional harmonization methods have focused on better matching techniques to ensure consistent appearance between foreground and background, by transferring handcrafted statistics such as color and texture [39, 33]. Recently, deep harmonization models and large-scale datasets have been developed to address this challenging task [35,

*Corresponding author: Haiyong Zheng (zhenghaiyong@ouc.edu.cn).
This work was supported by the National Natural Science Foundation of China under Grant Nos. 61771440 and 41776113.

10, 9], achieving better performance benefiting from deep model and big data. Current deep models mainly adopt an encoder-decoder CNN architecture, which employs an encoder to capture the context of the composite image and a decoder to reconstruct the harmonized image, trying to understand what it looks like in the surrounding background region near the foreground region.

Actually, the encoder-decoder CNN tackle image harmonization with a two-stage process: harmonizing foreground with background and reconstructing the harmonized image. Essentially, the first stage works on adjusting foreground color with background color to make them compatible, while the second stage devotes to recovering original structure and semantics. However, since CNN inherently has the inductive bias of locality, a shallow CNN can only capture the context of surrounding background near the foreground, and without global background context, it might be not enough for better adjustment to make the color of foreground and background consistent. Besides, previous methods adopt U-Net [32] with successive contraction, which has the ability to capture global context, but the inharmony might be introduced again to reconstruction via skip connections from encoder to decoder as a side effect.

Recently, Transformer [36] won renown as a new type of neural network, which can capture long-range context dependencies, thanks to the self-attention design. Instead of RNN and LSTM, Transformer was first applied to natural language processing (NLP) tasks where it achieved significant improvements [36, 12, 4]. Nowadays Transformer is also showing it is a viable alternative to CNN by being applied to computer vision (CV) tasks, such as object detection [5, 43], image recognition [14], and image processing [6]. Thus, in this work, we seek to solve image harmonization with Transformer, by leveraging its powerful ability of modeling long-range context, to satisfy the requirement of harmonization on capturing global context.

Inspired by the observation that adjusting light plays a key role in harmonizing images [16], we move one step forward. Based on intrinsic image [2] and Retinex theory [26, 25] with the assumption of ideal Lambertian surface, the light intensity values represented in an image actually encode all the characteristics of corresponding scene points, thus, in order to adjust the light of composite images, it is intuitive to separate material-dependent reflectance for light-dependent illumination re-rendering with disentangled background light for better harmonization. Therefore, in our work, we further devise to harmonize composite images by capturing the "light" from the background and put it on the "material" via disentangled harmonization Transformer.

Our contributions include: (1) we design and build the first harmonization Transformer frameworks without and with disentangled representation; (2) we explore and analyze the harmonization Transformer in the aspects of input, encoder/decoder, head, and layer; (3) we present comprehensive experiments to show the efficacy of both Transformer and disentanglement, achieving performance substantially better than previous methods on image harmonization; (4) we illustrate the utility of our framework in two extra vision tasks, *i.e*., image inpainting and image enhancement, both producing very competitive results.

2. Related Work

2.1. Image Harmonization

Early contributions in image harmonization have focused on using low-level image representations in color space to adjust foreground to background appearance, including color distribution matching [30, 31, 8], multi-scale statistics [33], and gradient-based methods [20, 29, 34]. Further studies have attended to assess and improve the realism of images [24, 38], among which Zhu *et al*. [42] fit a CNN model that distinguishes natural photographs from automatically generated composite images and adjusts color of composites by optimizing predicted visual realism score.

Recently, CNN models have been developed for end-to-end image harmonization. Tsai *et al*. [35] exploited encoder-decoder structure with skip connections to capture context and semantic information of composite images for harmonization. Cun *et al*. [10] also went with an encoder-decoder U-Net backbone equipped with an additional spatial-separated attention module to learn regional appearance changes in low-level features. Cong *et al*. [9] employed an attention enhanced U-Net generator with a global discriminator and a domain verification discriminator to transform foreground domain to background domain. Different from all existing methods, we devote to solve image harmonization with Transformer.

2.2. Vision Transformer

Transformer [36], first applied to NLP tasks [12, 4], is a new type of neural network mainly based on self-attention mechanism. Due to its strong representation capabilities, researchers are recently looking at ways to employ Transformer to CV tasks [17, 21]. Chen *et al*. [7] trained a sequence Transformer (iGPT) to auto-regressively predict pixels, achieving results comparable with CNNs on image classification. Dosovitskiy *et al*. [14] applied a pure Transformer directly to sequences of image patches (ViT) attaining excellent results compared to state-of-the-art CNNs. Carion *et al*. [5] redesigned the framework of object detection with Transformer (DETR) by treating the object detection task as an intuitive set prediction problem, opening up a new avenue to object detection [43, 11]. Besides, Transformer has been utilized to address a variety of other CV problems, including image processing [6], pose estimation [18], video inpainting [40]. Our work also contributes

to the study of vision Transformer, diving deeper into Transformer for image harmonization and beyond.

3. Methods

This work seeks to leverage Transformer for image harmonization, thus, we first analyze how to employ Transformer for vision, and then present our harmonization Transformer and disentangled harmonization Transformer.

3.1. Transformer for Vision

Image Input. Transformer is designed to handle sequential data, such as natural language, for tasks like translation, eschewing recurrence and instead relying entirely on an attention mechanism to draw global dependencies between input and output. Thus, to use Transformer for vision, we need to formulate 2D image to 1D sequence with tokens (words in NLP) and their embeddings as input. Actually we can tokenize an image into patches as tokens in order to avoid very long sequence with pixels as tokens. In this work, we preliminarily analyze the impact of different token numbers as well as different embedding types on the performance of Transformer in image harmonization. For token number, we consider to use different strides for adjustment while splitting image into patches. For embedding type, we consider to adopt linear (FC or CONV) and nonlinear (MLP or CNN with nonlinear activation function) projections. We empirically find that harmonization Transformer might be sensitive to token number while not sensitive to embedding type, see "Transformer Input" of Section 4.3 for analysis. We illustrate the image input pattern in Figure 2.

Transformer Encoder/Decoder. Transformer body contains an encoder $TRE(\cdot)$ to capture relations, and a decoder $TRD(\cdot)$ to produce outputs towards the task. TRE is composed of a stack of identical layers, where each layer has a multi-head self-attention sub-layer and a feed forward network sub-layer. TRD is also composed of a stack of identical layers, where each layer, in addition to the two sub-layers in each encoder layer, has a third encoder-decoder attention sub-layer that performs multi-head attention over the output of encoder stack. We can see that TRE employs self-attention to explore self-relation of its input, while TRD performs cross-attention to discover cross-relation between its input and encoder output. Thus, for an image input, TRE aims at outputting self-attention maps that encode the dependencies among input tokens (image patches), and TRD devotes to producing the mapping from source domain (TRE input) to target domain (TRD input/output). In this work, we investigate the efficacy of TRE and TRD on image harmonization task, as well as effect analysis of different heads and layers, see Section 4.3.

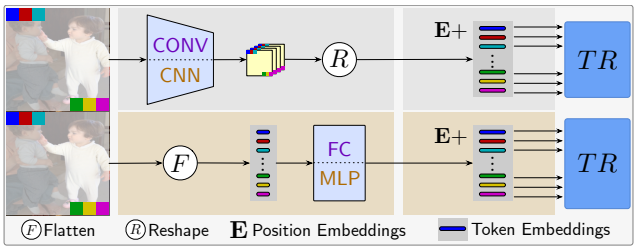

Figure 2. The image input pattern of using Transformer for vision.

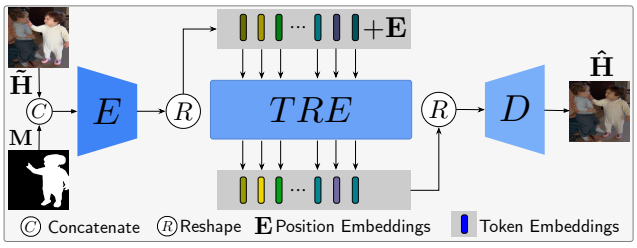

Figure 3. Our harmonization Transformer (HT) framework contains a convolutional encoder-decoder (E-D) involving a Transformer encoder (TRE) inside.

3.2. Harmonization Transformer

In order to eliminate the color inharmony caused by different light between foreground and background, we first design a simple basic harmonization Transformer framework, employing Transformer in a very basic convolutional encoder-decoder architecture, as shown in Figure 3.

The CNN encoder E and decoder D are responsible for compressing the input image to compact feature representation as Transformer input and reconstructing the Transformer output back to harmonized image, respectively. In this way, we actually utilize CNN embedding for Transformer under a basic encoder-decoder architecture. Noting that, for harmonization task with many information of input image unchanged, TRE and TRD can be considered to play similar roles in harmonization relying on self-attention, thus we only use TRE in our framework, see "Transformer Encoder/Decoder" of Section 4.3 for analysis.

Formally, given a composite image $\tilde{\mathbf{H}}$ and a foreground mask $\mathbf{M}$ which indicates the inharmonious region as input, our goal is to produce a harmonized image $\hat{\mathbf{H}}$ as output, where $\hat{\mathbf{H}}$ is expected to be as close to real image $\mathbf{H}$ as possible. Specifically, CNN encoder $E(\cdot)$ generates a lower-resolution feature map $\mathbf{F} \in \mathbb{R}^{h \times w \times c}$, where we use $h = \frac{H}{4}$, $w = \frac{W}{4}$, and $c = 256$. Then we reshape $\mathbf{F}$ to sequence $\mathbf{F}' \in \mathbb{R}^{hw \times c}$, with pixels (corresponding to image patches) as TRE input tokens and channel aggregation of each pixel as token embedding, also add with fixed position embeddings $\mathbf{E}$ using the sinusoidal version of vanilla Transformer [36]. Further we inversely reshape the output sequence of TRE back to feature map with the same size as $\mathbf{F}$, and feed it into CNN decoder $D(\cdot)$ to attain harmonized

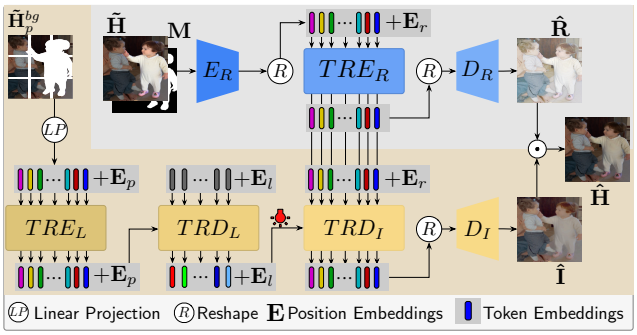

Figure 4. Our disentangled harmonization Transformer (D-HT) framework is a dual-pathway architecture for separating the composite image into pseudo-reflectance and pseudo-illumination intrinsic images. Refer to Section 3.3 for more details.

result $\hat{\mathbf{H}}$. We formulate the whole process as:

$$\hat{\mathbf{H}} = D\Big[\phi'\Big(TRE\big[\phi(E(\tilde{\mathbf{H}}, \mathbf{M})) + \mathbf{E}\big]\Big)\Big], \tag{1}$$

where ϕ and ϕ' represent reshape and inverse reshape operations respectively.

It is also noteworthy that, we only use a single $\mathcal{L}_1$ loss to encourage $\hat{\mathbf{H}} \approx \mathbf{H}$:

$$\mathcal{L}_1 = \mathbb{E}_{(\hat{\mathbf{H}},\mathbf{H})}\Big[\|\hat{\mathbf{H}} - \mathbf{H}\|_1\Big]. \tag{2}$$

3.3. Disentangled Harmonization Transformer

A further idea for better harmonizing composite images, according to intrinsic image [2] and Retinex theory [26, 25], is to separate light-dependent illumination and material-dependent reflectance [16]. Therefore, we then devise a dual-pathway framework for image harmonization, by separating the composite image into pseudo-reflectance and pseudo-illumination intrinsic images [1], and devote to disentangling the light from background and put it on reflectance for harmonization. Particularly, we employ Transformer in both pathways, to leverage its advantages of long-range dependencies learning for better harmonization.

We illustrate our framework of disentangled harmonization Transformer in Figure 4. The pseudo-reflectance pathway (top) is similar to the structure of harmonization Transformer (Section 3.2), since the output pseudo-reflectance $\hat{\mathbf{R}}$ can also be regarded as an image-to-image transformation. While the pseudo-illumination pathway (bottom) is quite different, where we intend to map the background image space to light latent space, and we choose FC embedding on patches of input masked composite image without overlapping, to acquire tokens with embeddings and position embeddings as TRE_L input, then we utilize a TRD_L connected to TRE_L with an initial zero light code and learnable light position as its input, to produce the background light code, finally we impose this light on the output pseudo-reflectance of TRE_R by employing a TRD_I, yielding the pseudo-illumination $\hat{\mathbf{I}}$. The harmonized image $\hat{\mathbf{H}}$ can finally be obtained via $\hat{\mathbf{H}} = \hat{\mathbf{R}} \odot \hat{\mathbf{I}}$ ($\odot$ is element-wise product) based on Retinex theory.

Formally, we first split background $\tilde{\mathbf{H}}^{bg} \in \mathbb{R}^{H\times W\times C}$ (channel number $C = 3$) into patch sequence $\tilde{\mathbf{H}}_p^{bg} \in \mathbb{R}^{T\times(P^2\cdot C)}$ (patch number $T = \frac{HW}{P^2}$, and we use patch size $P = 8$), then we flatten each patch (as token) and expand it to $C' = 256$ dimensions as its embedding though a linear projection $LP(\cdot)$. We also add fixed position embeddings $\mathbf{E}_p$ to token embeddings and feed them into $TRE_L(\cdot)$. And we further employ $TRD_L(\cdot)$ to receive $TRE_L(\cdot)$ output and light tokens $t_l \in \mathbb{R}^{d_l\times C'}$ (we set $d_l = 27$ referring to the 27 dimensional spherical harmonic coefficients of the lighting) with learned light position embeddings $\mathbf{E}_l$ as input, producing background light code $l^{bg} \in \mathbb{R}^{d_l\times C'}$ as output. Note that we use light code to represent light in latent space, and light token to represent input of corresponding Transformer. This process can be represented as:

$$l^{bg} = TRD_L\Big[TRE_L\big(LP(\tilde{\mathbf{H}}_p^{bg}) + \mathbf{E}_p\big), t_l + \mathbf{E}_l\Big]. \tag{3}$$

Moreover, we employ $TRD_I(\cdot)$ receiving background light tokens $t_l \in \mathbb{R}^{d_l\times C'}$ (l^{bg}) and pseudo-reflectance tokens $t_r \in \mathbb{R}^{hw\times c}$ from TRE_R with their corresponding position embeddings $\mathbf{E}_l$ and $\mathbf{E}_r$ as input, to produce pseudo-illumination tokens that will be reshaped and decoded by $D_I(\cdot)$, yielding the harmonized pseudo-illumination $\hat{\mathbf{I}}$:

$$\hat{\mathbf{I}} = D_I\Big[\phi'\big(TRD_I(t_l + \mathbf{E}_l, t_r + \mathbf{E}_r)\big)\Big]. \tag{4}$$

The harmonized pseudo-reflectance $\hat{\mathbf{R}}$ can be attained via Equation 1. So final harmonized image will be $\hat{\mathbf{H}} = \hat{\mathbf{R}} \odot \hat{\mathbf{I}}$. The only loss used is also a single $\mathcal{L}_1$ loss (Equation 2).

Overall, we devise to employ two encoders and two decoders of Transformer, where TRE_R receives patch CNN embeddings and produces pseudo-reflectance, yet TRE_L receives patch FC embeddings and produces output for TRD_L to capture background light, while TRD_I receives background light and pseudo-reflectance tokens from TRE_R to produce pseudo-illumination, and finally we combine pseudo-reflectance and pseudo-illumination to yield harmonization. We hope our work can provide meaningful reference for better harnessing vision Transformer.

4. Experiments on Image Harmonization

4.1. Datasets and Metrics

Synthesized iHarmony4 Dataset. We conduct experiments on public synthesized iHarmony4 dataset [9] to analyze and evaluate our harmonization Transformers on image

[1] We add "pseudo-" prefix to indicate that they are actually not physically-based but relative reflectance and illumination.

Dataset	Metric	Composite	E-D (U-Net)	E-D (CNN)	DIH [35]	S^{2}AM [10]	DoveNet [9]	Ours (HT)	Ours (D-HC)	Ours (D-HT)
HCOCO	PSNR↑	33.99	34.94	35.58	33.59	35.09	35.83	**37.87**	36.85	**38.76**
	fPSNR↑	19.86	21.66	21.73	20.67	22.45	22.48	**24.24**	23.11	**25.27**
	MSE↓	69.37	41.54	40.92	56.17	35.65	34.26	**20.99**	29.84	**16.89**
	fMSE↓	996.59	684.33	627.33	798.99	542.06	551.01	**377.11**	468.68	**299.30**
HAdobe5k	PSNR↑	28.52	33.72	34.58	32.36	34.23	35.13	**36.10**	35.08	**36.88**
	fPSNR↑	17.52	23.52	24.04	22.36	24.28	25.19	**25.80**	24.67	**26.78**
	MSE↓	345.54	72.09	66.46	94.89	53.93	56.86	**47.96**	64.35	**38.53**
	fMSE↓	2051.61	508.53	435.16	593.03	404.62	380.39	**321.14**	390.57	**265.11**
HFlickr	PSNR↑	28.43	30.11	29.98	29.08	30.53	30.75	**32.37**	31.30	**33.13**
	fPSNR↑	18.09	20.16	19.76	19.31	20.89	20.76	**22.25**	21.11	**23.06**
	MSE↓	264.35	135.16	156.62	168.35	123.36	125.85	**88.41**	109.60	**74.51**
	fMSE↓	1574.37	945.14	1002.23	1099.13	785.65	827.03	**617.26**	733.46	**515.45**
Hday2night	PSNR↑	34.36	34.17	34.50	33.59	34.48	34.87	36.38	**36.54**	**37.10**
	fPSNR↑	19.14	19.86	19.64	19.74	20.51	20.63	21.68	**21.86**	**22.51**
	MSE↓	109.65	62.60	95.79	86.25	54.39	57.17	58.14	**52.64**	**53.01**
	fMSE↓	1409.98	1114.96	1321.89	1129.40	989.07	1075.71	823.68	**716.04**	**704.42**
All	PSNR↑	31.78	34.03	34.64	32.73	34.32	35.04	**36.71**	35.71	**37.55**
	fPSNR↑	18.97	22.00	22.15	20.99	22.77	23.04	**24.43**	23.32	**25.41**
	MSE↓	172.47	61.30	62.29	80.55	51.13	51.51	**37.07**	49.24	**30.30**
	fMSE↓	1376.42	669.94	625.67	778.41	537.23	541.53	**395.66**	479.94	**320.78**

Note: we train DIH and S^{2}AM yet use pre-trained DoveNet to obtain the results for comparison.

Table 1. Quantitative comparison across four sub-datasets of iHarmony4 [9]. ↑ indicates the higher the better, and ↓ indicates the lower the better. **Bold** means the best, and **bold** means the next best. E-D means encoder-decoder, and HT represents our harmonization Transformer, while D-HC and D-HT denote our disentangled harmonization framework with CNN and Transformer respectively.

harmonization. iHarmony4 is composed of 4 sub-datasets: HCOCO, HAdobe5k, HFlickr, and Hday2night, each of which includes synthesized composite images, foreground masks of composite images, and corresponding real images. We follow the same settings of this dataset as DoveNet [9].

Real Composite Images. Following [35, 10, 9], we also evaluate our method on 99 real composite images used by [35] for subjective evaluation.

Objective Evaluation Metrics. Following [35, 9], we use Mean Squared Error (MSE) and Peak Signal-to-Noise Ratio (PSNR) as evaluation metrics. However, for image harmonization task, it is more suitable and more accurate to calculate the difference only in the foreground region due to unchanged background [9], thus we also report foreground MSE (fMSE) and foreground PSNR (fPSNR) as better metrics, measuring how well the foreground is harmonized. Noting that, we calculate fMSE and corresponding fPSNR over each single image and then take average across the dataset, so that they can be regarded as a better indicator in evaluating harmonization generalization ability of the method. Whereas, we argue that MSE and PSNR essentially measure the average errors over all pixels across the dataset, thus are not very suitable for tasks like harmonization with a number of pixels (background) unchanged. In our experiments, we use fMSE as the main metric.

Subjective Evaluation Metric. We invite 60 subjects to participate in user study and acquire a total of 29700 pairwise results for all 99 images, with 30 results for each pair of different methods on average. All subjects are not aware of image harmonization task, and are only required to select the visually better one corresponding to better method for each pair, then we record how many times one method is selected in each pair on all 99 images as the statistics for pairwise comparison of Bradley-Terry (B-T) model [3, 23], to calculate global ranking score for each method.

4.2. Implementation Details

We train all our models with only a single $\mathcal{L}_1$ loss, using Adam optimizer [22] with parameters of $\beta_1 = 0.5$, $\beta_2 = 0.999$ for total 60 epochs. Initial learning rate is set as e^{-4} and decayed to e^{-5} after 40 epochs. The final activation function is $\tanh$ for harmonized image $\hat{\mathbf{H}}$ in Section 3.2, pseudo intrinsic images $\hat{\mathbf{R}}$ and $\hat{\mathbf{I}}$ in Section 3.3. We resize input images as 256×256 for training and testing, and our models produce harmonized images with the same size. Specially, output pseudo-reflectance and pseudo-illumination are normalized to $[0, 1]$ to recover $\hat{\mathbf{H}}$. All our model architectures and details are in *supplementary file*.

4.3. Harmonization Transformer

Baseline and Comparison. For comparison, we first construct an encoder-decoder U-Net (E-D U-Net) and a basic encoder-decoder CNN (E-D CNN with Encoder-ResBlocks-Decoder structure) as baselines. Table 1 shows quantitative comparison of image harmonization across four sub-datasets of iHarmony4, comparing our harmonization Transformer (HT with 2-head and 9-layer TRE, Figure 3) with baselines and state-of-the-art methods: DIH [35], S^{2}AM [10] and DoveNet [9]. Besides, we also provide evaluation results of composite images as reference.

As can be seen, compared to E-D U-Net, E-D CNN performs better on HCOCO and HAdobe5K while worse on

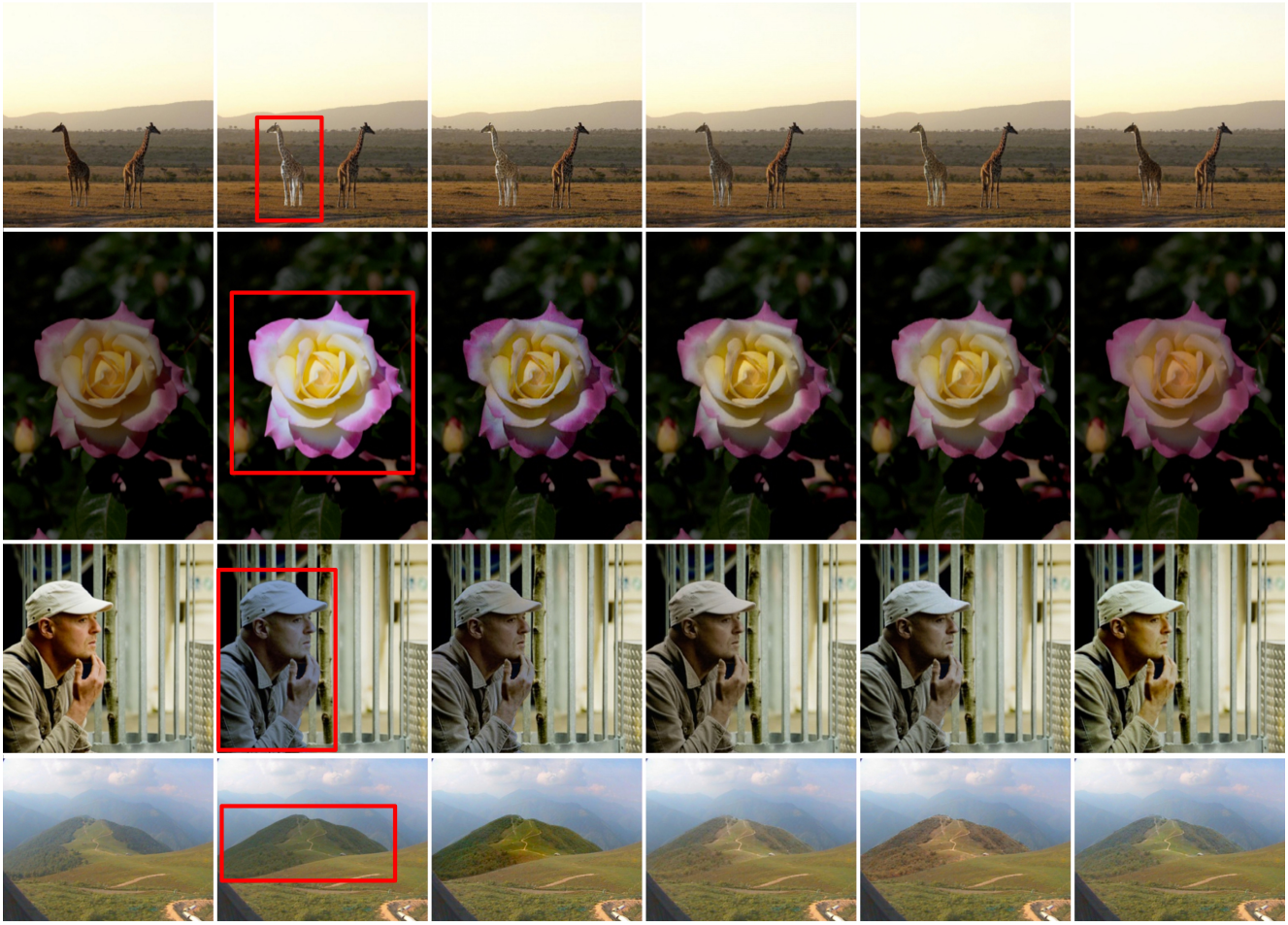

Figure 5. Qualitative comparison across four sub-datasets of iHarmony4 [9] (one example for each dataset). From top to bottom: HCOCO, HAdobe5k, HFlickr, and Hday2night. Red boxes in composite images mark foreground.

HFlickr and Hday2night, and the reason might be that, U-Net has global receptive field to capture global context but its skip connections may bring in inharmony to reconstruction, while CNN usually has limited receptive field due to its inductive bias of locality. In summary, CNN works better than U-Net with lower fMSE on the whole dataset. But our simple HT model outperforms not only the baselines but also the state-of-the-arts, indicating the efficacy of Transformer for modeling long-range context on harmonization.

MSE vs. fMSE. It is worth mentioning that our HT model is superior to S^2AM in fMSE, but inferior to S^2AM in MSE on Hday2night, mainly because that MSE evaluates harmonization performance at dataset level while fMSE reflects harmonization ability at image level which is more valuable and generalized, for instance, one method may obtain lower MSE yet higher fMSE because it harmonizes some images with big foreground very better while harmonizes some images with small foreground very worse, demonstrating unstable performance.

Transformer Input. We then conduct ablation study to investigate the impact of token number and embedding type with respect to Transformer performance based on the structure shown in Figure 2, where we use a 1-head and 3-layer TRE for TR following by a CNN decoder for reconstruction. We use stride S to adjust token numbers T. Table 2 presents that the performance is continuously improved with the increase of token number ($N \Rightarrow 4N \Rightarrow 16N$) for both linear and nonlinear token embedding. Besides, for a fixed token number, *e.g.* $4N$, the performance is similar, no matter which embedding type (linear FC or CONV, or nonlinear MLP or CNN) we choose. Thus we can speculate that Transformer performance might be sensitive to token number while insensitive to embedding type on harmonization. This makes sense that, Transformer can mine richer context if we provide long sequence with more tokens even redundancy may exist (overlapping patches), and the current different embedding methods can provide effective information for image patches so that they may not matter.

Embedding \ Token	$T=N$ $S=8$	$T\approx 4N$ $S=4$	$T\approx 16N$ $S=2$
FC	611.25	522.84	447.64
CONV	610.01	524.87	446.54
MLP	596.05	514.19	440.76
CNN	598.17	520.19	443.98

Table 2. Quantitative comparison of using different token numbers T adjusted by stride S, and embedding types (linear FC/CONV and nonlinear MLP/CNN) on fMSE↓.

6×layer		9×layer		12×layer	
$E(3)$+$D(3)$	$E(6)$	$E(3)$+$D(6)$	$E(9)$	$E(3)$+$D(9)$	$E(12)$
451.80	459.47	403.76	415.60	426.56	419.08

Table 3. Quantitative comparison of using different Transformer encoder (E) and decoder (D) layer numbers on fMSE↓.

	3×layer	6×layer	9×layer	12×layer
1×head	502.37	459.47	415.60	419.08
2×head	479.53	450.14	395.66	400.11
4×head	461.22	406.99	392.74	397.37

Table 4. Quantitative comparison of using different Transformer layer numbers and attention heads in HT model on fMSE↓.

Transformer Encoder/Decoder. We further design experiments to validate the effect of Transformer encoder and decoder layer numbers on harmonization based on the HT structure (Figure 3). Table 3 demonstrates that the performance is similar if encoder layer number is equal to total layer number of encoder and decoder, although the decoder has an extra attention sub-layer. Therefore, in our HT model, we only use the encoder TRE.

Transformer Head and Layer. We lastly conduct ablative experiments to analyze the impact of using different Transformer layer numbers and attention heads on harmonization with our HT model (Figure 3). Table 4 tells us that, both more layers and more heads are helpful for improving performance, but if we use more than 9 layers, the room for performance improvement will be limited.

4.4. Disentangled Harmonization Transformer

Comparison. We move on to our disentangled harmonization framework, where we build two disentangled models of using CNN (D-HC) and Transformer (D-HT with 2-head 9-layer TRE and TRD, Figure 3) respectively. To validate the effectiveness of our disentanglement, we construct D-HC model by replacing TRE_R with ResBlocks, TRE_L and TRD_L with Encoder and MLP, TRD_I with AdaIN [19] in D-HT model. Table 1 shows that, D-HC model achieves competitive or superior results compared to state-of-the-art methods, indicating that the disentanglement does contribute to harmonization. Also our D-HT model performs the best with a very low fMSE (320.78 vs. 537.23 of S^2AM and 541.53 of DoveNet). Note that D-HC outperforms HT on Hday2night, probably due to better harmonization ability of disentanglement (D-HC) and insufficient training data (only 311 images) for Transformer (HT) that lacks inductive bias.

Method	PSNR↑	fPSNR↑	MSE↓	fMSE↓
R_{CNN}+I_{CNN}	35.71	23.32	49.24	479.94
R_{TRE}+I_{CNN}	37.17	24.96	31.99	352.55
R_{CNN}+I_{TR}	37.26	25.07	32.22	348.80
R_{TRE}+I_{TR}	**37.55**	**25.41**	**30.30**	**320.78**

Table 5. Ablation study on our disentangled harmonization.

Figure 6. Image harmonization visual results with normal masks (middle row) and inverted masks (bottom row) on composite images (top row). Red boxes mark foreground of normal masks.

Analysis of Disentanglement. We conduct ablation study on our D-HT model, by replacing one pathway of reflectance (R) and illumination (I) with that (CNN) in D-HC model, resulting in four variants listed in Table 5, and the results show the strength of Transformer on harmonization.

Moreover, we design an additional experiment by inverting the normal masks, that is, exchanging foreground and background to yield inverted masks, so that our D-HT model tries to harmonize background according to foreground. Figure 6 presents harmonized results with normal masks (middle row) and inverted masks (bottom row) for contrast, indicating that D-HT can produce promising harmonized outputs from arbitrary foreground masks.

Analysis of Light. We then walk in light latent space to see if Transformer can learn relevant light representation. Given an image, we use D-HT model to obtain its light latent code, and then change it arbitrarily to produce results by recovery. Figure 7 illustrates examples with outputs under different light conditions, indicating efficacy of our design.

We further conduct experiment to employ D-HT model for transferring light from source images to target image. We change light code of target image by interpolating light latent codes of two source images for producing results shown in Figure 8, which demonstrate the light latent space.

4.5. Real Composite Image Harmonization

We also evaluate D-HT on real composite image harmonization compared with state-of-the-arts. Table 6 and Figure 9 demonstrate that our method achieves best performance with highest B-T score and best visual effect.

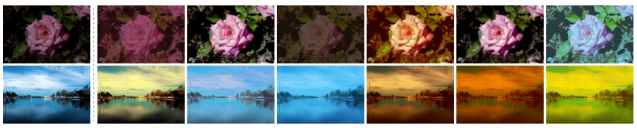

Figure 7. Changing light latent code of an image (left) from Transformer produces different results in different lighting conditions.

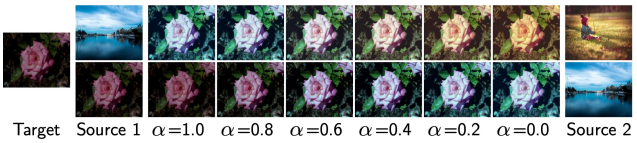

Figure 8. Changing light latent code of target image (L_t) produces different results, by interpolating light latent codes of two source images (L_{s1} and L_{s2}) with $L_t = \alpha L_{s1} + (1-\alpha)L_{s2}$.

Method	Composite	DIH [35]	S^{2}AM [10]	DoveNet [9]	Ours
B-T score↑	0.623	0.831	0.874	1.032	**2.248**

Table 6. User study comparison on 99 real composite images.

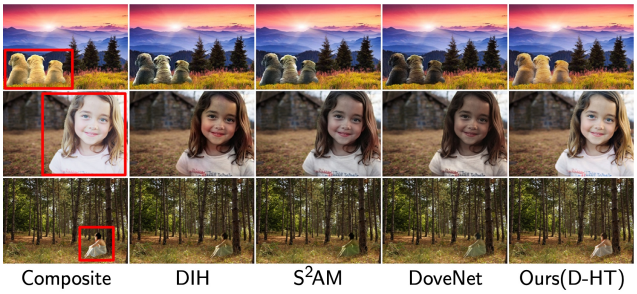

Figure 9. Visual comparison to harmonize real composite images.

5. Beyond Image Harmonization

5.1. Image Inpainting

We apply our HT model to free-form image inpainting task on Paris StreetView dataset [13], compared to state-of-the-art RFR-Net [27]. Image inpainting aims to fill missing pixels of an image, by synthesizing visually realistic and semantically plausible pixels that are coherent with existing ones. Table 7 and Figure 10 present superior performance of our HT model (with the same losses as RFR-Net), by giving full play to the advantage of Transformer in modeling long-term correlations between distant contextual information and the missing hole.

5.2. Image Enhancement

We also employ our D-HT model to image enhancement task on MIT-Adobe-5K-UPE dataset [37], compared to state-of-the-art DeepLPF [28]. Insufficient lighting while imaging results in degraded images, especially underexposed photos. Thus we use D-HT model to decompose observed images into reflectance and illumination via an extra reconstruction loss, and simply treat the reflectance as the final enhanced results refer to [15]. In this experiment, we

Method	ℓ_1 err.↓	PSNR↑	SSIM↑
RFR-Net [27]	0.028	28.42	0.8920
Ours (HT)	**0.021**	**29.55**	**0.9047**

Table 7. Quantitative comparison of image inpainting on Paris StreetView [13].

Figure 10. Visual comparison of image inpainting on Paris StreetView [13].

Method	PSNR↑	SSIM↑	LPIPS↓
DeepLPF [28]	23.00	0.726	0.050
Ours	**24.22**	**0.810**	**0.036**

Table 8. Quantitative comparison of image enhancement on MIT-Adobe-5K-UPE [37].

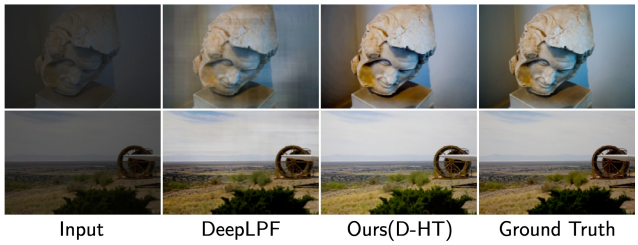

Figure 11. Visual comparison of image enhancement on MIT-Adobe-5K-UPE [37].

retrained DeepLPF model on MIT-Adobe-5K-UPE to obtain results for comparison. Table 8 shows that D-HT outperforms DeepLPF in terms of PSNR, SSIM, and LPIPS. Figure 11 further validate that our D-HT model can recover distinct contrast and natural color as well as clear details, thanks to the design of disentanglement with Transformer.

6. Conclusion

In this paper, we propose a novel way of image harmonization with Transformer, aiming to eliminate the inharmony by leveraging Transformer's modeling ability of long-range context dependencies. We not only build harmonization Transformer and disentangled harmonization Transformer frameworks, but also design comprehensive experiments to explore and analyze the Transformer on harmonization. We employ our methods on tasks beyond image harmonization, *i.e.*, image inpainting and image enhancement, further illustrating the superiority of our design. We hope that our work opens up new avenues for both image harmonization and vision Transformer.

References

[1] Jonathan T Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *IEEE TPAMI*, 37(8):1670–1687, 2014.

[2] H. G. Barrow and J. M. Tenenbaum. Recovering intrinsic scene characteristics from images. *Computer Vision Systems*, 1978.

[3] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[4] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229. Springer, 2020.

[6] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. *arXiv preprint arXiv:2012.00364*, 2020.

[7] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *ICML*, pages 1691–1703. PMLR, 2020.

[8] Daniel Cohen-Or, Olga Sorkine, Ran Gal, Tommer Leyvand, and Ying-Qing Xu. Color harmonization. In *SIGGRAPH*, pages 624–630, 2006.

[9] Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. DoveNet: Deep image harmonization via domain verification. In *CVPR*, pages 8394–8403, 2020.

[10] Xiaodong Cun and Chi-Man Pun. Improving the harmony of the composite image by spatial-separated attention module. *IEEE TIP*, 29:4759–4771, 2020.

[11] Zhigang Dai, Bolun Cai, Yugeng Lin, and Junying Chen. UP-DETR: Unsupervised pre-training for object detection with transformers. *arXiv preprint arXiv:2011.09094*, 2020.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186, 2019.

[13] Carl Doersch, Saurabh Singh, Abhinav Gupta, Josef Sivic, and Alexei A Efros. What makes paris look like paris? *ACM TOG*, 31(4):101, 2012.

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

[15] Xiaojie Guo, Yu Li, and Haibin Ling. LIME: Low-light image enhancement via illumination map estimation. *IEEE TIP*, 26(2):982–993, 2016.

[16] Zonghui Guo, Haiyong Zheng, Yufeng Jiang, Zhaorui Gu, and Bing Zheng. Intrinsic image harmonization. In *CVPR*, pages 16367–16376, 2021.

[17] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on visual transformer. *arXiv preprint arXiv:2012.12556*, 2020.

[18] Lin Huang, Jianchao Tan, Ji Liu, and Junsong Yuan. Hand-transformer: Non-autoregressive structured modeling for 3d hand pose estimation. In *ECCV*, pages 17–33. Springer, 2020.

[19] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, pages 1501–1510, 2017.

[20] Jiaya Jia, Jian Sun, Chi-Keung Tang, and Heung-Yeung Shum. Drag-and-drop pasting. *ACM TOG*, 25(3):631–637, 2006.

[21] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *arXiv preprint arXiv:2101.01169*, 2021.

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[23] Wei-Sheng Lai, Jia-Bin Huang, Zhe Hu, Narendra Ahuja, and Ming-Hsuan Yang. A comparative study for single image blind deblurring. In *CVPR*, pages 1701–1709, 2016.

[24] Jean-Francois Lalonde and Alexei A Efros. Using color compatibility for assessing image realism. In *ICCV*, pages 1–8, 2007.

[25] Edwin H Land. The retinex theory of color vision. *Scientific American*, 237(6):108–129, 1977.

[26] Edwin H Land and John J McCann. Lightness and retinex theory. *Journal of the Optical Society of America*, 61(1):1–11, 1971.

[27] Jingyuan Li, Ning Wang, Lefei Zhang, Bo Du, and Dacheng Tao. Recurrent feature reasoning for image inpainting. In *CVPR*, pages 7760–7768, 2020.

[28] Sean Moran, Pierre Marza, Steven McDonagh, Sarah Parisot, and Gregory Slabaugh. DeepLPF: Deep local parametric filters for image enhancement. In *CVPR*, pages 12826–12835, 2020.

[29] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. In *SIGGRAPH*, pages 313–318, 2003.

[30] François Pitié, Anil C Kokaram, and Rozenn Dahyot. N-dimensional probability density function transfer and its application to color transfer. In *ICCV*, pages 1434–1439, 2005.

[31] Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer Graphics and Applications*, 21(5):34–41, 2001.

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015.

[33] Kalyan Sunkavalli, Micah K Johnson, Wojciech Matusik, and Hanspeter Pfister. Multi-scale image harmonization. *ACM TOG*, 29:1–10, 2010.

[34] Michael W Tao, Micah K Johnson, and Sylvain Paris. Error-tolerant image compositing. *IJCV*, 103(2):178–189, 2013.

[35] Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. Deep image harmonization. In *CVPR*, pages 3789–3797, 2017.

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017.

[37] Ruixing Wang, Qing Zhang, Chi-Wing Fu, Xiaoyong Shen, Wei-Shi Zheng, and Jiaya Jia. Underexposed photo enhancement using deep illumination estimation. In *CVPR*, pages 6849–6857, 2019.

[38] Ning Xu, Brian Price, Scott Cohen, Jimei Yang, and Thomas S Huang. Deep interactive object selection. In *CVPR*, pages 373–381, 2016.

[39] Su Xue, Aseem Agarwala, Julie Dorsey, and Holly Rushmeier. Understanding and improving the realism of image composites. *ACM TOG*, 31(4):1–10, 2012.

[40] Yanhong Zeng, Jianlong Fu, and Hongyang Chao. Learning joint spatial-temporal transformations for video inpainting. In *ECCV*, pages 528–543. Springer, 2020.

[41] Ruo Zhang, Ping-Sing Tsai, James Edwin Cryer, and Mubarak Shah. Shape-from-shading: A survey. *IEEE TPAMI*, 21(8):690–706, 1999.

[42] Jun-Yan Zhu, Philipp Krahenbuhl, Eli Shechtman, and Alexei A Efros. Learning a discriminative model for the perception of realism in composite images. In *CVPR*, pages 3943–3951, 2015.

[43] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *ICLR*, 2021.