DCCF: Deep Comprehensible Color Filter Learning Framework for High-Resolution Image Harmonization

Ben Xue[1*], Shenghui Ran[2], Quan Chen[2**],
Rongfei Jia[2], Binqiang Zhao[2], and Xing Tang[2]

[1] Academy for Advanced Interdisciplinary Studies, Peking University, China
[2] Alibaba Group, China

Abstract. Image color harmonization algorithm aims to automatically match the color distribution of foreground and background images captured in different conditions. Previous deep learning based models neglect two issues that are critical for practical applications, namely high resolution (HR) image processing and model comprehensibility. In this paper, we propose a novel Deep Comprehensible Color Filter (DCCF) learning framework for high-resolution image harmonization. Specifically, DCCF first downsamples the original input image to its low-resolution (LR) counter-part, then learns four human comprehensible neural filters (i.e. hue, saturation, value and attentive rendering filters) in an end-to-end manner, finally applies these filters to the original input image to get the harmonized result. Benefiting from the comprehensible neural filters, we could provide a simple yet efficient handler for users to cooperate with deep model to get the desired results with very little effort when necessary. Extensive experiments demonstrate the effectiveness of DCCF learning framework and it outperforms state-of-the-art post-processing method on iHarmony4 dataset on images' full-resolutions by 7.63% and 1.69% relative improvements on MSE and PSNR, respectively. Our code is available at https://github.com/rockeyben/DCCF.

1 Introduction

Image composition, which aims at generating a realistic image with the given foreground and background, is one of the most widely used technology in photo editing. However, since the foreground and background may be captured in different conditions, simple cutting and pasting operations could not make them compatible in color space, as show in Fig. 1. Therefore, photo editors spend a lot of time in manual tuning the color distribution when they accomplish the real-world composition task.

In the past decades, a large amount of automatic color harmonization algorithms have been proposed. Traditional methods [3, 17, 22–25, 29, 30] tend to extract low-level handcrafted features to make the color statistics of the foreground

* Finish this work during an internship at Alibaba Group. Email: `xueben@pku.edu.cn`.
** Corresponding author. Email: `myctllmail@163.com`.

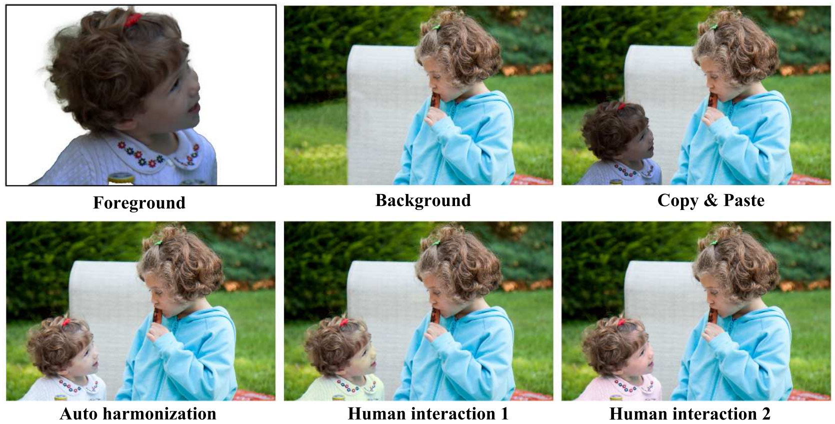


Fig. 1: Illustration of color harmonization

to match the background, which may have poor performance when the content of foreground and background are vastly different. Since Tsai et al. [31] propose a data-driven deep learning framework for color harmonization, the research community has made a large progress rapidly over a short period of time. Deep learning based methods have become the main stream. However, we argue that previous deep learning based color harmonization methods [4, 6, 7, 11, 20, 28, 31] have neglected two problems which are critical for practical applications.

First, high-resolution (HR) images are rarely taken into account in previous works when deep color harmonization models are designed and evaluated. Previous deep models in color harmonization follow the evaluation system proposed by Tsai et al. [31], which resizes the original images to 256×256 or 512×512 resolution and calculate objective metrics (i.e. MSE and PSNR) in this low-resolution to evaluate the performance of models, instead of the original image resolution. The principal reason is that these methods simply employ UNet-style [26] networks to directly predict pixel level RGB values, which are memory and computational costly, and even modern GPUs could not burden for HR images. However, color harmonization needs to be frequently applied to HR images in real-world applications whose resolution is 3000×3000 or even higher. Therefore, previous deep models which perform well on low-resolutions may have poor performance when be applied to real-world HR images.

Second, model comprehensibility and manual control mechanism are rarely considered in previous works. Imagine the scenario that the harmonization result of the network is flawed, and the photo editor wants to make some modifications based on the network's prediction to avoid tuning from scratch, such as hue adjustment in Fig. 1. Thus it is essential to provide human understandable cooperation mode with the deep models for a friendly color harmonization system. However, previous methods utilize variant networks following the common image-to-image translation framework [16] that directly predicts the harmonization result. It is nearly impossible to provide comprehensible tools for humans to interact with these deep models, because of the prediction processes are "black-box" and inscrutable for photo editors.

Inspired by the idea of learning desired image transformations that could reduce computing and memory burdens by a large margin for image enhancement [8], in this paper, we propose a novel Deep Comprehensible Color Filter (DCCF) learning framework for high-resolution image harmonization. Specifically, we first downsample the input to the low-resolution (such as 256×256) counter-part, then learn four comprehensible neural filters (i.e. hue, saturation, value and attentive rendering filters) in a novel end-to-end manner with the supervisions constructed from both RGB and HSV color spaces, finally apply these filters to the original input image to get the harmonized result. Compared with previous deep learning based color harmonization methods that may fail for high-resolution images, our neural filter learning framework is insensitive to image resolution and could perform well on dataset whose resolution range from 480p to 4K. Besides, benefiting from the mechanism that parameters in the filters (especially hue, saturation and value filters) are forced to learn decoupled meaningful chromatics functions, it makes it possible to provide comprehensible tools for humans to interact with these deep models in the traditional chromatics way they familiar with. It is worth noting that learning comprehensible neural filters is not easy. Our experiments show that learning weights directly from supervisions of hue, saturation and value channels could cause poor performance. To handle this, we construct three novel supervision maps that approximate the effects of HSV color space while making the deep model converge well.

We train and evaluate our approach in the open source iHarmony4 dataset [6] on the original image resolutions, which range from 480p (HCOCO) to 4K (HAdobe5k). Since previous deep learning based color harmonization models could perform poorly when they are directly applied to HR images, we compare to them with variant post-processing methods. Extensive experiments demonstrate that our approach can make the prediction process comprehensible and outperform these methods as well. We also provide a simple handler that humans could cooperate with the learned deep model to make some desired modifications based on the network's prediction capacity to avoid tuning from scratch.

In a nutshell, our contributions are three-folds.

- We propose an effective end-to-end deep neural filter learning framework that is insensitive to image resolution, which makes deep learning based color harmonization practical for real-world high-resolution images.
- To the best of our knowledge, we are the first to design four types of novel neural filter (i.e. hue, saturation, value and attentive rendering filters) learning functions and learning strategies that make the prediction process and result comprehensible for human in image harmonization task . Meanwhile, we provide a simple yet efficient handler for users to cooperate with deep model to get the desired results with very little effort when necessary.
- Our approach achieves state-of-the-art performance on the color harmonization benchmark for high-resolution images and outperforms state-of-the-art

post-processing method by 7.63% and 1.69% relative improvements on MSE and PSNR, respectively.

2 Related Work

Image harmonization In this subsection, we focus on the discussion of deep learning based methods. These methods regard color harmonization as a black box image-to-image translation task. [31] apply the well-known encoder-decoder U-net structure with skip-connection and train the network with multi-task learning, simultaneously predicting pixel value and semantic segmentation. [28] insert pretrained semantic segmentation branch into encoder backbone and introduce a learnable alpha-blending mask to borrow useful information from input image. They both use semantic features in networks. [4,6] tried to make composite image harmonious via domain transfer. [7,13] both used attention mechanism in networks. [1] propose a generative adversarial network (GAN) architecture for automatic image compositing, which considers geometric, color, and boundary consistency at the same time. [11] seek to solve image harmonization via separable harmonization of reflectance and illumination, where reflectance is harmonized through material-consistency penalty and illumination is harmonized by learning and transferring light from background to foreground. Note that recently some image harmonization works start to focus on high-resolution images. [18] use self-supervised learning strategy to train network with small local patches of high resolution images, but during inference it still follow the two stage post-processing strategy. [15, 27] learn global parameters to adjust image attributes such as lightness and saturation. [10] learns pixel-wise curves to perform low-light image enhancement.

Smart upsampling Processing high resolution image becomes difficult due to huge computational burden of deep-learning networks and limited GPU memory. A common approach to accelerate high resolution processing is to first downsample the image, apply time-consuming operator at low resolution and upsample back. To preserve edge gradients, guided filter upsampling [14] uses original high resolution input as guidance map. [9] fit transformation recipe from compressed input and output, then apply the recipe to high quality input. Bilateral guided upsampling [2] approximates the operator with grids of local affine transformations and apply them on high resolution input, thus control the operator complexity. [8] predict the local affine model with fully convolution networks, which is trained by end-to-end learning and obtain multi-scale semantic information. [32] propose a guided filter layer, using point-wise convolution to approximate median filter, thus can be plugged into networks and optimized jointly. [19] introduce extra networks to learn deformable offsets for each pixel, thus the interpolation neighbour is predicted online during upsampling. [5, 33] learn 3D lookup tables (LUT) to obtain high resolution results, but the learned transformation still lacks interpretable meanings.

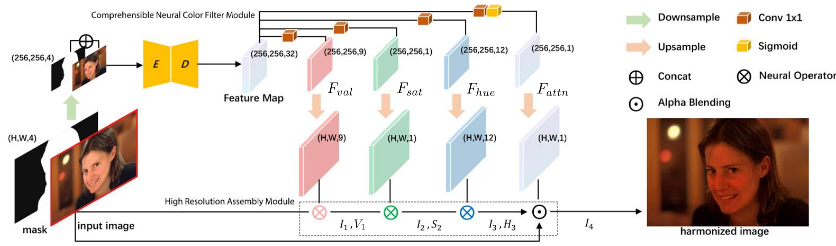


Fig. 2: An overview of our proposed color harmonization framework. It consists of two primary parts: *comprehensible neural color filter module* and *high resolution assembly module*. Given an input image and corresponding foreground mask, a low-resolution feature extraction backbone first downsamples them to a low-resolution version, such as 256×256, and employs an encoder-decoder network to extract foreground aware high-level semantic features. *Comprehensible neural color filter module* then learns value filter, saturation filter, hue filter and attentive rendering filter simultaneously based on the features extracted from the backbone. Each filter learns parameters of transformation function in per pixel manner. *High resolution assembly module* finally extracts and upsamples the specific channel of each DCCF's output to assemble the final result. In short, input image I is unharmonious, I_1 is V-harmonized, I_2 is V, S-harmonized, I_3 is V, S, H-harmonized, I_4 is the refinement of I_3 by an attention module.

3 Methodology

3.1 Framework Overview

The neural filter learning framework for high-resolution image color harmonization is illustrated in Fig. 2. It consists of two primary parts: *comprehensible neural color filter module* and *high resolution assembly module*.

Firstly, given an original input image ($H \times W \times 3$) and corresponding foreground mask ($H \times W \times 1$), low-resolution feature extraction backbone downsamples them to the low-resolution counterparts (256×256), then concatenates them as input (256×256×4) to extract foreground aware high-level semantic representations (256×256×32). The choice of backbone structure is flexible and iDIH-HRNet architecture [28] is used in this paper.

Subsequently, the *comprehensible neural color filter module* generates a series of deep comprehensible color filters (**DCCFs**) with the shape of ($256 \times 256 \times D$), where each pixel has D learnable parameters $\boldsymbol{q} = [q_1, q_2, ..., q_D]$ to construct a transformation function $f(I; \boldsymbol{q})$ which can be operated on input image I. The gathering of each pixel's functions f builds up a filter map F. The design of DCCFs and their cooperating mechanism will be detailed in Section 3.2.

Finally, the *high resolution assembly module* upsamples these filter maps to their full-resolution ($H \times W$) counterparts in order to be applied on the resolution of original input image. Meanwhile, since each DCCF only changes a specific aspect of image, an assembly strategy is thus required to ensure there is no conflicts between each filter's operating procedure. The details will be discussed in Section 3.3.

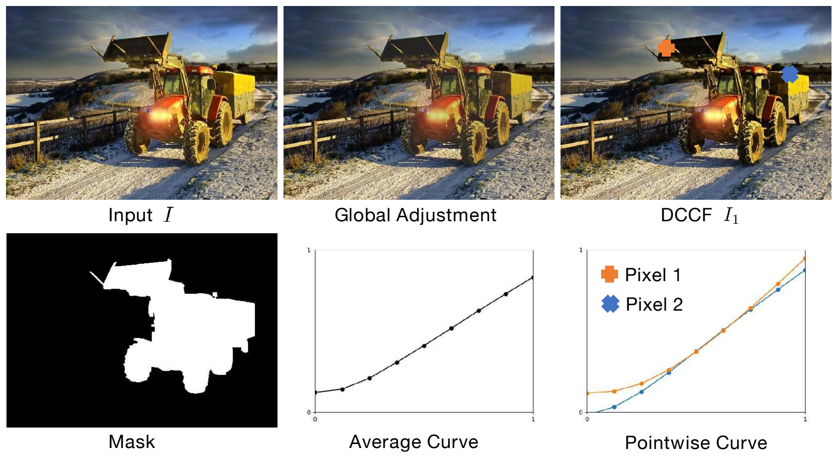


Fig. 3: Illustration of the pixel-level value adjustment function/curve. $I_1 = F_{val}(I)$ illustrated in Fig. 2 can be regarded as the result whose value is well tuned. Zoom for better view.

The entire network is trained in an end-to-end manner and benefited from the supervision of the full-resolution images. Moreover, we observe that traditional losses in RGB color space is not sufficient for achieving state-of-the-art quality. We therefore propose auxiliary losses in Section 3.4 for each DCCF's output to ensure that they are functioning as expected.

3.2 Comprehensible Neural Color Filter Module

The comprehensible neural color filter module plays a core rule in our proposed high-resolution image color harmonization framework. We take inspiration from the famous HSV color model which is widely used in photo editing community. Compared with RGB color space, HSV is much more intuitive and easier for humans to interact with computers for color tuning.

Our module consists of four neural filters, that is, *value filter*, *saturation filter*, *hue filter* and *attentive rendering filter* illustrated as F_{val}, F_{sat}, F_{hue} and F_{attn} respectively in Fig. 2. Each filter is generated by a 1×1 convolutional layer (expect for the attentive rendering filter has extra sigmoid layer for nomorlization) that builded on the low-resolution feature extraction backbone.

Value Filter The customized pointwise nonlinear value transformation function f_{val} is defined as:

$$f_{val}(x; \phi, V_{min}) = V_{min} + \sum_{i=1}^{m} \phi_i * \max(x - \frac{i-1}{m}, 0) \tag{1}$$

Where x indicates the V channel of input image in HSV color space, V_{min} and ϕ_i are learnable parameters and m is a hyper-parameter which we set as

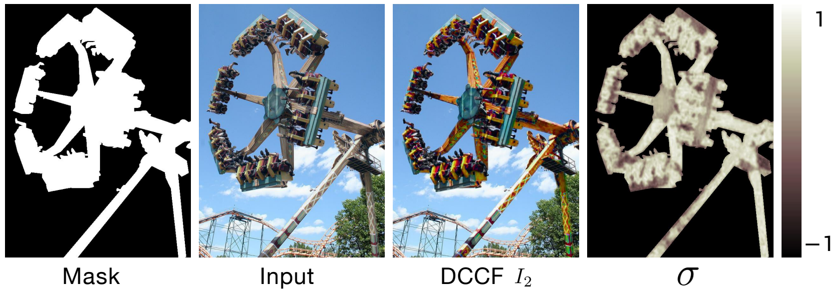


Fig. 4: Illustration of the pixel-level saturation adjustment. $I_2 = F_{sat}(I_1)$ illustrated in Fig. 2 is the intermediate result. The change of saturation is consistent with predicted σ distributions. Zoom for better view.

8 in this paper. It could be considered as an arbitrary nonlinear curve which is approximated by a stack of parameterized ReLUs. V_{min} controls lower-bound of value range, m and ϕ_i control nonlinearity of the curve. Parameters V_{min} and $\phi_i(i = 1, .., 8)$ are stored for each pixel in channel direction of value filter F_{val}.

We argue that different local regions should have different adjustment curves for better harmonization quality. As illustrated in Fig. 3, two marked points have large gap in original value distribution (the left is darker, the right is brighter), our DCCF F_{val} successfully allocates proper curves for these two regions, while the global adjustment degrades the overall aesthetic.

Saturation Filter We use a single parameter $\sigma \in [-1, 1]$ to control saturation for each pixel. The customized non-linear saturation transformation function f_{sat} for each pixel is defined as:

$$f_{sat}(x; \sigma) = x + (x - C_{med}) * clip(\sigma) \tag{2}$$

Where x indicates the R, G or B values in each pixel, $C_{max} = max(R, G, B)$, $C_{min} = min(R, G, B)$, $C_{med} = (C_{min} + C_{max})/2$, σ is our learned parameter and $clip(\sigma)$ is a monotonous function to avoid saturation overflow.

If $\sigma \to 1$, the values below median will be suppressed, while the value above median will be enhanced, as a result the saturation is increased and vice versa when $\sigma \to -1$. We visualize the effectiveness of σ in Fig. 4. DCCF allocated positive σ for most of the pixels in this de-saturated input image and obtained an enhanced result.

Hue Filter We define an affine color transformation function f_{col} for each pixel in RGB color space as:

$$
\begin{aligned}
f_{col}(x;\boldsymbol{\Delta}) &= \boldsymbol{R}x + t \\
&= \begin{bmatrix} \delta_{11} & \delta_{12} & \delta_{13} \\ \delta_{21} & \delta_{22} & \delta_{23} \\ \delta_{31} & \delta_{32} & \delta_{33} \end{bmatrix} \begin{bmatrix} x_R \\ x_G \\ x_B \end{bmatrix} + \begin{bmatrix} \delta_{14} \\ \delta_{24} \\ \delta_{34} \end{bmatrix}
\end{aligned} \tag{3}
$$

Where x indicates the RGB values for one pixel in image, and $\boldsymbol{\Delta}$ is a learnable 3x4 affine transformation matrix that contains a rotation matrix $\boldsymbol{R}$ and a translation vector t.

We suppose that one could find a suitable rotation matrix $\boldsymbol{R}$ in RGB color space that is equivalent to a corresponding radian moving $\boldsymbol{r}$ on the hue ring in HSV color space [12], which is futher discussed in supplementary. Based on this assumption, it is equivalent to learn an affine color transformation function $f_{col}(x;\boldsymbol{\Delta})$ in RGB color space, which contains a rotation function $\boldsymbol{R}$ that could be parameters for the corresponding hue rotation function $f_{hue}(h;\boldsymbol{R})$ in HSV color space. We suggest readers refer to [12] for technical details. Note that [12] needs extra linearization between sRGB and RGB space, which is mainly a gamma correction thus compatible with our learnable curve function f_{val}.

Attentive Rendering Filter We employ simple yet effective attentive rendering filter F_{attn} which is similar to the attention mask in [28] to further improve the harmonization result after hue filter.

For inference, we adopt the previous filters' harmonization result I_3 and input I to perform alpha blending as illustrated in Fig. 2

$$
I_4 = I * \alpha + W_{ref} * I_3 * (1 - \alpha) \tag{4}
$$

Where α is the per-pixel parameter on F_{attn} ranging in $[0, 1]$ to smartly borrow information from input image, W_{ref} is an extra affine matrix to refine the appearance of I_3.

3.3 High Resolution Assembly Module

The biggest reduction of computation comes from the design that each DCCF is generated at low-resolution branch. We then perform upsampling on DCCF's filter map to match the resolution of original input image. The effectiveness of this action is guaranteed by the common assumption that neighbourhood regions require similar tuning filters.

Afterwards, we propose a split-and-concat strategy to assemble the applying result of each filter. Specifically, as shown in Fig. 2, we utilize value filter F_{val}, saturation filter F_{sat} and hue filter H_{hue} to extract harmonized value channel V_1, saturation channel S_2 and hue channel H_3 respectively, then assemble V_1, S_2 and H_3 as harmonized image I_3, finally use attentive rendering filter to get the final harmonized image I_4. We illustrate the implementation details of saturation assembling as example in Fig. 5.

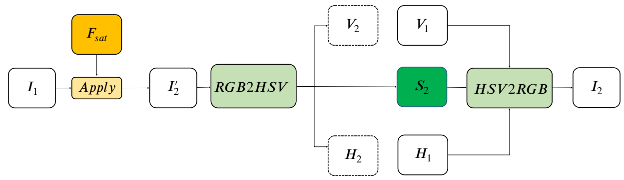


Fig. 5: Illustration of assembly module details. We take the procedure of saturation filter F_{sat} as example. The engaged channel (i.e. S_2) are colored for visualization.

3.4 Training Loss

In the following description, we will use the superscript l for low-resolution and h for high-resolution.

High Resolution Supervision Since the area of foreground region varies a lot among training examples, we adopt foreground-normalized MSE loss [28] between ground-truth I_{gt} and intermediate result I_3, final predicted result I_4. This loss uses the area of foreground mask as a normalization factor to stablize the gradient on foreground object. Differently, our loss can be calculated on both low-resolution and high-resolution streams, namely $\mathcal{L}^l_{rgb}$ and $\mathcal{L}^h_{rgb}$.

Auxiliary HSV Loss A straight forward solution to supervise $F_{val}, F_{sat}, F_{hue}$ is using the standard HSV decomposition equations to get HSV channels. However, we observe that this strategy could contain high frequency contents in the output channel as visualized in Fig. 6a $\sim$ Fig. 6f, which may degrade the convergence of network according to our experiments in Fig. 6g.

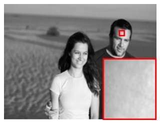
Therefore we heuristicly designed an approximated version of HSV loss to stabilize network training. It is mainly based on a combination of several differentiable basic image processing filters (e.g. whitening, blurring, blending) to obtain smooth approximations of these three attributes H, S, V, which benifit training procedure. The implementation details are shown in supplementary.

Auxiliary HSV losses $\mathcal{L}^l_{val}$, $\mathcal{L}^l_{sat}$, $\mathcal{L}^l_{hue}$ are calculated with MSE in low resolution stream only due to memory consideration. We also apply total variation regularization on predicted filters to increase smoothness. The overall training loss is defined as follows, where $\lambda_i (i = 1, ..., 5)$ is hyper-parameters:

$$\mathcal{L} = \lambda_1 \mathcal{L}^l_{rgb} + \lambda_2 \mathcal{L}^h_{rgb} + \lambda_3 \mathcal{L}^l_{val} + \lambda_4 \mathcal{L}^l_{sat} + \lambda_5 \mathcal{L}^l_{hue} \tag{5}$$

4 Experiments

In this section, we first describle experimental setups and implementation details, then compare our approach with the state-of-the-arts quantitatively and qualitatively. Finally, we carry out some ablation studies and provide a simple comprehensible interface to interact with our model. We also present more results and potential limitations in supplementary materials.

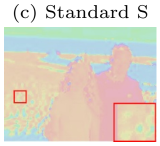


(a) Standard V

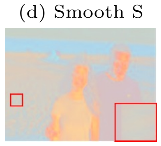
(b) Smooth V

(c) Standard S

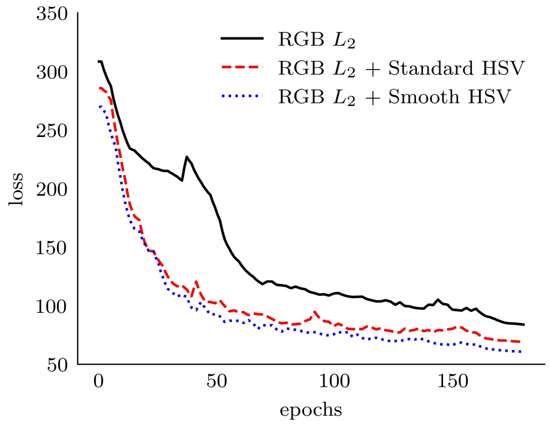
(d) Smooth S



(e) Standard H



(f) Smooth H



(g) Ablations on loss function. Testing errors on iHarmony4 with different loss contributions.

Fig. 6: Visualzation of standard HSV and our ad-hoc smoothed version. The smoothed version of V, S, H keeps global chromological properties, meanwhile makes the network converge better, which is demonstrated in sub-figure (g).

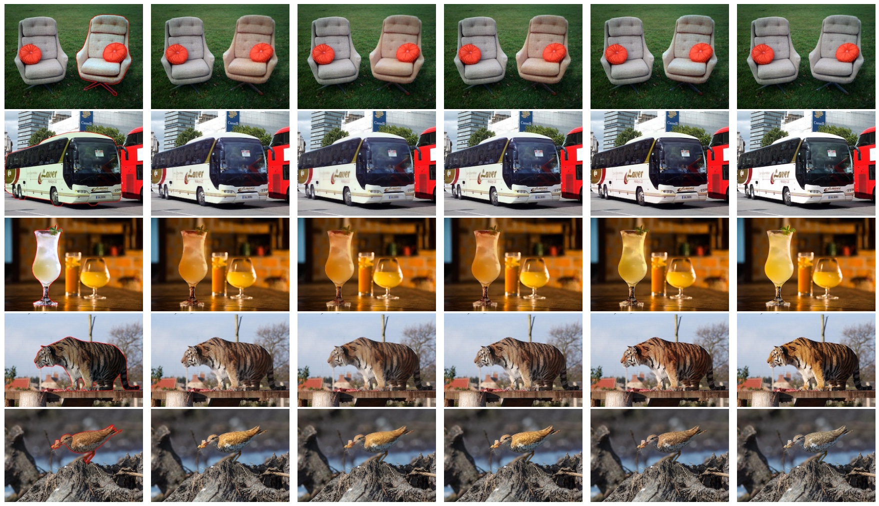
Table 1: **Quantative performance comparison** on the iHarmony4 test sets. We are the first to evaluate on original resolution in this dataset. The best results are in bold. '-' means not able to obtain results due to memory limitation. Our method is trained in an end-to-end manner and outperforms post-upsampling baselines by comparison. More quantative results with different backbones are shown in supplementary.

	Entire Dataset		HCOCO		HAdobe5k		HFlickr		Hday2night	
Method	MSE ↓	PSNR ↑	MSE ↓	PSNR ↑	MSE ↓	PSNR ↑	MSE ↓	PSNR ↑	MSE ↓	PSNR ↑
Input image	177.99	31.22	73.03	33.53	354.46	27.63	270.99	28.20	113.07	33.91
iDIH-HRNet [28]	-	-	19.96	38.25	-	-	93.50	32.42	71.01	35.77
iDIH-HRNet [28]+BU	43.56	34.98	34.40	35.45	37.82	35.47	104.69	30.91	50.87	37.41
iDIH-HRNet [28]+GF [14]	35.47	36.00	25.93	36.70	34.51	36.03	85.05	32.01	**49.90**	**37.67**
iDIH-HRNet [28]+BGU [2]	26.85	37.24	18.53	37.90	26.71	37.50	66.26	33.19	51.96	37.23
DCCF	**24.65**	**37.87**	**17.07**	**38.66**	**23.34**	**37.75**	**64.77**	**33.60**	55.76	37.40

4.1 Experimental Setups

We use iHarmony4 [6] as our experiment dataset which contains 73146 images. It consists of 4 subsets: HCOCO, HFlickr, HAdobe5k, HDay2night. The image resolution varies from 640×480 to 6048×4032, which is difficult for learning based color harmonization algorithms to process on the original images' full resolution. We suggest readers refer to [6] for dataset details.

Since the lack of high-resolution process ability, previous methods [4, 6, 7, 11, 20, 28, 31] resize all images in the dataset to 256×256 to process and evaluate their performance via Mean Square Error (MSE) and Peak Signal To Noise Ratio (PSNR) in this extremely low-resolution. However, we argue that evaluate algorithms on the image's original full-resolution is much more scientific for practical applications. In this paper, we adopt MSE and PSNR as our objective metrics on the image's original full-resolution instead of 256×256.



(a) Input (b) BU (c) GF (d) BGU (e) DCCF (f) GT

Fig. 7: **Visualization of high-resolution results**. Foregrounds are marked in red contour. Bilinear upsampling, guided filter upsampling and bilateral guided upsampling are represented as BU, GF [14] and BGU [2] respectively. GT represents ground truths. Our method DCCF has not only better global appearance but also refined high resolution details. Zoom for better view. More visual results please refer to supplementary materials.

4.2 Implementation Details

Our DCCF learning framework is differentiable and could be stacked on the head of any deep feature extraction networks. In this paper, we adopt the recent state-of-the-art harmonization network iDIH-HRNet [28] as our backbone to carry out experiments. For feature extraction backbone, we downsample inputs (images and corresponding foreground mask) to 256×256 following the previous deep harmonization models' common setting. For detailed training procedure and hyper-parameter setting, please refer to our official Pytorch [21] code[3].

4.3 Comparison with Baselines

In order to evalute the effectiveness of our proposed DCCF learning framework, we construct two kinds of baselines. (1) Applying recent state-of-the-art methods directly on the original input images to get the full-resolution harmonized results. (2) Applying recent state-of-the-art methods on the low-resolution inputs (256×256) to predict low-resolution harmonized images and adopting variates of state-of-the-art post-processing methods to get the final full-resolution harmonized results. In this paper, we choose iDIH-HRNet [28] as the deep model

[3] https://github.com/rockeyben/DCCF

Table 2: **Qualitative results**. We evaluate visual perceptual quality by DNN-based image quality accessment LPIPS [34] and a user study.

Method	iDIH-HRNet [28] +BU	iDIH-HRNet [28]+GF [14]	iDIH-HRNet [28]+BGU [2]	DCCF
LPIPS [34] ↓	0.0459	0.0291	0.0201	**0.0186**
User Score ↑	2.0541	2.6583	3.3041	**3.5583**

provided by Sofiiuk et al. [28] and Bilinear Upsampling (BU), Guided Filter Upsampling [14] (GF), Bilateral Guided Upsampling [2] (BGU) as post-processing methods. For fair comparison, we adopt the same low-resolution (i.e. 256×256) feature extractor as [28] for our DCCF learning framework. The performance comparison is shown in Table 1. Some harmonization results are shown in Fig. 7. For the comparison of efficiency metrics like inference time and memory usage, please refer to supplementary for details.

The method of applying [28] directly on the full-resolution (first row in Table 1) performs pooly. The principal reason is that [28] is designed and trained on the resolution of 256×256, directly applying this model in testing phase to the original image full-resolution would lead to serious feature misalignment. Moreover this strategy failed on HAdobe5k subset (max resolution: 6048×4032) due to memory limitation.

The method of applying post-processing after the low-resolutional prediction results from [28] with low-resolution inputs solves the memory problem. However, BU would lead to blurring effect, especially for high-resolution subset HAdobe5k, see Fig 7. Therefore, we adopt more advanced post-processing algorithms GF [14] and BGU [2] that take original full-resolution image as detail guidance to mitigate the blurring effect from upsampling operation. Table 1 shows that these upsampling methods outperform bilinear upsampling methods by a large margin and the best one BGU [2] achieves 26.85 on MSE and 37.24 on PSNR. However, the best performance of post-processing methods is behind our approach DCCF. Our approach achieves 24.65 on MSE and 37.87 on PSNR, 7.63% and 1.69% relative improvements on MSE and PSNR respectively compared with [28]+ BGU [2].

4.4 Qualitative Results

We conduct two evaluations to compare the subjective visual quality of DCCF with other methods, which is presented in Table 2. First, we adopt LPIPS [34] to evalute visual perceptual similarity of harmonized image and ground truth reference. It computes the feature distance between two images and the lower score indicates better result. Second, we randomly select 20 images then present DCCF result with baseline results on the screen after shuffling, and ask 12 users to judge images' global appearance and detail texture then give scores from 1 to 5, the higher the better. Our DCCF achieves the best result in both metrics which is consistent the quantative performance.

Table 3: **Ablation studies**. (a) As for filter design, DBL [8] is a "black-box" per-pixel linear filter that directly applied to RGB images. DCCFs with attention achieves the best result. (b) As for losses, supervision we constructed from HSV (i.e. smooth $\mathcal{L}_{hsv}$) is essential and improves standard HSV by 3.21 (11.65%) on MSE.

(a) Method	MSE ↓	PSNR ↑
DBL [8]	27.92	37.48
DCCFs w.o. attention	26.36	37.80
DCCFs with attention	**24.65**	**37.87**

(b) Method	MSE ↓	PSNR ↑
$\mathcal{L}_{rgb}$	35.17	36.81
$\mathcal{L}_{rgb}$ + standard $\mathcal{L}_{hsv}$	27.86	37.39
$\mathcal{L}_{rgb}$ + smooth $\mathcal{L}_{hsv}$	**24.65**	**37.87**

4.5 Ablation Studies

Filter Design An evaluation of filter design is shown in Table 3a. DBL [8] is an end-to-end "black-box" bilateral learning method that proposed in image enhancement. We adapt it to our DCCF learning framework to process high-resolution image harmonization. DCCFs w.o. attention is our DCCF learning method that exclude attentive rendering filter. Even DCCFs w.o. attention improves the performance of DBL filter [8] by 1.56 (5.58%) on MSE. It demonstrate that the performance of our model is not just from end-to-end training, our divide, conquer and assemble strategy that learns explicit meaningful parameters also benefit a lot for color harmonization task. DCCFs with attn further improve the DBL filter [8] by 3.27 (11.71%) on MSE.

Loss Functions The impact of loss functions for our DCCF learning framework is shown in Table 3b. Note that standard H channel is an angle value while our approximated H is a scalar value, so we train standard $\mathcal{L}_h$ with cosine distance while training approximated smooth $\mathcal{L}_h$ with euclidean distance. Numerical results show that supervisions from HSV color space is essential for our DCCF learning framework, which is manifested in simply adding loss from standard HSV channels will remarkably decrease MSE from 35.17 to 27.86. The principal reason may be the parameters of our DCCFs (expect for the last attentive rendering filter) are designed from the inspiration of practical tuning criteria in HSV color space used by color artists and has explicit chromatics meaning. Therefore model converges better when supervisory signals from HSV color space are added, which is demonstrate in Fig. 6g. It is worth noting that adding smooth approximated HSV loss described in subsection 3.4 instead of standard HSV loss will further decrease MSE to 24.65 which demonstrates the effectiveness of proposed smoothing HSV loss.

4.6 Comprehensible Interaction with Deep Model

Benefiting from the comprehensible neural filters, we could provide a simple yet efficient handler for users to cooperate with deep model to get the desired results with very little effort when necessary. We provide two adjustable parameters in the three dimensions of hue, saturation and value respectively for users to express their color adjustment intentions. For space limitation, we only explain

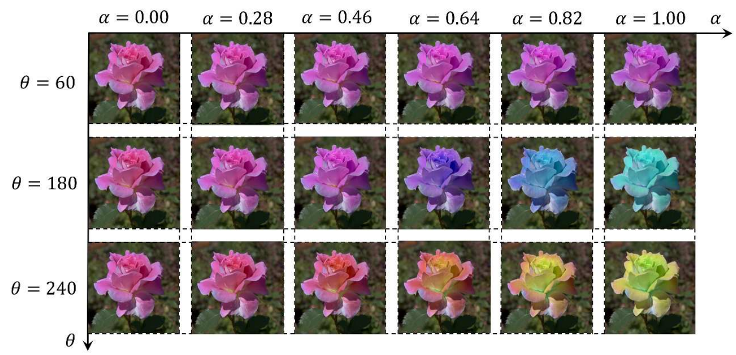


Fig. 8: Illustration of comprehensible interaction with deep harmonization model on parameter space of hue adjustment. Abscissa represents parameter α and ordinate represents parameter θ. Sampling values in (α, θ) and their results are listed. Zoom for better view.

hue adjustment for example. The other two dimensions are similar and will be detailed in supplementary.

For hue, we define parameter $\theta \in [0, 360]$ and $\alpha \in [0, 1]$ to represent the angle for Hue circle and the amount of user color intentions respectively. We calculate the desired rotation matrix R mentioned in Eq. (3) as:

$$\begin{bmatrix} \frac{1}{3} - \frac{2\cos\theta}{3} & \frac{1-\cos\theta}{3} - \frac{\sin\theta}{\sqrt{3}} & \frac{1-\cos\theta}{3} + \frac{\sin\theta}{\sqrt{3}} \\ \frac{1-\cos\theta}{3} + \frac{\sin\theta}{\sqrt{3}} & \frac{1}{3} - \frac{2\cos\theta}{3} & \frac{1-\cos\theta}{3} - \frac{\sin\theta}{\sqrt{3}} \\ \frac{1-\cos\theta}{3} + \frac{\sin\theta}{\sqrt{3}} & \frac{1-\cos\theta}{3} + \frac{\sin\theta}{\sqrt{3}} & \frac{1}{3} - \frac{2\cos\theta}{3} \end{bmatrix} \tag{6}$$

Then we could get the final rotation matrix R: $F'_{hue} = \alpha * R + (1 - \alpha) * F_{hue}$, which can be applied on image that takes global user intentions and local complex self-adaptions from deep model in mind.

In one word, users can express their color intentions by parameter $\theta \in [0, 360]$ and decide the amount of injected color by controlling $\alpha \in [0, 1]$, which is illustrated in Fig. 8. It is worth noting that when users interact with deep model in one dimension (such as hue above), they need not worry about the side-effect changes of other two dimensions from the network's prediction.

5 Conclusion

In this paper, we propose comprehensible image processing filters to deal with image harmonization problem. By gradually modifying image's attributes: value, saturation and hue, we can obtain results not only high-quality but also understandable. This also facilitate human to cooperate with deep models to perform image harmonization. We also leverage these filters to tackle high resolution images in a simple yet effective way. We hope that DCCF can set up a brand new direction for image harmonization.

References

1. Chen, B.C., Kae, A.: Toward realistic image compositing with adversarial learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8415–8424 (2019)
2. Chen, J., Adams, A., Wadhwa, N., Hasinoff, S.W.: Bilateral guided upsampling. ACM Transactions on Graphics (TOG) **35**(6), 1–8 (2016)
3. Cohen-Or, D., Sorkine, O., Gal, R., Leyvand, T., Xu, Y.Q.: Color harmonization. In: ACM SIGGRAPH, pp. 624–630 (2006)
4. Cong, W., Niu, L., Zhang, J., Liang, J., Zhang, L.: Bargainnet: Background-guided domain translation for image harmonization. In: IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6 (2021)
5. Cong, W., Tao, X., Niu, L., Liang, J., Gao, X., Sun, Q., Zhang, L.: High-resolution image harmonization via collaborative dual transformations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18470–18479 (2022)
6. Cong, W., Zhang, J., Niu, L., Liu, L., Ling, Z., Li, W., Zhang, L.: Dovenet: Deep image harmonization via domain verification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8394–8403 (2020)
7. Cun, X., Pun, C.M.: Improving the harmony of the composite image by spatial-separated attention module. IEEE Transactions on Image Processing (TIP) **29**, 4759–4771 (2020)
8. Gharbi, M., Chen, J., Barron, J.T., Hasinoff, S.W., Durand, F.: Deep bilateral learning for real-time image enhancement. ACM Transactions on Graphics (TOG) **36**(4), 1–12 (2017)
9. Gharbi, M., Shih, Y., Chaurasia, G., Ragan-Kelley, J., Paris, S., Durand, F.: Transform recipes for efficient cloud photo enhancement. ACM Transactions on Graphics (TOG) **34**(6), 1–12 (2015)
10. Guo, C., Li, C., Guo, J., Loy, C.C., Hou, J., Kwong, S., Runmin, C.: Zero-reference deep curve estimation for low-light image enhancement. pp. 1780–1789 (2020)
11. Guo, Z., Zheng, H., Jiang, Y., Gu, Z., Zheng, B.: Intrinsic image harmonization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16367–16376 (2021)
12. Haeberli, P.: Matrix operations for image processing. Grafica Obscura website (1993), http://graficaobscura.com/matrix/index.html
13. Hao, G., Iizuka, S., Fukui, K.: Image harmonization with attention-based deep feature modulation. In: British Machine Vision Conference (BMVC) (2020)
14. He, K., Sun, J., Tang, X.: Guided image filtering. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **35**(6), 1397–1409 (2013)
15. Hu, Y., He, H., Xu, C., Wang, B., Lin, S.: Exposure: A white-box photo post-processing framework. ACM Transactions on Graphics (TOG) **37**(2), 26 (2018)
16. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1125–1134 (2017)
17. Jia, J., Sun, J., Tang, C.K., Shum, H.Y.: Drag-and-drop pasting. ACM Transactions on Graphics (TOG) **25**(3), 631–637 (2006)
18. Jiang, Y., Zhang, H., Zhang, J., Wang, Y., Lin, Z., Sunkavalli, K., Chen, S., Amirghodsi, S., Kong, S., Wang, Z.: Ssh: A self-supervised framework for image harmonization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4832–4841 (2021)

19. Kim, B., Ponce, J., Ham, B.: Deformable kernel networks for guided depth map upsampling. CoRR **abs/1903.11286** (2019)
20. Ling, J., Xue, H., Song, L., Xie, R., Gu, X.: Region-aware adaptive instance normalization for image harmonization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9361–9370 (June 2021)
21. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems (NeurIPS) **32**, 8026–8037 (2019)
22. Pérez, P., Gangnet, M., Blake, A.: Poisson image editing. In: ACM SIGGRAPH, pp. 313–318 (2003)
23. Pitié, F., Kokaram, A.: The linear monge-kantorovitch linear colour mapping for example-based colour transfer. In: Proceedings of the European Conference on Visual Media Production (CVMP). pp. 1–9 (2007)
24. Pitie, F., Kokaram, A.C., Dahyot, R.: N-dimensional probability density function transfer and its application to color transfer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). vol. 2, pp. 1434–1439 (2005)
25. Reinhard, E., Adhikhmin, M., Gooch, B., Shirley, P.: Color transfer between images. IEEE Computer Graphics and Applications (CG&A) **21**(5), 34–41 (2001)
26. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 234–241 (2015)
27. Shi, J., Xu, N., Xu, Y., Bui, T., Dernoncourt, F., Xu, C.: Learning by planning: Language-guided global image editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13590–13599 (2021)
28. Sofiiuk, K., Popenova, P., Konushin, A.: Foreground-aware semantic representations for image harmonization. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1620–1629 (2021)
29. Sunkavalli, K., Johnson, M.K., Matusik, W., Pfister, H.: Multi-scale image harmonization. ACM Transactions on Graphics (TOG) **29**(4), 1–10 (2010)
30. Tao, M.W., Johnson, M.K., Paris, S.: Error-tolerant image compositing. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 31–44 (2010)
31. Tsai, Y.H., Shen, X., Lin, Z., Sunkavalli, K., Lu, X., Yang, M.H.: Deep image harmonization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3789–3797 (2017)
32. Wu, H., Zheng, S., Zhang, J., Huang, K.: Fast end-to-end trainable guided filter. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1838–1847 (2018)
33. Zeng, H., Cai, J., Li, L., Cao, Z., Zhang, L.: Learning image-adaptive 3d lookup tables for high performance photo enhancement in real-time. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2020)
34. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)