Who looks like me: Semantic Routed Image Harmonization

Jinsheng Sun[1] and **Chao Yao**[2*] and **Xiaokun Wang**[1] and **Yu Guo**[1] and **Yalan Zhang**[1] and **Xiaojuan Ban**[1,3,4*]

[1]Beijing Advanced Innovation Center for Materials Genome Engineering, School of Intelligence Science and Technology, University of Science and Technology Beijing, Beijing 100083, China.

[2]School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China.

[3]Key Laboratory of Intelligent Bionic Unmanned Systems, Ministry of Education, University of Science and Technology Beijing, Beijing 100083, China.

[4]Institute of Materials Intelligent Technology, Liaoning Academy of Materials, Shenyang 110004, China

b20180318@xs.ustb.edu.com, {yaochao, wangxiaokun, guoyu, zhangyl, banxj}@ustb.edu.com

Abstract

Image harmonization, aiming to seamlessly blend extraneous foreground objects with background images, is a promising and challenging task. Ensuring a synthetic image appears realistic requires maintaining consistency in visual characteristics, such as texture and style, across global and semantic regions. In this paper, We approach image harmonization as a semantic routed style transfer problem, and propose an image harmonization model by routing semantic similarity explicitly to enhance the consistency of appearance characteristics. To refine calculate the similarity between the composed foreground and background instance, we propose an Instance Similarity Evaluation Module(ISEM). To harness analogous semantic information effectively, we further introduce Style Transfer Block(STB) to establish fine-grained foreground-background semantic correlation. Our method has achieved excellent experimental results on existing datasets and our model outperforms the state-of-the-art by a margin of 0.45 dB on iHarmony4 dataset.

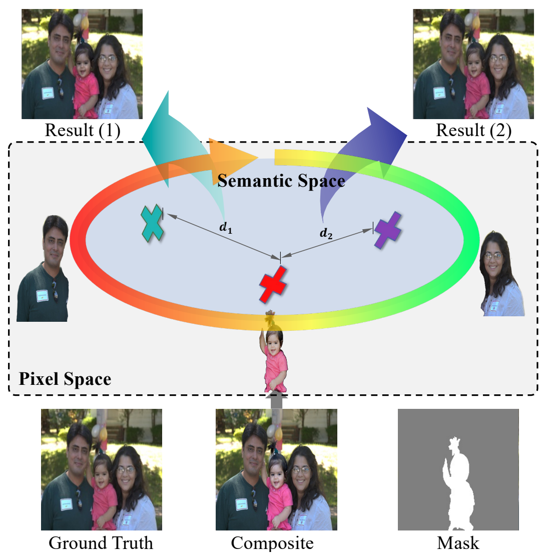

Figure 1: Illustration of image harmonization guided by semantic similarity. The appearance characteristics and semantic similarity of foreground and background objects are more related. The little girl could be related to multiple instances in the background. A stronger influence from the left-side instance leads to a more subdued color profile, whereas a stronger influence from the right-side instance results in a more vibrant color profile.

1 Introduction

Image editing technology is extensively utilized across various aspects of our daily lives, encompassing areas such as commercial promotion, social sharing, digital entertainment, and even the emerging realm of the Metaverse [Kaur *et al.*, 2023; Ren and Liu, 2022]. Notably, AIGC [Ho *et al.*, 2020; Kim *et al.*, 2022] technology empowers the direct generation of a diverse array of images, although many synthetic images require subsequent editing to enhance realism. However, individuals lacking professional photo-editing expertise may find that composited images face challenges in terms of evaluation credibility, stemming from issues such as inharmonious color, texture, or illumination. Consequently, the process of image harmonization becomes imperative for elevating the overall quality of composite images.

Numerous methods have been developed with the objective of harmonizing composite images, addressing the discordance between foreground and background [Cong *et al.*, 2020; Liang and Pun, 2022; Ren and Liu, 2022; Zhu *et al.*, 2022; Chen *et al.*, 2022; Niu *et al.*, 2023]. *Zhu et al.* [Zhu *et al.*, 2022] proposed a technique to align the representation of each foreground location with corresponding background elements. In a different approach, *Tsai et al.* [Tsai *et al.*, 2017] introduced an end-to-end learning method for image harmonization, primarily focusing on constraining semantic infor-

*Correspongding author

mation learning in the encoder. *Cun et al.* [Cun and Pun, 2020] integrated a spatial-separated attention module to compel the network to learn foreground and background features separately, but this approach falls short in ensuring style consistency between the two components. However, these existing methods predominantly emphasize visual style consistency between foreground and background regions, lacking realism derived from instance similarity.

Based on the human perception process for image harmonization, the appearance characteristics and semantic similarity of foreground and background objects are highly relevant. As illustrated in Figure 1, the little girl could be related to multiple instances in the background, including the man on the left and the woman on the right, with varying degrees of semantic similarity, When the appearance characteristics are influenced by semantic similarity, the resulting harmonization exhibits distinct characteristics. A stronger influence from the left-side instance leads to a more subdued color profile, whereas a stronger influence from the right-side instance results in a more vibrant color profile.

To alleviate the ambiguity derived from different semantic information, we propose an image harmonization model by measuring semantic similarity explicitly to enhance the consistency of appearance characteristics. As the saying goes, "who looks like me". We approach image harmonization as a semantic routed style transfer problem, focusing on refining the appearance of foreground objects using the style guidance of the most similar instance. Specifically, an Instance Similarity Evaluation Module (ISEM) is designed to compute the similarity matrices of components between the composed foreground object and the background instances. To harness analogous semantic information more effectively, we further introduce the Style Transfer Block (STB). On one hand, this module is specifically crafted to query the most akin background instance. On the other hand, corresponding style characteristics are seamlessly transferred onto the composed foreground object, enhancing the overall harmonization process. Extensive experiments including human perception experiments demonstrate the superior performance of our proposed method in improving image harmonization.

In summary, our contributions are given as follows:

- We design an image harmonization framework by evaluating the instance-similarity
- We propose an instance similarity evaluation module (ISEM), designed to assess the similarity of components within both the semantic and stylistic domains of instances in the foreground and background.
- We introduce a style transfer block(STB) that captures the global style information of the input image and transfers it to the latent space of the style encoder.

2 Related Work

Most early studies on image harmonization aimed to design and match low-level color statistical information of foreground and background, such as color histograms [Xue *et al.*, 2012], gradient information [Perez *et al.*, 2023] and image pyramids [Sunkavalli *et al.*, 2010]. The utilization scenarios of these methods are significantly constrained due to limitations in representing high-level features. Paired images and harmonized training data [Tsai *et al.*, 2017; Cong *et al.*, 2020] have been constructed by adjusting the color and illumination of the foreground objects in real images. Based on these datasets, large numbers of image harmonization models based on supervised deep learning models have been proposed and achieved more reliable results using these datasets. DIH [Tsai *et al.*, 2017] and *Sofiiuk et al.* [Sofiiuk *et al.*, 2021] use semantic information to capture image context, which aids in harmonizing the composite foreground. RainNet[Ling *et al.*, 2021] treats the mean and variance of the deep features as appearance information and adjusts the mean and variance of the foreground to match those of the background. In addition, several endeavors have attempted to apply models that have achieved outstanding performance in other domains, such as Transformer [Guo *et al.*, 2021a] and diffusion models [Lu *et al.*, 2023; Li *et al.*, 2023], to address the task of image harmonization.

Furthermore, in the pursuit of context consistency, recent notable works have approached image harmonization as a style transfer problem [Song *et al.*, 2023]. These endeavors aim to precisely transfer the global features of the background onto the composed foreground object. *Hao et al.* [Hao *et al.*, 2020] align the standard deviation of the foreground features with that of the background features, capturing global dependencies in the entire image. BargainNet [Cong *et al.*, 2021] uses a domain code extractor to capture background domain information, guiding the foreground's harmonization. Recently, Hang et al. [Hang *et al.*, 2022] has achieved state-of-the-art results by incorporating background and foreground style consistency constraints and dynamically sampling negative examples in a contrastive learning paradigm. These methods leverage network models to learn the relationship between foreground and background feature representations implicitly.

In this paper, the background elements that exert a more pronounced influence on the appearance characteristics of foreground objects are concerned. We explicitly extract the semantic relationship between the background and foreground elements, and employ this information to guide and inform the image harmonization process.

3 Methods

3.1 Overall Pipeline

The objective of our paper is to maintain consistent appearance characteristics between the foreground and background of synthetic images. Consequently, forming a substantial association between the composite foreground instance and other background instances is vital for crafting harmonious appearance uniformity. As depicted in Figure 2, we initially deploy a pre-trained SAM model to divide the synthetic image into a semantic space, with the mask of the foreground functioning as the model's prompt. Subsequently, semantic mapping takes place to transform the SAM model's output into the semantic and location data of the background instances. We introduce the Instance Similarity Evaluation Module (ISEM), designed to compute a similarity matrix between the composite foreground instance and the various

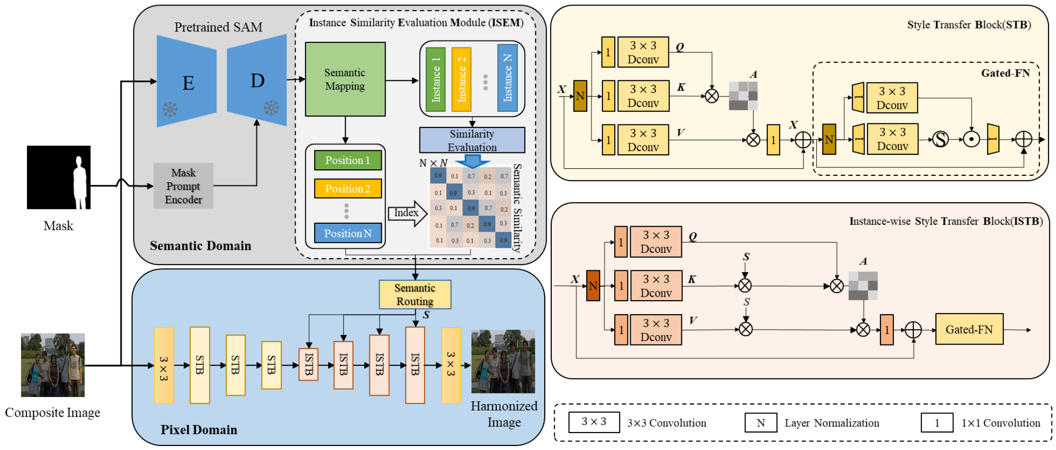

Figure 2: The overall structure of the Image harmonization model. The composite image first acquires instance information based on the SAM model and estimates the similarity matrix between instances. The harmonization model adopts an encoder-decoder structure. To build the global relationship between the background and foreground and explicitly utilize the instance similarity matrix, we design the STB and ISTB modules in the encoding and decoding stages, respectively.

background instances. As part of the harmonization procedure, we utilize a semantic routing technique that utilizes semantic similarity, which incorporates instance location and a semantic similarity matrix to deliberately adjust the feature representations within the image. To bolster the influence of analogous semantics, we employ an encoder-decoder network architecture. Here, the composite image is subject to convolutional encoding and then processed through three strata of the STB encoder. During decoding, to leverage the semantic similarity matrix in guiding the harmonization process, we introduce the Style Transfer Block (STB). This block shares a similar framework with STB, with a distinction in the attention mechanism where the Key-value matrix is modulated by the corresponding scale instance similarity matrix. This adjustment ensures alignment with semantic similarity and the subsequent refinement of the harmonization results. We apply a feature transformation function to ensure feature dimension consistency following each multiplication process. The process is formulated as:

$$K' = Reshape(K \times S) \tag{1}$$

$$V' = Reshape(V \times S) \tag{2}$$

Where K and K' are the input and output feature map, same to V and V'; S is the same scale instance similarity metrix obtained from the semantic routing module. Finally, following the traversal of a convolutional layer, we can get the harmonized image.

3.2 Instance Similarity Evaluation Module

We employ the pre-trained Segment Anything Model (SAM) [Kirillov *et al.*, 2023] on a comprehensive dataset for decomposing the composite image. SAM leverages foreground/background points, bounding boxes, or masks as prompts to produce segmentation results. It incorporates three primary components: an image encoder, a prompt encoder, and a mask decoder. Utilizing a pre-trained mask self-encoder based on the Vision Transformer (ViT), SAM processes the image into intermediary features while transforming the prompts into embedding tokens. The mask decoder's cross-attention mechanism then enables interactions between image features and prompt embeddings, culminating in the generation of the mask output. This process can be expressed as:

$$F_i = \phi(I_i) \tag{3}$$

$$F_p = \phi_{prompt}(Mask) \tag{4}$$

$$\hat{M} = \phi_{m_dec}(F_{img} + F_{c-mask}, [T_{out}, T_{prompt}]) \tag{5}$$

where F_i is the image feature, F_p is the prompt feature, $\hat{M}$ is the mask output, T_{out} and T_{prompt} are the output and prompt embedding tokens, respectively.

To derive the semantic representation of each instance, we initially employ the "full image" mode of SAM for segmenting all possible instance targets within the image. Subsequently, we introduce a semantic mapping module that ascertains the location and semantic details of instances, drawing from the image embedding produced by the SAM decoder.

Specifically, following the SAM decoder, the image embedding undergoes an up-sampling by a factor of $4\times$ via two transposed convolutional layers. The image tokens, labeled as E_{im} and incorporating prompt and output tokens, engage with the image embedding. The refreshed token embedding is then directed through three-layer MLP (Multi-Layer Perceptron) [Riedmiller and Lernen, 2014] modules to yield the instance embedding, represented as E_{in}. A spatial point-wise product is performed between the up-scaled image embedding and the instance embedding to predict the position of the instance, signified as P. This process can be expressed as:

$$E'_{im} = conv.Trans(E_{im}) \tag{6}$$

$$T_u = Attn(E_{im}, T) \tag{7}$$

$$P = E'_{im} \cdot MLP(T_u) \tag{8}$$

$$E_{in} = MLP(T_u) \tag{9}$$

Furthermore, we use a cross-similarity module to calculate the similarity between N instances. We use global average pooling to generate mean query feature $\bar{F}(E_{in})$. Then we copy it and make it have the same shape with the target feature E^i_{in}. The cross similarity map S has the same width/height with the number of instances detected. Mathematically, the similarity metric can be expressed as

$$q = \bar{F}(E_{in}) = GAP(F(E_{in})) \tag{10}$$

$$cos(E^i_{in}, q) = \frac{{E^i_{in}}^T \cdot q}{||E^i_{in}|| \cdot ||q||} \tag{11}$$

where $cos(\cdot)$ indicates the cosine similarity.

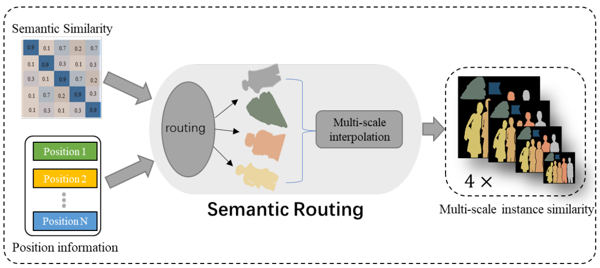

Figure 3: The illustration of the semantic routing.

3.3 Semantic Routing

To preserve the pronounced impact of background regions with analogous semantics on the foreground object, we introduce a semantic routing strategy predicated on assessing semantic similarity within the semantic space. As depicted in Figure 3, the semantic similarity matrix coupled with instance location data is employed to identify all feasible instances. By aligning semantic information with spatial location indices, we compute the correlation coefficient between background instances and foreground objects, subsequently generating a spatial importance map. In detail, the instance index of the position embedding is denoted as i and the corresponding value as S_i, it can be formulated as:

$$S_i = M_j, where\ i = j \tag{12}$$

where M is the semantic similar value from the semantic similarity matrix.

Upon finalizing the semantic-location mapping, the semantic similarity matrix is transformed into an instance similarity matrix. This matrix not only embeds instance location information but also encompasses correlation coefficients between background instances and foreground targets. To align with the Key-Value pairing mechanism in the multi-level STB, the similarity matrix is subject to interpolation operations, which yield a multi-scale matrix pyramid mirroring the scale structure of the STB.

3.4 Style Transfer Block

Style Transfer Block(STB) aims to integrate the spatial semantic and similarity information, which involves applying Self-Attention (SA) across channels instead of the spatial dimension. This allows us to compute cross-covariance across channels, resulting in the generation of an attention map that implicitly encodes the global context. We further enhance STB by introducing depth-wise convolutions, which emphasize the local context before calculating the feature covariance for producing the global attention map.

From a layer normalized tensor $Y \in \mathbb{R}^{H \times W \times C}$, our STB first generates query (Q), key (K) and value (V) projections, enriched with the local context. It is achieved by applying 1×1 convolutions to aggregate pixel-wise cross-channel context followed by 3×3 depth-wise convolutions to encode channel-wise spatial context, yielding $Q = W^Q_d W^Q_p Y$, $K = W^K_d W^K_p Y$ and $V = W^V_d W^V_p Y$. Where $W^{(\cdot)}_p$ is the 1×1 point-wise convolution and $W^{(\cdot)}_d$ is the 3×3 depth-wise convolution. We use bias-free convolutional layers in the network. Next, we reshape query and key projections such that their dot-product interaction generates a transposed-attention map A of size $\mathbb{R}^{C \times C}$, instead of the huge regular attention map of size $\mathbb{R}^{HW \times HW}$. Overall, the STB process is defined as:

$$\hat{X} = W_p \text{Attention}(\hat{Q}, \hat{K}, \hat{V}) + X, \tag{13}$$

$$\text{Attention}(\hat{Q}, \hat{K}, \hat{V}) = \hat{V} \cdot \text{Softmax}(\hat{K} \cdot \hat{Q} \alpha) \tag{14}$$

where X and $\hat{X}$ are the input and output feature maps; Q, K, V matrices are obtained after reshaping tensors from the original size $\mathbb{R}^{H \times W \times C}$. Here, α is a learnable scaling parameter to control the magnitude of the dot product of $\hat{K}$ and $\hat{Q}$ before applying the softmax function. Similar to the conventional multi-head SA, we divide the number of channels into heads and learn separate attention maps in parallel. To transform style features, the regular feed-forward network (FN) operates on each pixel location separately and identically. It uses two 1×1 convolutions, the first is used to expand the feature channels (usually by factor $\gamma = 4$) and the second is to reduce channels back to the original input dimension. A non-linearity is applied in the hidden layer.

In this work, we propose two fundamental modifications in FN to improve representation learning: (1) gating mechanism, and (2) depth-wise convolutions. The gating mechanism is formulated as the element-wise product of two parallel paths of linear transformation layers, one of which is activated with the GELU non-linearity. We include depth-wise convolutions to encode information from spatially neighboring pixel positions, useful for learning local image structure for effective restoration. Given an input tensor $X \in \mathbb{R}^{H \times W \times C}$, it is formulated as:

$$\hat{X} = W^0_p \text{Gating}(X) + X \tag{15}$$

$$Gating(X) = \phi(W^1_d W^1_p (LN(X))) \cdot W^2_d W^2_p (LN(X)) \tag{16}$$

where $(\cdot)$ denotes element-wise multiplication, ϕ represents the non-linearity, and LN is the layer normalization. Overall, the module controls the information flow through the respective hierarchical levels in our pipeline, thereby allowing each level to focus on the fine details complementary to the other levels.

Figure 4: Comparison with SOTA methods. Our results can obtain the similarity of instances in the background image and harmonize based on instances with high similarity. Therefore, they are able to better eliminate interference factors in the background.

4 Experiments

4.1 Datasets

Our experiments use the iHarmony4 dataset, a publicly available synthesized dataset referenced by Cong et al. [Cong *et al.*, 2020], which includes four sub-datasets: HCOCO, HAdobe5k, HFlickr, and Hday2night. These sub-datasets encompass synthesized composite images, foreground masks for these images, and their corresponding real images. We employed the same processing method as HDNet [Chen *et al.*, 2022] for the dataset. Additionally, to validation the performance of our methods in real-world scenarios, we employed 100 real-world images from CDTNet [Cong *et al.*, 2022], which are processed in the format of the iHarmony4 dataset.

Objective Evaluation Metrics. We evaluated the performance of our method using MSE, PSNR, fMSE, as suggested by [Cong *et al.*, 2020; Ling *et al.*, 2021; Niu *et al.*, 2023], in which fMSE means MSE within the foreground region. To illustrate performance, we qualitatively compare our method with following harmonization methods, including DoveNet [Cong *et al.*, 2020], Intrinsic [Guo *et al.*, 2021b], Bargainnet [Cong *et al.*, 2021], RainNet [Ling *et al.*, 2021], D-HT [Guo *et al.*, 2021a], Harmonizer [Ke *et al.*, 2022], SCS-Co [Hang *et al.*, 2022], CDTNet [Cong *et al.*, 2022], HDNet [Chen *et al.*, 2022], GKNet [Shen *et al.*, 2023], and LEMaRT [Liu *et al.*, 2023].

4.2 Implementation Details

Our model is trained by AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay $1e^{-4}$. We train the model for 200 epochs with input images resized to 256×256 and batch size set to 8. The initial learning rate is set to $3e^{-4}$ and gradually reduced to $1e^{-6}$ with the cosine annealing [Loshchilov and Hutter, 2017]. We use PyTorch to implement our models with NVIDIA GeForce RTX 4090.

4.3 Comparison with Existing Methods

Quantitative comparison Table 1 shows the quantitative results of previous image harmonization methods as well as our method. It is evident that our method surpasses the comparative methods across all datasets with the exception of MSE and fMSE on HCOCO. Furthermore, when contrasted with the second-best performing method, ours realizes a substantial average enhancement of $0.52dB$ in PSNR, a 0.55 reduction in MSE, and an improvement of 77.26 in fMSE.

Influence of fore-ground ratios Following [Cong *et al.*, 2020], we examine the influence of different fore-ground ratios on the harmonization models, i.e., 0% to 5%, 5% to 15%, 15% to 100%, and overall results. The comparative results of previous methods and our method are tabulated in Table 2. Upon scrutiny, it is evident that our method exhibits superior performance, outperforming all other approaches.

Qualitative comparison In Figure 4, Additionally, we provide a qualitative comparison of results on the iHarmony4 dataset. It is readily apparent that our method secures a more

model	venue	HCOCO		HAdobe5k		HFlickr		Hday2night		All	
		PSNR↑	MSE↓	PSNR↑	MSE↓	PSNR↑	MSE↓	PSNR↑	MSE↓	PSNR↑	MSE↓
Comp	-	33.99	69.66	28.48	347.52	28.41	266.05	34.3	110.95	31.76	173.43
Dovenet	CVPR'20	35.83	36.72	34.34	52.32	30.21	133.14	35.18	54.05	34.75	52.36
intrinsic	CVPR'21	37.21	24.92	36.01	43.02	36.23	105.13	34.03	55.53	35.01	38.71
BargainNet	ICME'21	37.03	24.84	39.94	35.34	31.34	97.32	35.67	50.98	35.88	37.82
RainNet	CVPR'21	37.08	29.52	36.22	43.35	31.64	110.59	34.83	57.4	36.12	40.29
D-HT	ICCV'21	38.33	16.89	36.11	38.53	33.13	75.51	37.1	53.01	37.55	30.3
Harmonizer	ECCV'22	38.77	17.34	37.64	21.89	33.63	64.81	37.56	33.14	37.84	24.26
SCS-Co	CVPR'22	39.88	13.58	38.29	21.01	34.22	55.83	37.83	41.75	38.75	21.33
CDTNet	CVPR'22	39.15	16.25	38.24	20.62	33.55	68.61	37.95	36.72	38.23	23.75
HDNet	MM'23	39.49	15.59	38.56	22.67	33.96	63.85	38.11	35.92	38.58	23.42
GKNet	ICCV'23	40.32	12.95	39.97	17.84	34.45	57.58	38.47	42.76	39.53	19.90
LEMaRT	CVPR'23	**41.0**	**10.1**	39.4	18.8	35.3	**40.7**	38.1	42.3	39.8	16.8
Ours	-	40.94	12.15	**40.91**	**14.77**	**35.79**	48.57	**39.30**	**27.00**	**40.32**	**17.25**

Table 1: Quantitative comparison across four sub-datasets of iHarmony4. **Bold** and underline indicate the best and second best performance, respectively.

model	$0\% \sim 5\%$		$5\% \sim 15\%$		$15\% \sim 100\%$		Average	
	MSE↓	fMSE↓	MSE↓	fMSE↓	MSE↓	fMSE↓	MSE↓	fMSE↓
Composite	28.51	1208.86	119.19	1323.23	577.58	1887.05	172.47	1387.30
DIH	18.92	799.17	64.23	725.86	228.86	768.89	76.77	773.18
$\mathrm{S^2AM}$	13.51	509.41	41.79	454.21	137.12	449.81	48.00	481.79
DoveNet	14.03	591.88	44.90	504.42	152.07	505.82	52.36	549.96
RainNet	11.66	550.38	32.05	378.69	117.41	389.80	40.29	469.60
BargainNet	10.55	450.33	32.13	359.49	109.23	353.84	37.82	405.23
Intrinsic	9.97	441.02	31.51	363.61	110.22	354.84	38.71	400.29
HDNet	5.95	230.75	20.32	265.31	68.95	318.15	23.42	258.80
ours	**4.37**	**198.47**	**13.50**	**155.61**	**52.55**	**172.11**	**17.25**	**181.54**

Table 2: We measure the error of different methods in fore-ground ratio range based on the whole test set. fMSE indicates the mean square error of the fore-ground region. Top performance are shown in **bold**.

uniform visual style across the entire composite image, resulting in a more photorealistic outcome. For example, as shown in the second row of Figure 4, the visual style of the foreground and the background are quite different, resulting in obvious image distortion. The other three methods cannot adjust the style of the foreground, especially the overall tone and the contrast of lighting and shadows. Unlike them, our method produces a more photo-realistic result and is closer to the ground-truth real image.

Overall Inference Time In Table 4, we present the inference time, parameter count, and FLOPs required for harmonizing a single image during testing. our approach does not show efficiency advantages, as indicated in the last row of Table 4, due to utilizing the pretrained SAM model for instance information retrieval. Yet, when relying solely on pixel domain architecture without ISEM, our model demonstrates comparable inference speed, with each step taking $20.4ms$ and a parameter count of $25.28M$, as shown in the third row of Table 4. In this study, we intentionally sacrificed some speed advantages to prioritize the realism of the harmonized images. Nonetheless, there is significant potential to enhance both the speed and parameter count of the SAM model, a direction we aim to pursue in future research.

4.4 Ablation Study

Effectiveness of each component In this section, we investigate the effectiveness of each component in our model.

Figure 5: Ablation study on ISEM and STB. Full model means baseline with both ISEM and STB

The results of ablating each component are reported in Table 3. Our ISEM module enables assess the similarity of components within both the semantic and stylistic domains of instances in the foreground and background. In Table 3, we can see that adding ISEM to the baseline brings 0.56 dB and 5.12 average performance improvement in terms of PSNR and MSE.

The STB effectively learns global style features and applies them to foreground objects. The addition of the STB enhances the overall coherence between foreground objects and background images. However, it also introduces a limitation in the form of excessive reliance on the background, which limits the effectiveness of improvement. In Table 3, we

Metric	HCOCO		HAdobe5k		HFlickr		Hday2night		All	
	PSNR↑	MSE↓	PSNR↑	MSE↓	PSNR↑	MSE↓	PSNR↑	MSE↓	PSNR↑	MSE↓
Comp	33.99	69.66	28.48	347.52	28.41	266.05	34.3	110.95	31.76	173.43
Basic	38.65	17.10	36.02	38.42	33.25	75.68	37.76	54.12	37.87	30.10
+ISEM	39.12	16.28	38.14	20.53	33.24	68.42	38.02	36.22	38.33	24.98
+STB	39.62	15.71	38.87	23.88	34.10	65.76	38.11	35.98	38.58	23.86
Total	**40.94**	**12.15**	**40.91**	**14.77**	**35.79**	**48.57**	**39.30**	**27.00**	**40.32**	**17.25**

Table 3: Ablation study across four sub-datasets of iHarmony4, Top performance are shown in **bold**

Method	Time(ms)	Params(M)	FLOPs(G)
RainNet	12.06	54.75	3.79
HDNet	15.08	10.41	48.04
CDTNet	10.8	24.36	78.05
Ours w/o ISEM	20.4	25.28	87.7
Ours	160.72	112.3	356.4

Table 4: Quantitative efficiency comparison of different methods.

can see that adding STB to the baseline brings 0.71dB and 5.24 average performance improvement in terms of PSNR and MSE.

By concurrently incorporating the ISEM and STB modules, our method effectively establishes correlations between various components of the target object and background instances, thus enhancing overall coherence. Consequently, the improvement is significantly pronounced. In Table 3, we can see that adding both ISEM and STB to the baseline brings 2.45 and 12.85 average performance improvement in terms of PSNR and MSE.

Visual comparison To further illustrate the effectiveness of our mothods, we show some output results of ablation experiments in Figure 5. It can be found that compared with the distortion results produced by the module, the full model's results performe more consistent in lighting and color with background regions.

4.5 User Study

We extend our evaluation by comparing various methods using a dataset of 100 real composite images provided by CDTNet [Cong *et al.*, 2022]. To gauge the performance against competitive baselines, we conduct a user study. This study involves the construction of 600 image pairs, in which we randomly select two images from each composite image and its 3 corresponding harmonized results across the 100 real composite images. Subsequently, we allocate 60 pairs for each of the 20 participants, who are tasked with viewing one image pair at a time and selecting the image they perceive as more harmonious. This process generates a total of 1200 pairwise results. Following the methodology adopted in GiftNet [Niu *et al.*, 2023], we computed the Bradley-Terry(B-T) scores for all methods, as detailed in Table 5. Notably, our approach emerges with the highest B-T score (which is 0.413) concerning realism, underscoring the efficacy of the method proposed in this paper. The visualization results pertaining to real composite images are presented in Figure 6. Compared to previous methods, our results demonstrate enhanced realism, particularly evident when similar instances are present in the background, as illustrated in the first three rows. Furthermore, when there

Method	Composite	RainNet	HDNet	CDTNet	Ours
B-T Score	-0.972	0.084	0.177	0.298	**0.413**

Table 5: B-T scores of different methods on 100 real composite images.

(a) Comp (b) Masks (c) RainNet (d) HDNet (e) CDTNet (f) Ours

Figure 6: The visualization of different methods on real composite images.

are $N(N > 0)$ related instances in the background, the model constructs an N-dimensional similarity matrix to represent the degree of similarity between instances. These instances affect the foreground through weighted accumulation across the matrix, and the foreground maintains good consistency with the most relevant instances, such as the color of sunflowers in the 3rd row of Figure 6. Furthermore, in the absence of similar instances, the proposed STB and ISTB, which can capture and transfer global color information into the foreground, can maintain overall appearance consistency throughout the image, as illustrated in the 4th row of Figure 6.

5 Conclusion

In this paper, we propose a image harmonization model utilizing instance similarity to maintain consistency uniformity in global and similar regions. We propose an instance similarity evaluation module (ISEM), which can assess the similarity of components within both the semantic and stylistic domains of instances in the foreground and background. We introduce a style transfer block(STB) that captures the global style information of the input image and transfers it to the latent space of the style encoder. Our method has achieved excellent experimental results on existing datasets and has more significant advantages in user visual reality evaluation.

Acknowledgments

This work is supported in part by the National Science and Technology Major Project of China under Grant 2022ZD0118001; and in part by the National Science Foundation of China under Grant 62332017, Grant 62303043, Grant U22A2022, Grant 62372036, Grant 62120106009.

References

[Chen *et al.*, 2022] Haoxing Chen, Zhangxuan Gu, Yaohui Li, Jun Lan, Changhua Meng, Weiqiang Wang, and Huaxiong Li. Hierarchical dynamic image harmonization. *arXiv preprint arXiv:2211.08639*, 2022.

[Cong *et al.*, 2020] Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. Dovenet: Deep image harmonization via domain verification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8391–8400. IEEE, 2020.

[Cong *et al.*, 2021] Wenyan Cong, Li Niu, Jianfu Zhang, Jing Liang, and Liqing Zhang. Bargainnet: Background-guided domain translation for image harmonization. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021.

[Cong *et al.*, 2022] Wenyan Cong, Xinhao Tao, Li Niu, Jing Liang, Xuesong Gao, Qihao Sun, and Liqing Zhang. High-resolution image harmonization via collaborative dual transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18470–18479, 2022.

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

[Guo *et al.*, 2021a] Zonghui Guo, Dongsheng Guo, Haiyong Zheng, Zhaorui Gu, Bing Zheng, and Junyu Dong. Image harmonization with transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14870–14879, 2021.

[Guo *et al.*, 2021b] Zonghui Guo, Haiyong Zheng, Yufeng Jiang, Zhaorui Gu, and Bing Zheng. Intrinsic image harmonization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16367–16376, 2021.

[Hang *et al.*, 2022] Yucheng Hang, Bin Xia, Wenming Yang, and Qingmin Liao. Scs-co: Self-consistent style contrastive learning for image harmonization, 2022.

[Hao *et al.*, 2020] Guoqing Hao, Satoshi Iizuka, and Kazuhiro Fukui. Image harmonization with attention-based deep feature modulation. In *The British Machine Vision Conference (BMCV)*, page 4, 2020.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models, 2020.

[Kaur *et al.*, 2023] Gurpreet Kaur, Navdeep Singh, and Munish Kumar. Image forgery techniques: a review. *Artificial Intelligence Review*, 56(2):1577–1625, 2023.

[Ke *et al.*, 2022] Zhanghan Ke, Chunyi Sun, Lei Zhu, Ke Xu, and Rynson WH Lau. Harmonizer: Learning to perform white-box image and video harmonization. In *European Conference on Computer Vision*, pages 690–706. Springer, 2022.

[Kim *et al.*, 2022] Kihong Kim, Yunho Kim, Seokju Cho, Junyoung Seo, Jisu Nam, Kychul Lee, Seungryong Kim, and KwangHee Lee. DiffFace: Diffusion-based Face Swapping with Facial Guidance, 2022.

[Kirillov *et al.*, 2023] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[Li *et al.*, 2023] Ruibin Li, Jingcai Guo, Song Guo, Qihua Zhou, and Jie Zhang. Freepih: Training-free painterly image harmonization with diffusion model. *arXiv preprint arXiv:2311.14926*, 2023.

[Liang and Pun, 2022] Jingtang Liang and Chi-Man Pun. Image harmonization with region-wise contrastive learning, 2022.

[Ling *et al.*, 2021] Jun Ling, Han Xue, Li Song, Rong Xie, and Xiao Gu. Region-aware adaptive instance normalization for image harmonization. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9357–9366. IEEE, 2021.

[Liu *et al.*, 2023] Sheng Liu, Cong Phuoc Huynh, Cong Chen, Maxim Arap, and Raffay Hamid. Lemart: Label-efficient masked region transform for image harmonization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18290–18299, 2023.

[Loshchilov and Hutter, 2017] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017.

[Lu *et al.*, 2023] Lingxiao Lu, Jiangtong Li, Junyan Cao, Li Niu, and Liqing Zhang. Painterly image harmonization using diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 233–241, 2023.

[Niu *et al.*, 2023] Li Niu, Linfeng Tan, Xinhao Tao, Junyan Cao, Fengjun Guo, Teng Long, and Liqing Zhang. Deep image harmonization with globally guided feature transformation and relation distillation, 2023.

[Perez *et al.*, 2023] Patrick Perez, Michel Gangnet, and Andrew Blake. Poisson image editing. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 577–582. 2023.

[Ren and Liu, 2022] Xuqian Ren and Yifan Liu. Semantic-guided multi-mask image harmonization. In *European Conference on Computer Vision*, pages 564–579. Springer, 2022.

[Riedmiller and Lernen, 2014] Martin Riedmiller and A Lernen. Multi layer perceptron. *Machine Learning Lab Special Lecture, University of Freiburg*, 24, 2014.

[Shen *et al.*, 2023] Xintian Shen, Jiangning Zhang, Jun Chen, Shipeng Bai, Yue Han, Yabiao Wang, Chengjie Wang, and Yong Liu. Learning global-aware kernel for image harmonization. *arXiv preprint arXiv:2305.11676*, 2023.

[Sofiiuk *et al.*, 2021] Konstantin Sofiiuk, Polina Popenova, and Anton Konushin. Foreground-aware semantic representations for image harmonization. In *Foreground-Aware Semantic Representations for Image Harmonization*, pages 1619–1628. IEEE, 2021.

[Song *et al.*, 2023] Seokbeom Song, Suhyeon Lee, Hongje Seong, Kyoungwon Min, and Euntai Kim. Shunit: Style harmonization for unpaired image-to-image translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2292–2302, 2023.

[Sunkavalli *et al.*, 2010] Kalyan Sunkavalli, Micah K Johnson, Wojciech Matusik, and Hanspeter Pfister. Multi-scale image harmonization. *ACM Transactions on Graphics (TOG)*, 29(4):1–10, 2010.

[Tsai *et al.*, 2017] Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. Deep image harmonization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3789–3797, 2017.

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

[Zhu *et al.*, 2022] Ziyue Zhu, Zhao Zhang, Zheng Lin, Ruiqi Wu, Zhi Chai, and Chun-Le Guo. Image harmonization by matching regional references, 2022.