LEMaRT: Label-Efficient Masked Region Transform for Image Harmonization

Sheng Liu Cong Phuoc Huynh Cong Chen Maxim Arap Raffay Hamid
Amazon Prime Video
{shenlu, conghuyn, checongt, maxarap, raffay}@amazon.com

Abstract

*We present a simple yet effective self-supervised pre-training method for image harmonization which can leverage large-scale unannotated image datasets. To achieve this goal, we first generate pre-training data online with our **L**abel-**E**fficient **Ma**sked **R**egion **T**ransform (LEMaRT) pipeline. Given an image, LEMaRT generates a foreground mask and then applies a set of transformations to perturb various visual attributes,* e.g.*, defocus blur, contrast, saturation, of the region specified by the generated mask. We then pre-train image harmonization models by recovering the original image from the perturbed image. Secondly, we introduce an image harmonization model, namely SwinIH, by retrofitting the Swin Transformer [27] with a combination of local and global self-attention mechanisms. Pre-training SwinIH with LEMaRT results in a new state of the art for image harmonization, while being label-efficient,* i.e.*, consuming less annotated data for fine-tuning than existing methods. Notably, on iHarmony4 dataset [8], SwinIH outperforms the state of the art,* i.e.*, SCS-Co [16] by a margin of* 0.4 *dB when it is fine-tuned on only* 50% *of the training data, and by* 1.0 *dB when it is trained on the full training dataset.*

1. Introduction

The goal of image harmonization is to synthesize photo-realistic images by extracting and transferring foreground regions from an image to another (background) image. The main challenge is the appearance mismatch between the foreground and the surrounding background, due to differences in camera and lens settings, capturing conditions, such as illumination, and post-capture image processing. Image harmonization aims to resolve this mismatch by adjusting the appearance of the foreground in a composite image to make it compatible with the background. Research in image harmonization has relevant applications in photo-realistic image editing and enhancement [42,44], video synthesis [23,37] and data augmentation for various computer vision tasks [11,12,35].

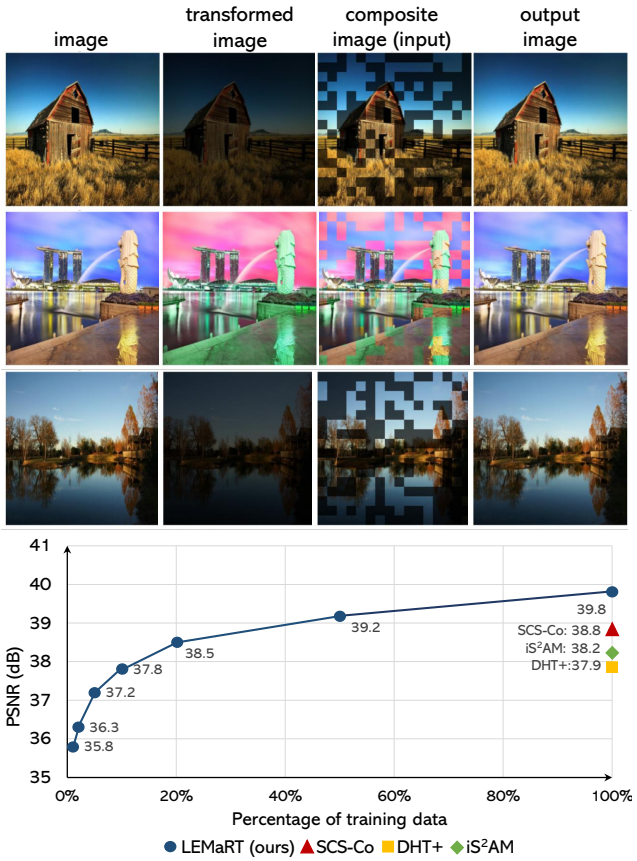

Figure 1. Top: given an image, LEMaRT applies a set of transformations, *e.g.*, brightness, hue adjustment, to obtain a transformed image. The transformed image is then combined with the original image to form a composite image, which is used to pre-train our SwinIH image harmonization model. As shown in the right-hand column, SwinIH is capable of reconstructing photo-realistic output images after pre-training and fine-tuning. Bottom: using our *LEMaRT* pre-training scheme, our image harmonization model (*SwinIH*) surpasses state of the art (SOTA) counterparts with less than 40% of the training data from iHarmony4 for fine-tuning.

Traditional image harmonization approaches perform color transforms to match the low-level color statistics of the foreground to the background with the aim to achieve photorealism [22, 31, 33, 39]. However, the generalization

ability of these methods is questionable because the evaluation was only conducted at a small scale, mainly using human judgement. More recent works [8] have constructed real image harmonization datasets with tens to thousands of images to train learning-based methods. However, due to the bottleneck of manual editing, these datasets do not match the scale often required to train large-scale neural networks. Rendered image datasets [3, 15] are more scalable but they suffer from the domain gap between synthetic and real images. As a result, the performance of image harmonization models is constrained by the limited size of a few existing datasets [8, 20] on which they can be trained.

Inspired by the impressive performance leap achieved by pre-trained models [17, 29] on various downstream tasks, *e.g.*, image classification, object detection, image captioning, in this work, we introduce a novel self-supervised pre-training method to boost the performance of image harmonization models while being label-efficient, *i.e.*, consuming small amounts of fine-tuning data. The novelty of our technique lies in the use of foreground masking strategies and the perturbation of foreground visual attributes to self-generate training data without annotations. Hence, we name our pre-training method as **L**abel-**E**fficient **Ma**sked **R**egion **T**ransform (LEMaRT). In the first step, LEMaRT proposes pseudo foreground regions in an image. Subsequently, it applies a set of transformations to perturb visual attributes of the foreground, including contrast, sharpness, blur and saturation. These transformations aim to mimic the appearance discrepancy between the foreground and the background. Using the transformed image, *i.e.*, image with the perturbed foreground, as the input, LEMaRT pre-trains image harmonization models to reconstruct the original image, as shown in the top half of Figure 1.

Subsequently, we design an image harmonization model based on Swin Transformer [27], namely SwinIH, which is short for Swin Image Harmonization. We build our model upon Swin Transformer instead of the ViT model [10] mainly due to the efficiency gain offered by its local shifted window (Swin) attention. Similar to the design of the original Swin Transformer, we keep the local self-attention mechanism in all the Transformer blocks up except the last one, where we employ global self-attention. We introduce global self-attention into SwinIH to alleviate block boundary artifacts produced by the Swin Transformer model when it is directly trained for image harmonization.

We verify that LEMaRT consistently improves the performance of models with a range of vision Transformer and CNN architectures compared to training only on the target dataset, *e.g.*, iHarmony4. When we pre-train our SwinIH model on MS-COCO dataset with LEMaRT and then fine-tune it on iHarmony4 [8], it outperforms the state of the art [16] by 0.4 dB while using only *50*% of the samples from iHarmony4 for fine-tuning, and by 1.0 dB when using all the samples (see the plot in the bottom half of Figure 1).

The **key contributions** of our work are summarized below.

- We introduce Label-Efficient Masked Region Transform (LEMaRT), a novel pre-training method for image harmonization, which is able to leverage large-scale unannotated image datasets.
- We design SwinIH, an image harmonization model based on the Swin Transformer architecture [27].
- LEMaRT (SwinIH) establishes new state of the art on iHarmony4 dataset, while consuming significantly less amount of training data. LEMaRT also boosts the performance of models with various network architectures.

2. Related Work

a. Image Harmonization: Most early works extract and match low-level color statistics of the foreground and its surrounding background. These works rely on color histograms [39], multi-level pyramid representations [33], color clusters [22], etc. The limited representation power of low-level features negatively affects their performance.

More recent works [8, 20] have constructed datasets at a reasonable scale to advance learning-based methods. Numerous supervised deep learning-based image harmonization models have been trained on these datasets [9, 14, 15, 26]. Tsai *et al*. [34] combine image harmonization and semantic segmentation under a multi-task setting. S^2AM [9] proposes to predict a foreground mask and to adjust the appearance of foreground with spatial-separated attention. RainNet [26] transfers statistics of instance normalization layers from the background to the foreground. In addition, generative models have also been trained for image harmonization [4, 8, 45].

Some state of the art (SOTA) methods formulate image harmonization as a style transfer problem. These methods learn a domain representation of the foreground and background with contrastive learning [7] or by maximizing mutual information between the foreground and background [24]. More recently, Hang *et al*. [16] have advanced state of the art results by adding background and foreground style consistency constraints and dynamically sampling negative examples within a contrastive learning paradigm. Using only a reconstruction loss during pre-training and fine-tuning, our method is able to outperform [16] with a much simpler training set up.

b. Transfer Learning: Transfer learning is a well-known and effective technique for adapting a pre-trained model to a downstream task, especially with limited training data [5, 6, 18]. Recent advances in foundation models [1, 19, 29, 36, 40, 41, 43] have resulted in models that can be adapted to a wide range of downstream tasks. Sofiiuk *et al*. [30] propose an image harmonization model which takes visual features extracted from a pre-trained segmentation model as an

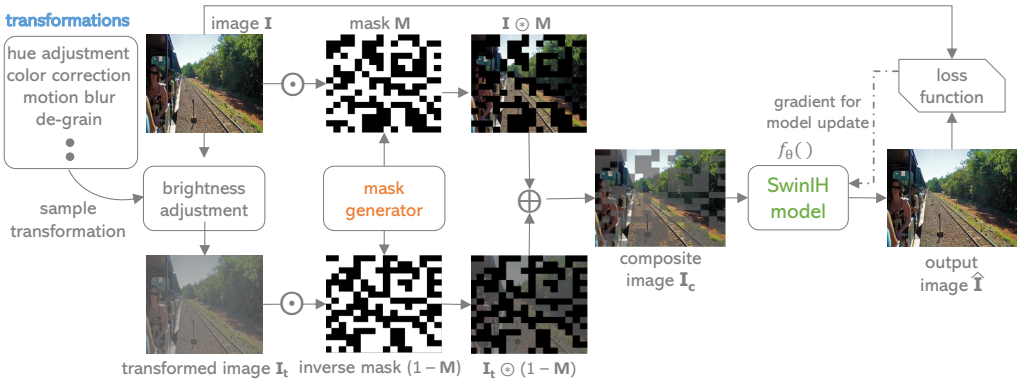

Figure 2. Our online data generation and pre-training pipeline (LEMaRT). LEMaRT generates the input composite image $\mathbf{I}_c$ for the pre-training process via masked region transform. The goal of pre-training is to learn an image harmonization model $f_\theta(\cdot)$, e.g. our SwinIH model, that can reconstruct the original image $\mathbf{I}$ from the composite image $\mathbf{I}_c$.

auxiliary input. Instead of leveraging a pre-trained segmentation model for feature extraction, we specifically pre-train a model for image harmonization. We opt for this direction based on the hypothesis that pre-training for the same target task results in better performance than pre-training for a different task. Inspired by [18], our LEMaRT method is more suitable for image harmonization than [18] because LEMaRT creates training samples by applying transformations to the foreground rather than masking the foreground, which makes the pre-training task closer to image harmonization. In addition, [18] introduces an asymmetric encoder-decoder architecture, while our SwinIH model is specifically designed for image harmonization and does not have an explicit encoder or a decoder.

3. Method

3.1. Problem Formulation

The goal of image harmonization is to synthesize photo-realistic images by extracting and transferring foreground regions from an image $\mathbf{I}_1$, specified by a binary mask $\mathbf{M}$, to another (background) image $\mathbf{I}_2$. Let $\mathbf{I}_c = \mathbf{M} \odot \mathbf{I}_1 \oplus (1 - \mathbf{M}) \odot \mathbf{I}_2$ be the composite image generated by a direct copy and paste of the foreground region from $\mathbf{I}_1$ on top of $\mathbf{I}_2$. The operators $\odot$ and $\oplus$ denote element-wise multiplication and addition, respectively. Subsequently, an image harmonization function $f(\cdot)$ transforms the composite image $\mathbf{I}_c$ into a harmonized image $\hat{\mathbf{I}} = f(\mathbf{I}_c)$, such that the latter is photo-realistic. Deep learning-based image harmonization methods implement this function as a neural network $f_\theta(\cdot)$ with parameters denoted by θ. Our goal is to learn θ via self-supervised pre-training, so that the function $f_\theta(\cdot)$ can generate photo-realistic images.

3.2. Online Pre-training Data Generation

We first introduce our data generation and pre-training pipeline, *i.e.*, LEMaRT that generates the input and the ground truth for the pre-training process without relying on any manual annotations. As shown in Figure 2, LEMaRT applies a set of random transformations such as hue, contrast, brightness adjustment and defocus blur to perturb the original image $\mathbf{I}$. The generated image is referred to as the transformed image $\mathbf{I}_t$. The random transformations are designed to mimic different kinds of visual mismatches between a foreground region and a background image. In addition, LEMaRT employs a mask generation strategy to propose a foreground mask $\mathbf{M}$ (please refer to § 3.3 for more details). The mask and the set of transformations are generated on the fly for each input image $\mathbf{I}$. With these ingredients, we generate a composite image $\mathbf{I}_c$ from the original image $\mathbf{I}$ and the transformed image $\mathbf{I}_t$ as $\mathbf{I}_c = \mathbf{M} \odot \mathbf{I}_t \oplus (1 - \mathbf{M}) \odot \mathbf{I}$. We note that the composite image $\mathbf{I}_c$ is generated and fed into the network in an *online* fashion.

The goal of pre-training is to learn a harmonization function $f_\theta(\cdot)$ to resolve the mismatch of visual appearance between the foreground and the background of the composite image $\mathbf{I}_c$. We formulate this task as the reconstruction of the original image $\mathbf{I}$ from the composite image $\mathbf{I}_c$. The original image $\mathbf{I}$ serves as the supervision signal for the pre-training process under the assumption that visual elements of real images are in harmony. This formulation is applicable to a wide range of network architectures, *e.g.*, SwinIH (see § 3.4 for details), ViT [10] and CNN models [23, 32].

3.3. Mask Generation

We now present three foreground mask generation strategies, which we refer to as *random*, *grid* and *block*.

• *random*: as shown in Figure 4, this strategy first partitions an image into a regular pattern that consists of $m \times m$ even patches (labelled by white pixels). It then generates a mask by randomly selecting a subset of the image patches.

• *grid*: similar to *random*, this strategy first partitions an image into regular pattern of $m \times m$ even patches. It then

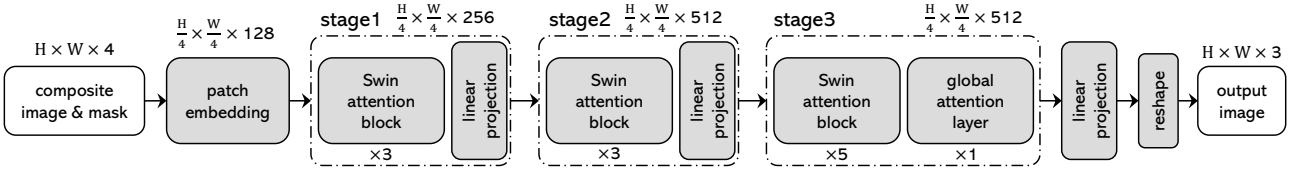

Figure 3. Illustration of our SwinIH model, *i.e.*, a Transformer-based image harmonization model.

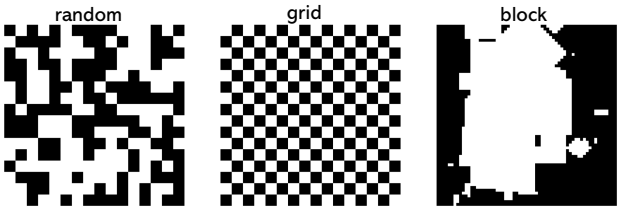

Figure 4. Sample masks generated by the three strategies, *i.e.*, random, grid and block, introduced in § 3.3.

generates a *fixed* mask (same for all images) by selecting image patches following the pattern shown in Figure 4.

• *block*: inspired by [2], we design a mask generation strategy that attempts to mimic the *shape* of objects. It generates a mask in an iterative manner. Each iteration has two steps, *i.e.*, (1) generating a rectangular region; (2) applying a random homography to the rectangular region to make the boundary of the region be composed of slanted lines. If the rectangular region is smaller than the size of the desired region to be masked, we generate a new region by executing the two steps once more and merge the newly generated region with previously generated regions (please refer to the supplementary materials for more details).

3.4. Network Architecture

Since the pre-training process is agnostic to network architecture, the only constraint for model design is that it needs to generate an output image of the same size as the input. To this end, we choose to implement our image harmonization model based on a Transformer architecture, due to the recent successes of vision transformer models in various tasks, including image harmonization [13, 14].

We choose to build our SwinIH model upon Swin attention blocks [27] due to its improved performance and efficiency over global attention layers used by ViT-style models [10, 13, 14]. Let H and W denote the height and width of an input image $\mathbf{I}_c$ and N denote the size of an image patch. The length of the sequence of visual tokens is $\frac{H \cdot W}{N^2}$. A global attention layer has a space and time complexity of $\mathcal{O}(\frac{H^2 \cdot W^2}{N^2})$. On the contrary, the space and time complexity of a Swin attention block is $\mathcal{O}(\frac{H \cdot W \cdot K^2}{N^2})$, where the shifted window size K is smaller than H and W.

The architecture of our SwinIH model is shown in Figure 3. It takes a four-channel image (a channel-wise concatenation of a composite image $\mathbf{I}_c$ and a foreground mask $\mathbf{M}$) as input and generates an output image $\hat{\mathbf{I}}$. Our SwinIH model is composed of three stages. The first two stages consist of three Swin attention blocks. The third stage has five Swin attention blocks and a global attention layer. We set the dimension of the patch embedding to 128 and double it at the end of the first two stages with linear projections.

Unlike Swin Transformer [27] which processes its input in a multi-scale manner[1], we choose to preserve the original resolution of the input. As will be shown in § 4.4, such a design choice is important as the information loss due to the reduced resolution hurts model accuracy.

3.5. Objective Function

We adopt the mean squared error (MSE) between the network's output $\hat{\mathbf{I}}$ and the original image $\mathbf{I}$ as the objective function for pre-training. When we fine-tune the pre-trained network, we follow [30] and use a foreground-normalized MSE loss as the objective function.

4. Experiments

We evaluate our method by comparing its performance with other state-of-the-art (SOTA) methods and provide insights into our method through ablation studies. We adopt four metrics, *i.e.*, mean squared error (MSE), peak signal to noise ratio (PSNR), foreground mean squared error (fMSE) [8], and foreground peak signal to noise ratio (fPSNR) [8].

4.1. Datasets

Following previous works [7,8,15], we evaluate our method on iHarmony4 dataset [8]. For completeness, we also evaluate our method on RealHM dataset [20]. Unless otherwise stated, we pre-train our LEMaRT model on the set of 120K unlabeled images from the MS COCO dataset [25] and fine-tune on iHarmony4. There is no overlap between the images used for pre-training and the images used for fine-tuning and evaluation. Images in iHarmony4 either come from the set of labeled images in MS COCO, which is disjoint from the unlabeled images in MS COCO, or from other datasets. Following [7, 8, 15, 30, 34], we resize the input images and the ground truth images to 256×256.

[1] Swin Transformer [27] uses a Patch Merging layer to reduce the spatial size of its input by a factor of two at the end of each stage.

dataset	metric	composite image	DIH [34]	S^{2}AM [9]	DoveNet [8]	BargNet [7]	IntrHarm [15]	RainNet [26]	iS^{2}AM [30]	DHT+ [13]	SCS-Co [16]	LEMaRT (SwinIH)
HCOCO	PSNR↑	33.9	34.7	35.5	35.8	37.0	37.2	37.1	39.2	39.2	*39.9*	**41.0** ↑1.1
	MSE↓	69.4	51.9	41.1	36.7	24.8	24.9	29.5	16.5	15.0	*13.6*	**10.1** ↓3.5
HAdobe	PSNR↑	28.2	32.3	33.8	34.3	35.3	35.2	36.2	38.1	37.2	*38.3*	**39.4** ↑1.1
	MSE↓	345.5	92.7	63.4	52.3	39.9	43.0	43.4	21.9	36.8	*21.0*	**18.8** ↓2.2
HFlickr	PSNR↑	28.3	29.6	30.0	30.2	31.3	31.3	31.6	33.6	33.6	*34.2*	**35.3** ↑1.1
	MSE↓	264.4	163.4	143.5	133.1	97.3	105.1	110.6	69.7	67.9	*55.8*	**40.7** ↓15.1
HD2N	PSNR↑	34.0	34.6	34.5	35.3	35.7	36.0	34.8	37.7	36.4	*37.8*	**38.1** ↑0.3
	MSE↓	109.7	82.3	76.6	52.0	51.0	55.5	57.4	**40.6**	49.7	*41.8*	42.3 ↑1.7
all	PSNR	31.6	33.4	34.3	34.8	35.9	35.9	36.1	38.2	37.9	*38.8*	**39.8** ↑1.0
	MSE↓	172.5	76.8	59.7	52.3	37.8	38.7	40.3	24.4	27.9	*21.3*	**16.8** ↓4.5

Table 1. Our pre-trained image harmonization model, LEMaRT, outperforms state-of-the-art (SOTA) models on iHarmony4. The column named *composite image* shows the result for the direct copy and paste of foreground regions on top of background images.

4.2. Implementation Details

We use an AdamW optimizer [28] both during pre-training and fine-tuning. We set $\beta_1 = 0.9$, $\beta_2 = 0.95$, $\epsilon = 1e^{-8}$ and weight decay to 0.05. The window size and the patch size of SwinIH are set to 32 and 4, respectively. We pre-train our model for 30 epochs with a batch size of 192 and a learning rate of $2.7e^{-2}$. We then fine-tune the pre-trained model for 120 epochs with a learning rate of $2.7e^{-3}$. A cosine annealing scheduler controls the change of learning rate. The minimum learning rate is set to 0.0. We adopt the random mask generation strategy and set mask ratio to 50% during pre-training. This is the default setting for the experiments.

4.3. Comparison with SOTA Methods

a. On iHarmony4 Dataset

In Table 1, we present a comparison between the performance of our method, LEMaRT (SwinIH), and the performance of existing methods on iHarmony4. Overall, LEMaRT comprehensively outperforms existing methods across the two metrics (PSNR and MSE). Most notably, our method achieves a PSNR of 39.8 dB, which is 1.0 dB higher than the previous best method. The MSE of our method is 16.8, which is 4.5 lower (21.1% relative improvement) than the previous best method [16] [2].

We notice that our method, LEMaRT, consistently achieves better performance than SOTA methods [7–9, 13, 15, 16, 26, 30, 34] on three of the four subsets, *i.e.*, HCOCO, HAdobe and HFlickr of iHarmony4. Meanwhile, on the HD2N subset, the performance of our method is on par with SOTA methods. While our method yields higher PSNR, the MSE of our method is higher. We hypothesize that the domain of MS COCO, the dataset which we use to pre-train LEMaRT, is not closely aligned with that of HD2N. For example, mountains and buildings are the salient objects in most images in the HD2N subset. However, they do not often appear as the main objects in MS COCO images.

In Figure 5, we compare the harmonized images generated by three SOTA methods, *i.e.*, RainNet, iS^{2}AM, DHT+, and our method, *i.e.*, LEMaRT (SwinIH). We see that LEMaRT can generate photo-realistic images. Compared to other methods, LEMaRT is better at making color corrections, thanks to the pre-training process during which LEMaRT learns the distribution of photo-realistic images.

b. On RealHM Dataset

In Table 2, we compare the performance of our method, LEMaRT, with multiple SOTA methods on RealHM dataset. We pre-train our model on 120K images from Open Images V6 [21] for 22 epochs and then fine-tune our model on iHarmony4 for 1 epoch with a learning rate of $5.3e^{-3}$. We see that LEMaRT comfortably outperforms DoveNet [8] and S^{2}AM [9], and achieves comparable results to SSH [20]. A comparison of the harmonized images generated by our LEMaRT method and existing methods can be found in the supplementary materials.

method	DoveNet [8]	S^{2}AM [9]	SSH [20]	LEMaRT
MSE ↓	214.1	283.3	206.9	**206.1**
PSNR ↑	27.4	26.8	**27.9**	27.6

Table 2. Comparison between our pre-trained model, LEMaRT, and SOTA models on RealHM.

4.4. Ablation Studies

We conduct ablation studies to gain insights into various aspects of our method. These aspects include the general-

[2] Our method also outperforms SOTA methods across the iHarmony4 dataset in terms of fPSNR and fMSE. For brevity, we omit them in Table 1 and include them in supplementary materials instead.

dataset	metric	SwinIH w/o	SwinIH w/	ViT [10] w/o	ViT [10] w/	ResNet [38] w/o	ResNet [38] w/	HRNet [32] w/o	HRNet [32] w/	HT+ [13] w/o	HT+ [13] w/
all	PSNR↑	37.0	39.0	35.7	38.4	34.6	36.3	33.2	35.3	37.7	38.9
	MSE↓	35.5	20.9	48.2	24.1	64.4	44.2	78.5	48.4	31.4	22.3
	fPSNR↑	24.5	26.6	23.1	25.9	21.9	23.4	20.6	22.7	25.1	26.3
	fMSE↓	386.6	250.0	499.9	282.6	645.3	459.9	811.9	510.6	342.6	266.5

Table 3. Effect of pre-training on different image harmonization models, *i.e.*, our SwinIH, ViT [10], ResNet [38], HRNet [32] and HT+ [13]. We compare their performance when they are trained on iHarmony4 from scratch (*w/o* columns), and when they are fine-tuned after being pre-trained with LEMaRT using the unlabeled images in MS COCO (*w/* columns).

ization ability of our LEMaRT method across different network architectures, its efficiency in terms of data and annotation consumption, the design choices of our SwinIH model, and the sensitivity of its performance to the mask generation strategy and the mask size.

a. Generalization Across Network Architectures

The goal of the first ablation study is to understand the effectiveness of the proposed pre-training method, *i.e.*, LEMaRT, on various network architectures, including vision Transformers and convolutional neural networks (CNNs). Specifically, we adopt five different networks, *i.e.*, SwinIH, ViT [10], ResNet [38], HRNet [32] and HT+ [13]. SwinIH refers to our model introduced in § 3.4. ViT refers to the vision Transformer model that adopts global attention. ResNet is a variant of the ResNet generator introduced in pix2pixHD [38]. We remove the down sampling operators to make it suitable for image harmonization. We re-implement HT+ [13], a ViT-style Transformer model designed for image harmonization. Our implementation has comparable results (0.3 dB higher PSNR, and 0.9 higher MSE) with those reported in [13]. We compare the performance of the five networks when they are trained on iHarmony4 from scratch (*w/o* columns in Table 3), and when they are fine-tuned after being pre-trained with LEMaRT (*w/* columns). We train (or fine-tune) SwinIH, ViT, ResNet and HRNet on iHarmony4 for 30 epochs, and HT+ for 120 epochs to be consistent with the results reported in [13]. Other settings are kept the same as the default setting.

As shown in Table 3, pre-training on MS COCO with LEMaRT significantly improves performance of the models under study over training from scratch on iHarmony4. Specifically, the performance boost ranges from 1.2 to 2.7 dB in terms of PSNR and 1.2 to 2.8 dB in terms of fPSNR. In particular, LEMaRT improves the PSNR of our SwinIH model by 2.0 dB and its MSE by 14.6. Moreover, LEMaRT is effective not only for models adapted from other vision tasks, but also for those specifically designed for image harmonization, such as HT+ [13].

b. Data Efficiency

Next, we evaluate the effectiveness of LEMaRT with respect to the amount of fine-tuning data. As before, we pre-train our SwinIH model in two settings: training from scratch on iHarmony4 only and pre-training followed by fine-tuning. For both settings, we vary the amount of fine-tuning data by uniformly sampling between 1% and 100% of the iHarmony4 training set.

The results in Figure 6 are consistent with the previous section, in the sense that pre-training improves image harmonization accuracy by a large margin (up to 2.4 dB in terms of PSNR and 299.1 in terms of MSE) regardless of the amount of fine-tuning data. More importantly, the LEMaRT pre-training scheme is more beneficial to the low data regime than the high data regime. For example, when using no more than 10% of the fine-tuning data, the performance boost attributed to pre-training ranges between 2.3 and 2.4 dB, whereas the improvement at 100% of fine-tuning data declines to 1.4 dB. We observe a similar trend in the MSE measure, where the MSE improvement drops from around 300.0 at 1% of iHarmony4 training data to less than 90.0 when using the full training set.

c. Model Design Choices

In this experiment, we pay attention to two design choices of our SwinIH model. The first choice is to maintain the input resolution across all the transformer blocks or to adopt a bottleneck layer similar to encoder-decoder models. The second choice is whether to use efficient local attention, *i.e.*, Swin attention, across all the blocks or to use global attention as well. This choice stems from a visual observation that the Swin attention *occasionally* induces block-shaped visual artifacts in harmonized images, as shown in Figure 7. Therefore, it prompts the necessity to modify model architecture to maintain a balance between efficiency and visual quality.

To gain insights into these aspects, we compare SwinIH, its two variants, *i.e.*, SwinIH-MS, SwinIH-Local, and Swin Transformer (Swin-T) [27]. SwinIH is introduced in § 3.4. SwinIH-MS differs from SwinIH in that it first reduces and then enlarges the resolution of feature maps at deeper layers. SwinIH-Local replaces the global attention layer of SwinIH with a Swin attention block. As discussed in § 3.4, Swin-T is composed of Swin attention blocks and uses a Patch Merging layer to reduce the spatial size of feature maps.

Figure 5. Qualitative comparison between our method, LEMaRT (SwinIH), and three SOTA methods (RainNet [26], iS^{2}AM [30], DHT+ [13]) on the iHarmony4 dataset. Compared to other methods, LEMaRT is better at color correction, thanks to the pre-training process during which LEMaRT learns the distribution of photo-realistic images.

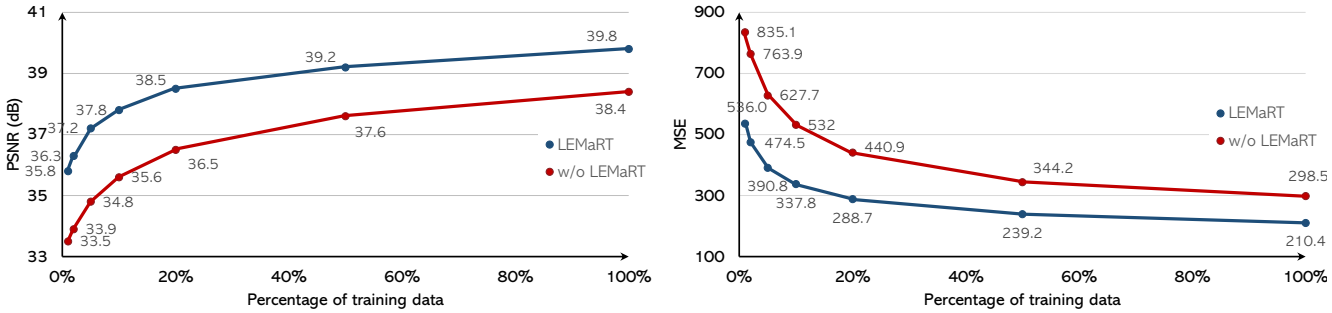

Figure 6. Performance of our SwinIH model when it is fine-tuned on a portion (1–50%) of the iHarmony4 dataset. One variant is trained from scratch using the all training data from the iHarmony4 dataset (referred to as *w/o LEMaRT*). The other is pre-trained with LEMaRT and then fine-tuned using the all training data from the iHarmony4 dataset (referred to as LEMaRT) .

We add a Patch Splitting layer (does the opposite of a Patch Merging layer) to enlarge the size of feature maps to make it suitable for image harmonization.

As shown in Table 4, SwinIH significantly outperforms SwinIH-MS and Swin-T across all four metrics, *e.g.*, by 0.7 dB and 0.9 dB in terms of PSNR, and by 5.8 and 11.5 in terms of MSE, respectively. We hypothesize that the performance drop is caused by the information loss when the resolution of a feature map is reduced. The performance of SwinIH-Local and that of SwinIH are comparable in terms of PSNR and MSE. However, as shown in Figure 7, SwinIH produces results that are of higher visual quality. As shown

Figure 7. Qualitative comparison between the results of SwinIH-Local and SwinIH. SwinIH-Local occasionally generates images with block-shaped visual artifacts, while SwinIH does not.

dataset	metric	model			
		SwinIH	MS	Local	Swin-T
all	PSNR↑	37.0	36.3	37.0	36.1
	MSE↓	35.5	41.3	35.1	47.0
	fPSNR↑	24.5	23.7	24.5	23.2
	fMSE↓	386.6	454.5	385.2	499.3

Table 4. Performance comparison of SwinIH, its two variants, *i.e.*, SwinIH-Local (denoted as Local), SwinIH-MS (denoted as MS), and Swin Transformer (denoted as Swin-T) [27] on iHarmony4.

in Figure 7, SwinIH-Local produces visible block boundaries. This is caused by the shifted window (Swin) attention, which prevents visual tokens at the border of each window to attend to its neighboring visual tokens in adjacent windows. SwinIH is able to remove these block-shaped artifacts. This demonstrates the benefit of using a combination of global and local attention. To maintain high computational and memory efficiency, we only employ it in the last layer of our model.

d. Mask Generation Strategy

dataset	metric	mask generation strategy		
		random	grid	block
all	PSNR↑	39.0	37.1	38.6
	MSE↓	20.9	33.5	23.2
	fPSNR↑	26.6	24.5	26.1
	fMSE↓	250.0	380.3	273.8

Table 5. Comparison of three mask generation strategies introduced in § 3.3: *random*, *grid*, *block*, on iHarmony4.

Here we study the sensitivity of harmonization performance with respect to the mask generation strategy. To this end, we compare the three strategies discussed in § 3.3, *i.e.*, *random*, *grid* and *block*. We measure the performance of our model after being pre-trained on MS COCO and fine-tuned on iHarmony4 for 30 epochs.

As seen in Table 5, the performance of *grid* strategy is worse that of the other two strategies. This result is expected, as the *grid* strategy can only transform image patches at specific locations. Therefore, it is not flexible for cases where there are multiple foreground regions or they cover an area larger than a grid cell. To our surprise, the *random* strategy achieves comparable performance to the *block* strategy, which is designed to mimic test cases. This result confirms that there is no need for a special mask generation algorithm that is tuned for the LEMaRT pre-training scheme. In other words, this simplifies the design and broadens the applicability of LEMaRT to new datasets.

e. Foreground Mask Size

dataset	metric	mask ratio		
		30%	50%	70%
all	PSNR↑	38.8	39.0	39.0
	MSE↓	21.8	20.9	21.2
	fPSNR↑	26.4	26.6	26.5
	fMSE↓	255.0	250.0	250.5

Table 6. Image harmonization metrics corresponding to three different foreground mask ratios on iHarmony4.

We examine the sensitivity of image harmonization results to the foreground mask size. Here, the foreground size is measured by the ratio of the foreground mask size to the image size. In this experiment, we fine-tune the pre-trained models for 30 epochs. In Table 6, we show the quantitative metrics at three different foreground mask ratios, 30%, 50% and 70%. We see that these metrics do not vary significantly between the three ratios. For example, fPSNR varies within a range with 0.2 dB width and fMSE varies within a range whose width is smaller than 5.0. This indicates that the size of generated foreground masks does not have significant impact on performance of pre-trained models.

5. Conclusion

In this work, we introduced Label-Efficient Masked Region Transform (LEMaRT), an effective technique of online data generation for self-supervised pre-training of image harmonization models. LEMaRT provides a simple, yet effective way to leverage large-scale unannotated datasets. In addition, we derived a Swin Transformer-based model that is more efficient than ViT-style Transformer networks for image harmonization. Extensive experiments on the iHarmony4 dataset validate the effectiveness of both our pre-training method and our model. We set a new state of the art for image harmonization, while showing that our pre-training method is much more label-efficient than the existing methods and is consistently applicable to a wide range of network architectures for image harmonization.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

[2] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEit: BERT pre-training of image transformers. In *International Conference on Learning Representations*, 2022.

[3] Junyan Cao, Wenyan Cong, Li Niu, Jianfu Zhang, and Liqing Zhang. Deep image harmonization by bridging the reality gap. In *BMVC*, June 2022.

[4] Bor-Chun Chen and Andrew Kae. Toward realistic image compositing with adversarial learning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8407–8416, 2019.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607, 2020.

[6] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9640–9649, 2021.

[7] Wenyan Cong, Li Niu, Jianfu Zhang, Jing Liang, and Liqing Zhang. Bargainnet: Background-guided domain translation for image harmonization. In *ICME*, pages 1–6, 2021.

[8] Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. Dovenet: Deep image harmonization via domain verification. In *CVPR*, 2020.

[9] 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.

[10] 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 *International Conference on Learning Representations*, 2021.

[11] Debidatta Dwibedi, Ishan Misra, and Martial Hebert. Cut, paste and learn: Surprisingly easy synthesis for instance detection. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[12] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D. Cubuk, Quoc V. Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2918–2928, June 2021.

[13] Zonghui Guo, Zhaorui Gu, Bing Zheng, Junyu Dong, and Haiyong Zheng. Transformer for image harmonization and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[14] Zonghui Guo, Dongsheng Guo, Haiyong Zheng, Zhaorui Gu, Bing Zheng, and Junyu Dong. Image harmonization with transformer. In *ICCV*, pages 14870–14879, October 2021.

[15] 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 (CVPR)*, pages 16367–16376, June 2021.

[16] Yucheng Hang, Bin Xia, Wenming Yang, and Qingmin Liao. Scs-co: Self-consistent style contrastive learning for image harmonization. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19678–19687, 2022.

[17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 15979–15988. IEEE, 2022.

[18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020.

[19] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR, 18–24 Jul 2021.

[20] Yifan Jiang, He Zhang, Jianming Zhang, Yilin Wang, Zhe Lin, Kalyan Sunkavalli, Simon Chen, Sohrab Amirghodsi, Sarah Kong, and Zhangyang Wang. Ssh: A self-supervised framework for image harmonization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4832–4841, 2021.

[21] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, 2020.

[22] Jean-Francois Lalonde and Alexei A. Efros. Using color compatibility for assessing image realism. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8, 2007.

[23] Donghoon Lee, Tomas Pfister, and Ming-Hsuan Yang. Inserting videos into videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

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

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C Lawrence Zitnick. Deep high-resolution representation learning for human pose estimation. In *ECCV*, 2014.

[26] Jun Ling, Han Xue, Li Song, Rong Xie, and Xiao Gu. Region-aware adaptive instance normalization for image harmonization. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition (CVPR), pages 9361–9370, June 2021.

[27] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[28] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam, 2018.

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021.

[30] Konstantin Sofiiuk, Polina Popenova, and Anton Konushin. Foreground-aware semantic representations for image harmonization. In *WACV*, pages 1620–1629, 2021.

[31] Shuangbing Song, Fan Zhong, Xueying Qin, and Changhe Tu. Illumination harmonization with gray mean scale. In *Advances in Computer Graphics: 37th Computer Graphics International Conference, CGI 2020, Geneva, Switzerland, October 20–23, 2020, Proceedings*, page 193–205, Berlin, Heidelberg, 2020. Springer-Verlag.

[32] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019.

[33] Kalyan Sunkavalli, Micah K. Johnson, Wojciech Matusik, and Hanspeter Pfister. Multi-scale image harmonization. *ACM Trans. Graph.*, 29(4), jul 2010.

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

[35] Hao Wang, Qilong Wang, Fan Yang, Weiqi Zhang, and Wangmeng Zuo. Data augmentation for object detection via progressive and selective instance-switching. *CoRR*, abs/1906.00358, 2019.

[36] Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Luowei Zhou, Yucheng Zhao, Yujia Xie, Ce Liu, Yu-Gang Jiang, and Lu Yuan. Omnivl: One foundation model for image-language and video-language tasks. *arXiv preprint arXiv:2209.07526*, 2022.

[37] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[38] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[39] Su Xue, Aseem Agarwala, Julie Dorsey, and Holly Rushmeier. Understanding and improving the realism of image composites. *ACM Trans. Graph.*, 31(4), July 2012.

[40] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

[41] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803, 2022.

[42] Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David W. Jacobs. Deep single portrait image relighting. In *International Conference on Computer Vision (ICCV)*, 2019.

[43] Mingyang Zhou, Luowei Zhou, Shuohang Wang, Yu Cheng, Linjie Li, Zhou Yu, and Jingjing Liu. Uc2: Universal cross-lingual cross-modal vision-and-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4155–4165, 2021.

[44] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2020.

[45] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A. Efros. Learning a discriminative model for the perception of realism in composite images. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.