TSA^2: Temporal Segment Adaptation and Aggregation for Video Harmonization

Zeyu Xiao Yurui Zhu Xueyang Fu Zhiwei Xiong*
University of Science and Technology of China
{zeyuxiao,zyr}@mail.ustc.edu.cn {xyfu,zwxiong}@ustc.edu.cn

Abstract

Video composition merges the foreground and background of different videos, presenting challenges due to variations in capture conditions (e.g., saturation, brightness, and contrast). Video harmonization is a vital process in achieving a realistic composite by seamlessly adjusting the foreground's appearance to match the background. In this paper, we propose TSA^2, a novel method for video harmonization that incorporates temporal segment adaptation and aggregation. TSA^2 divides the inharmonious input sequence into temporal segments, each corresponding to a different frame rate, allowing effective utilization of complementary information within each segment. The method includes the Temporal Segment Adaptation module, which learns and remaps the distribution difference between background and foreground regions, and the Temporal Segment Aggregation module, which emphasizes and aggregates cross-segment information through element-wise correlations. Experimental results demonstrate that TSA^2 outperforms advanced image and video harmonization methods quantitatively and qualitatively.

1. Introduction

The generation of realistic composite videos has gained significant attention in both academia and industry, offering a wide range of applications in modern society, particularly in the field of multimedia and AIGC [5]. For instance, in online conferences, there is a demand for changing the background of self-portraits [26, 29, 33]. In general, video composition involves merging the foreground of one video with the background of another [1, 28]. In practice, however, due to variations in capture conditions, such as saturation, brightness, and contrast, composite videos often appear unrealistic. To overcome this challenge, image and video harmonization techniques have been proposed to generate more realistic composite videos with minimal human intervention, eliminating the need for specialized editing software like Adobe Photoshop and Adobe Premiere Pro.

To generate harmonious video frames, one approach is to apply image harmonization methods on a frame-by-frame basis. Early techniques focused on color and tone matching, involving the transfer of global statistics [32, 34], gradient-domain methods [31], and multi-scale statistics matching [38]. More recently, convolutional neural networks (CNNs) and vision Transformers have been utilized [8, 10, 14–16, 18, 36, 40]. These methods have achieved notable progress by learning dense pixel-to-pixel transformations between composite images and ground-truth harmonized images. However, image harmonization methods typically neglect the temporal relationship between frames, leading to potential inconsistencies and flickering artifacts in consecutive frames. As illustrated in Figure 1, even advanced image harmonization methods like HarmonyTransformer [15] exhibit unsatisfactory results, with noticeable temporal discontinuity in the inharmonious car area.

To address the temporal consistency in video harmonization, Huang *et al.* [19] propose an end-to-end network trained on a synthetic dataset. However, its limited temporal receptive field (2 frames) and inadequate consideration of internal correlations and distributions within the foreground and background regions lead to unsatisfactory visual results, as shown in Figure 1. Recently, Lu *et al.* [27] propose CO$_2$Net, a video harmonization method based on the assumption of color mapping consistency. CO$_2$Net consists of an image harmony network for generating harmonized images and a refinement module that enhances the results using color mapping consistency from a lookup table. However, the video harmonization performance of CO$_2$Net is influenced by its first stage, and the two-stage training scheme introduces complexity compared to training from scratch. Additionally, the lookup table focuses on pixel-level color consistency, neglecting global color information in the foreground and background regions.

In this paper, we propose TSA^2, a novel video harmonization method that addresses the aforementioned issues by leveraging long-term temporal information and exploiting background information across frames. Taking inspiration from notable works such as TSN [44] and Slowfast Network [12], which adeptly sample temporal information for action recognition and segment videos into various parts

[*] Corresponding author.

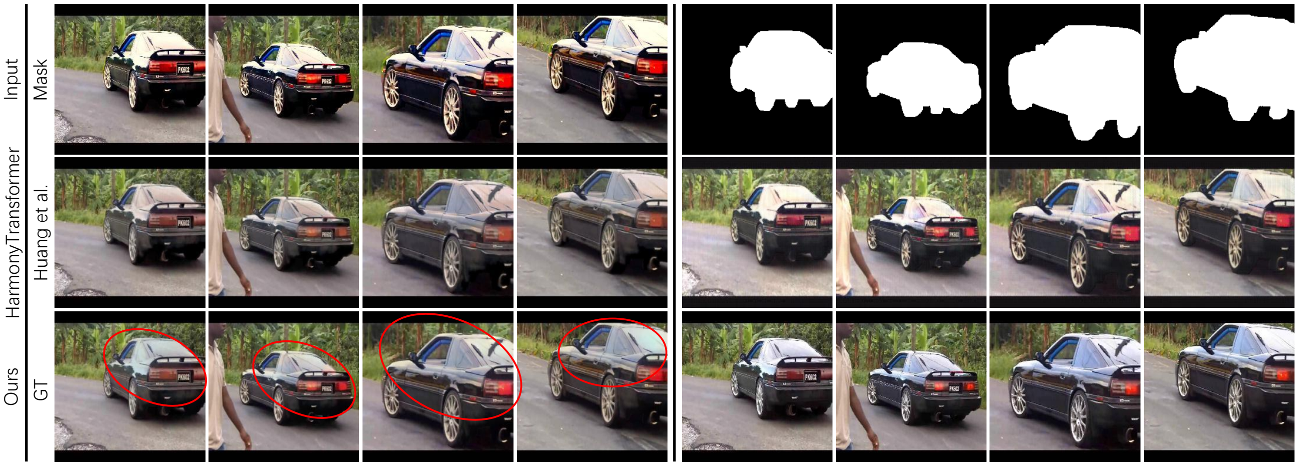


Figure 1. Examples of video harmonization. Our proposed TSA^2 can adjust the appearances of the composite foreground video sequence to make them compatible with the background regions. We show the harmonized results generated by the advanced image harmonization method HarmonyTransformer [15] and the video harmonization method Huang *et al.* [19]. Best viewed on screen.

with distinct frame rates for video recognition, we adopt a similar strategy. This involves partitioning the original inharmonious input sequence into multiple temporal segments, each encapsulating a diverse frame rate representation. By dividing the sequence into temporal segments, TSA^2 surpasses methods that rely on two confined adjacent frames or a single image harmonization technique, as it can effectively employ temporal information within each segment and capture complementary temporal information across segments. We introduce two key modules in TSA^2: the Temporal Segment Adaptation (TSAda) module and the Temporal Segment Aggregation (TSAgg) module.

The TSAda module integrates inharmonious foreground and harmonious background within each temporal segment. It learns the distribution difference between foreground and background regions and remaps the distribution of foreground features to match that of background features. The TSAgg module further aggregates information across temporal segments. It emphasizes and aggregates the complementary information by utilizing element-wise correlations calculated through a segment attention mechanism. This adaptive aggregation process enhances the harmonization results. Moreover, TSA^2 can be trained from scratch, avoiding the need for a complex two-stage training scheme used by CO$_2$Net. This simplifies the training process while achieving effective video harmonization. As shown in Figure 1, with the two elaborate modules, our TSA^2 achieves improved visual quality compared with existing advanced image and video harmonization methods.

Our main contributions are summarized as follows. (1) We introduce TSA^2, a novel video harmonization method that effectively addresses the challenges of inharmonious video sequences. By dividing the input into frame-rate-aware temporal segments, TSA^2 leverages long-term temporal information for harmonization. (2) We propose the TSAda module, which facilitates the mapping of foreground features to match the distribution of background features within each temporal segment. (3) We propose the TSAgg module, which can emphasize and aggregate information across different temporal segments. (4) Comparative evaluations demonstrate the superiority of our TSA^2 over existing image and video harmonization techniques.

2. Related Work

Image harmonization. Early research in image harmonization focused on utilizing low-level image representations to adjust the appearance of the foreground to match the background. These methods employed techniques such as alpha matting [37, 42], gradient [20, 31], color distribution [7, 32, 34] and multi-scale statistics [38]. However, these approaches often prioritize matching the appearance without giving enough consideration to visual realism [29]. Recent advancements in image harmonization have leveraged CNNs and vision Transformers to achieve promising results [2–4, 6, 8–10, 13–18, 21–25, 35, 36, 40, 41, 43, 48, 50]. For instance, Tsai *et al.* [40] introduce an end-to-end deep network that utilizes a segmentation mask as semantic information for training. Cong *et al.* [8] formulate image harmonization as background-guided domain translation, using a background domain code to guide the harmonization process. However, since the image harmonization methods do not consider the temporal relationship between frames, consecutive frames are not connected naturally, resulting in sub-optimal results when they are applied directly to the video harmonization task.

Video harmonization. Video harmonization can be seen as an extension of image harmonization, where the goal is to adjust the appearance of the foreground video to match the background video. Huang *et al.* [19] introduce a dataset of synthetic composite videos generated from real images and proposed an end-to-end network for video harmonization. Their method achieves realistic results by incorporat-

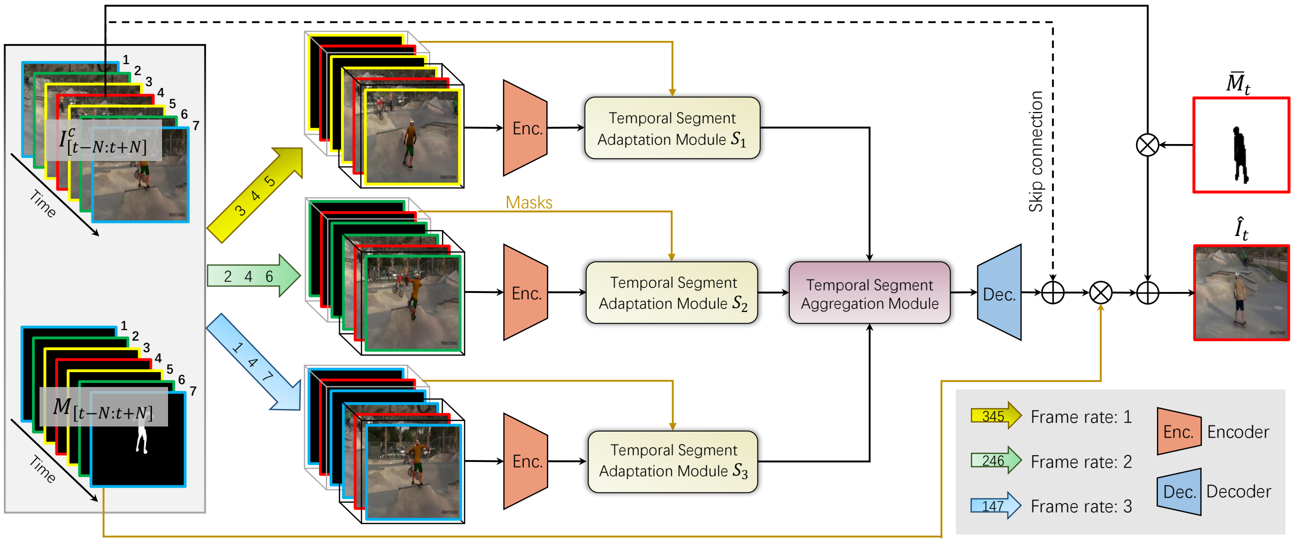


Figure 2. The framework of the proposed TSA^2, which consists of three feature encoders, three TSAda modules, a TSAgg module, and a feature decoder. Here we take an inharmonious input sequence with seven frames as an example.

ing a pixel-wise disharmony discriminator and considering the temporal consistency between adjacent frames through a temporal consistency loss. However, their network has limitations in terms of capturing long-term temporal contexts and utilizing the internal correlations and distributions of foreground and background information. More recently, Lu *et al.* [27] propose CO$_2$Net for video harmonization. CO$_2$Net consists of an image harmonization network and a refinement module. However, the performance of CO$_2$Net is dependent on the image harmonization network, and the utilization of local pixel-level information through a lookup table may limit its capabilities. In this paper, we introduce TSA^2, a method that utilizes temporal segment adaptation and aggregation to tackle video harmonization. Our TSA^2 effectively leverages internal correlations and distributions within each temporal segment and across different segments. Moreover, TSA^2 can be trained from scratch, eliminating the need for a complex two-stage training scheme as in [27].

3. Method

3.1. Overview

Given $2N{+}1$ consecutive composite inharmonious frame sequence $I^C_{[t-N:t+N]} \in \mathbb{R}^{(2N+1)\times 3\times H\times W}$, and the corresponding binary mask sequence of composite foreground regions $M_{[t-N:t+N]} \in \mathbb{R}^{(2N+1)\times 1\times H\times W}$, we denote I^C_t as the central frame and the other frames $I^C_{[t-N:t-1,t+1:t+N]}$ as neighboring frames, and M_t denotes the central frame foreground mask. H and W denote the height and the weight, respectively. Accordingly, the background mask sequence can be denoted as $\bar{M}_{[t-N:t+N]} = 1 - M_{[t-N:t+N]}$. The goal of our method is to generate a realistic central frame $\hat{I}_t$ which is consistent with the background region $I^B_t = \bar{M}_t \circ I^C_t$ visually and temporally, and should be close to the ground-truth frame I^{GT}_t. $\bar{M}_t$ denotes the central frame background mask. The overall pipeline of our proposed method is shown in Figure 2.

Without loss of generality, we take the composite inharmonious frame sequence with 7 frames $\{I^C_1, I^C_2, \ldots, I^C_7\}$ as an example. As shown in Figure 2, the seven input frames are first divided into three temporal segments based on different temporal intervals (*i.e.*, $\{I^C_4, I^C_4, I^C_5\}$, $\{I^C_2, I^C_4, I^C_6\}$ and $\{I^C_1, I^C_4, I^C_7\}$), with each temporal segment representing a frame rate (from one to three, respectively) and containing various temporal information for video harmonization. To process these temporal segments, we employ a shared-weight feature encoder to convert the frames into the feature domain. The TSAda module is then applied to transfer important background information to the inharmonious foreground region in the central frame I^C_4 within each segment. Information across three temporal segments is further emphasized and aggregated through the TSAgg module. The aggregated features are then fed into the feature decoder to generate the harmonized central frame $\hat{I}_4$. The structures of the feature encoder and the feature decoder are shown in the supplementary document.

3.2. Dividing Sequence into Temporal Segments

Utilizing long-term temporal information from distant harmonious background regions is crucial for video harmonization [19]. However, the network proposed in [19] falls short in this aspect. It stacks two adjacent inharmonious frames along the channel dimension and directly applies several multi-scale 2D convolution layers to these stacked frames. The use of only two frames limits the temporal context and hinders the inharmonious regions from accessing harmonious background information from distant frames.

In this paper, we draw inspiration from TSN [44] and Slowfast Network [12] to address the video harmonization problem. We divide the neighboring $2N$ frames in an inhar-

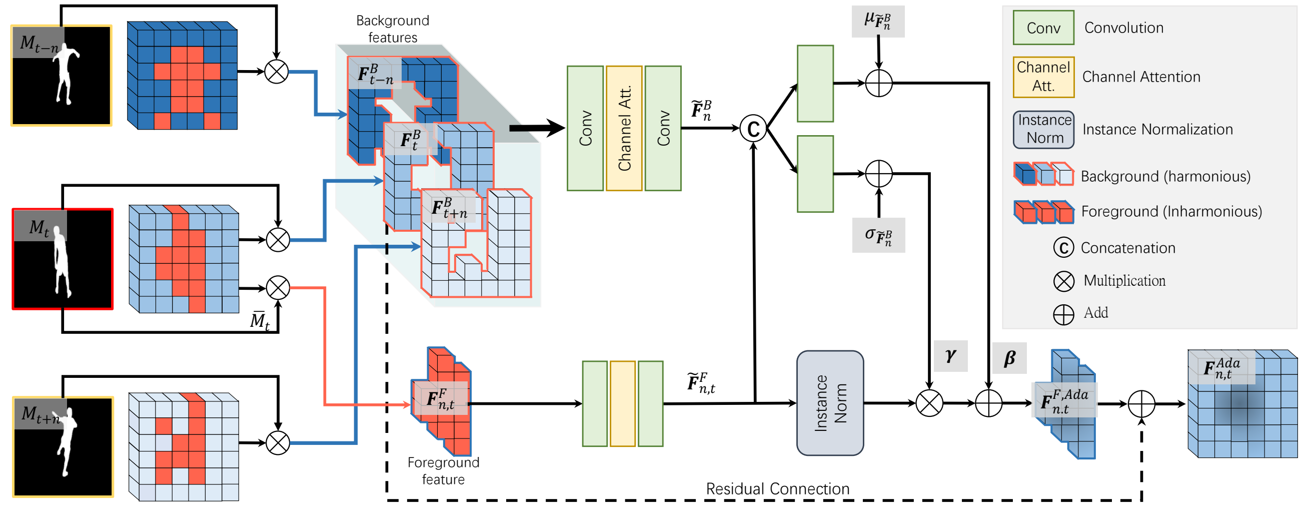


Figure 3. Structure of the TSAda module. This module is utilized to remap the distribution of extracted foreground features to that of the background features.

monious sequence into N temporal segments based on their temporal distances from the central frame, with each segment representing a specific frame rate. The segmentation is predicated upon their temporal proximity to the central frame, effectively encapsulating various frame rates within each segment. This segmentation equips us to harness temporal cues more effectively. By recognizing the differing significance of individual frames, and by orchestrating the harmonious collaboration of insights across segments, we optimize the utilization of temporal data. Expanding upon this foundational segmentation, we introduce the TSAda module and the TSAgg module.

3.3. Temporal Segment Adaptation

Due to variations in capture conditions such as season, weather, and time of the day, there is a significant divergence in brightness and color distributions between inharmonious foreground regions and harmonious background regions [27]. Simply concatenating foreground and background features within each temporal segment and passing them through subsequent convolution layers is insufficient, leading to suboptimal utilization of foreground and background information. To address this issue, we propose the TSAda module inspired by [30] and [24], which aims to remap the distribution of extracted foreground features to that of the background features. By doing so, we enable more effective utilization and transfer of important information and color distribution from the background features. The structure of the TSAda module is depicted in Figure 3.

The input video sequence is divided into N temporal segments $\{S_1, \dots, S_n\}, n \in [1 : N]$, where $S_n = \{I^C_{t-n}, I^C_t, I^C_{t+n}\}$ is a temporal segment consisting of a former neighboring frame I^C_{t-n}, the central frame I^C_t and a latter neighboring frame I^C_{t+n}. Since we need to harmonize the central frame, the reference frame I^C_t appears in each temporal segment. After converting the input segment S_n into the feature domain $\boldsymbol{F}_n = \left\{\boldsymbol{F}^C_{t-n}, \boldsymbol{F}^C_t, \boldsymbol{F}^C_{t+n}\right\}$, we multiply the input features by the background masks $\bar{M}_n = \left\{\bar{M}_{t-n}, \bar{M}_t, \bar{M}_{t+n}\right\}$ and multiply the central feature by its foreground mask M_t, getting the background features $\boldsymbol{F}^B_n = \left\{\boldsymbol{F}^B_{t-n}, \boldsymbol{F}^B_t, \boldsymbol{F}^B_{t+n}\right\}$ and the central foreground feature $\boldsymbol{F}^F_{n,t}$. Then, $\boldsymbol{F}^B_n$ and $\boldsymbol{F}^F_{n,t}$ are then fed into two convolution layers with a channel attention operation, yielding $\widetilde{\boldsymbol{F}}^B_n$ and $\widetilde{\boldsymbol{F}}^F_t$, respectively. $\widetilde{\boldsymbol{F}}^B_n$ and $\widetilde{\boldsymbol{F}}^F_t$ are with equal channel numbers. Afterward, $\widetilde{\boldsymbol{F}}^B_n$ and $\widetilde{\boldsymbol{F}}^F_{n,t}$ are first concatenated before feeding into convolution layers to produce two parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, which are with the same size as $\widetilde{\boldsymbol{F}}^F_t$. Then instance normalization is applied to the central foreground feature $\widetilde{\boldsymbol{F}}^F_{n,t}$ as

$$\widetilde{\boldsymbol{F}}^{F,c}_{n,t} \leftarrow (\widetilde{\boldsymbol{F}}^{F,c}_{n,t} - \boldsymbol{\mu}^{F,c}_{n,t})/\boldsymbol{\sigma}^{F,c}_{n,t}, \tag{1}$$

where $\boldsymbol{\mu}^{F,c}_{n,t}$ and $\boldsymbol{\sigma}^{F,c}_{n,t}$ are the mean and standard deviation of $\widetilde{\boldsymbol{F}}^{F,c}_{n,t}$ in channel c as

$$\boldsymbol{\mu}^{F,c}_{n,t} = \frac{1}{HW}\sum_{y,x}\widetilde{\boldsymbol{F}}^{F,c,y,x}_{n,t} , \tag{2}$$

$$\boldsymbol{\sigma}^{F,c}_{n,t} = \sqrt{\frac{1}{HW}\sum_{y,x}\left(\widetilde{\boldsymbol{F}}^{F,c,y,x}_{n,t} - \boldsymbol{\mu}^{F,c}_{n,t}\right)^2} , \tag{3}$$

where H and W are the height and width of $\widetilde{\boldsymbol{F}}^{F,c}_{n,t}$. Then we update $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ as

$$\boldsymbol{\beta} \leftarrow \boldsymbol{\beta} + \boldsymbol{\mu}_{\widetilde{\boldsymbol{F}}^B_n} , \boldsymbol{\gamma} \leftarrow \boldsymbol{\gamma} + \boldsymbol{\sigma}_{\widetilde{\boldsymbol{F}}^B_n} , \tag{4}$$

where $\boldsymbol{\mu}_{\widetilde{\boldsymbol{F}}^B_n}$ and $\boldsymbol{\sigma}_{\widetilde{\boldsymbol{F}}^B_n}$ are calculated in a similar way as Equations (2) and (3). Then, γ and $\boldsymbol{\beta}$ are multiplied and added to the normalized foreground feature in an element-wise manner as

$$\boldsymbol{F}^{F,Ada}_{n,t} \leftarrow \widetilde{\boldsymbol{F}}^F_{n,t} \cdot \boldsymbol{\gamma} + \boldsymbol{\beta} . \tag{5}$$

Finally, we add $\boldsymbol{F}^{F,Ada}_{n,t}$ with the extracted harmonious region of the central frame $\boldsymbol{F}^{B}_{t}$ as the output

$$\boldsymbol{F}^{Ada}_{n,t} \leftarrow \boldsymbol{F}^{F,Ada}_{n,t} + \boldsymbol{F}^{B}_{t}\ . \tag{6}$$

Since the difference between the foreground and the background features varies with respect to the different spatial location, while the statistics $\boldsymbol{\mu}_{\widetilde{\boldsymbol{F}}^{B}_{n}}$, $\boldsymbol{\sigma}_{\widetilde{\boldsymbol{F}}^{B}_{n}}$, $\boldsymbol{\mu}_{\widetilde{\boldsymbol{F}}^{F,c}_{n,t}}$ and $\boldsymbol{\sigma}_{\widetilde{\boldsymbol{F}}^{F,c}_{n,t}}$ are of size $C \times 1 \times 1$, we use convolution layers to predict two spatial-wise adaptation parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$.

In contrast to [30], which employs segmentation maps for parameter generation, and unlike [24], where foreground and background regions are treated separately, the convolutional layers within the TSAda module jointly process foreground and background features to capture their distinctions. Subsequent to acquiring $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ from the convolutional layers, we seamlessly integrate them into the background features' mean and standard deviation. This synergistic fusion enables us to harness significant background information during the remapping process, consequently amplifying the effective utilization and seamless transference of pertinent features. This harmonious integration considerably enhances the feature remapping mechanism, thereby elevating its efficiency and efficacy.

3.4. Temporal Segment Aggregation

To address the varying levels of informativeness in different temporal segments and effectively integrate features from multiple segments, we introduce the TSAgg module. Unlike existing methods that use 2D or 3D convolution layers for multi-frame aggregation, our TSAgg module adopts a novel approach. It leverages multi-segment attention maps to assign adaptive pixel-level aggregation weights to each temporal segment. These attention maps, calculated in an embedding space, represent the similarity between different adapted segments. By incorporating the attention maps, the TSAgg module can effectively aggregate information across different temporal segments, taking into account both the informative low-frame-rate segments and the fine-detail information captured by high-frame-rate segments. This adaptable aggregation mechanism substantially bolsters the holistic video harmonization procedure. This enhancement empowers the model to efficiently capitalize on the available temporal cues, ultimately leading to enhanced performance. Detailed struture of the TSAgg module can be found in the supplementary document.

For each adapted temporal segment output from the TSAda module (*i.e.*, $\boldsymbol{F}^{Ada}_{n,t} = \left\{\boldsymbol{F}^{Ada}_{1,t}, \boldsymbol{F}^{Ada}_{2,t}, \boldsymbol{F}^{Ada}_{3,t}\right\}$), we convert $\boldsymbol{F}^{Ada}_{n,t}, n = [1:3]$ into the embedding space. In our implementation, we use a 3×3 convolution layer for the converting operation. Intuitively, regions that are more similar to the adjusted central frame should be paid more attention in the embedding space. They are further concatenated, followed by a *sigmoid* operation applied to each position across channels to calculate the multi-segment attention maps M_n as

$$M_n = \textit{sigmoid}(\theta(\boldsymbol{F}^{Ada}_{n,t})^T \phi(\boldsymbol{F}^{Ada}_{2,t})), n = 1, 2, 3, \tag{7}$$

where $\theta(\cdot)$ and $\phi(\cdot)$ are convolution layers. The multi-segment attention maps are then multiplied in a pixel-wise adaptive manner to $\boldsymbol{F}^{Ada}_{n,t}$, followed by a fusion convolution layer to aggregate these segment-attention-modulated features as

$$\boldsymbol{F}^{Agg}_{t} = \mathrm{Conv}([M_1 \otimes \boldsymbol{F}^{Ada}_{1,t}, M_2 \otimes \boldsymbol{F}^{Ada}_{2,t}, M_3 \otimes \boldsymbol{F}^{Ada}_{3,t}]). \tag{8}$$

In order to fully utilize temporal information over different temporal segments, we further concatenate them along the temporal dimension and feed it into a 3D convolution layer. The output from the 3D convolution layer is added with $\boldsymbol{F}^{Agg}_{t}$ to generate $\widetilde{\boldsymbol{F}}^{Agg}_{t}$ as the final output of the TSAgg module. In this way, important information can be emphasized and aggregated across different segments.

4. Experiments

4.1. Experimental Settings

Training settings. Huang *et al.* [19] collect a private synthetic dataset named Dancing MSCOCO on their own, which is not suitable for a fair comparison with other methods, and there is a massive gap between the simulated movement and the complex movement in real videos [27]. In this paper, we use HYouTube as the training set [27], the same as CO$_2$Net. HYouTube is a recent proposed large dataset for the video harmonization task based on the existing large-scale video object segmentation dataset YouTubeVOS 2018 [49], which contains more than 25,000 20-frame video sequences with various motions and diverse scenes [27]. Given the ground-truth frame I^{GT}_{t} and the harmonized result $\hat{I}_t$ generated by our method TSA^2, we adopt the simple but effective Charbonnier loss [45–47] to train our method from scratch, which is different from the two-stage training scheme proposed in CO$_2$Net. The training loss function is

$$\mathcal{L} = \sqrt{\left\|I^{GT}_{t} - \hat{I}_t\right\|^2 + \varepsilon^2}, \tag{9}$$

where ε is set to $1e{-}6$ in our experiments. All video frames are resized to 256×256 pixels for training and testing, which is the same setting as CO$_2$Net [27]. Rotation and flipping are applied for data augmentation.

Inference settings. We evaluate our method on HYouTube-Test using the sliding-window scheme [45]. HYouTube-Test includes 636 20-frame video sequences. To quantitatively evaluate the generated frames, we utilize MSE, PSNR, SSIM, foreground MSE (fMSE), foreground PSNR (fPSNR) and foreground SSIM (fSSIM) as metrics [15, 16]. We utilize NIQE to evaluate the visual quality and TL to evalutate the temporal consistency. We employ RAFT [39] to compute optical flow for TL.

Implementation details. TSA^2 takes seven consecutive frames (*i.e.*, $N = 3$) as inputs unless otherwise specified.

Table 1. Quantitative comparisons of different methods on HYouTube-Test [27]. The best results are marked in **bold** while the second ones are marked with underlines. The unit of #Params is million (M). The Runtime is the average running time (seconds) which is measured on HYouTube with the spatial resolution of 256×256 in a per-frame manner.

Method	Parameters (M)	Runtime (s)	PSNR↑	SSIM↑	fPSNR↑	fSSIM↑	NIQE↓	MSE↓	fMSE↓	TL↓
DIH [40]	24.79	0.0066	29.95	0.6292	24.17	0.6933	7.0711	105.70	614.07	11.75
S^2AM [11]	67.00	0.1636	30.05	0.6695	24.26	0.7657	5.4439	103.93	611.22	9.64
DoveNet [10]	54.76	0.0113	30.19	0.6856	24.35	0.7677	5.8637	113.15	444.92	9.13
FeatureModulation [18]	11.59	0.1512	30.26	0.6960	24.64	0.7876	5.7941	82.04	474.13	8.64
RainNet [24]	54.75	0.0162	30.75	0.8006	24.36	0.7681	5.6167	56.90	298.85	6.24
IntrinsicHarmony [16]	40.86	0.0440	30.51	0.7984	23.88	0.7497	5.7580	67.27	359.81	5.89
BargainNet [8]	3.92	0.0113	30.21	0.7955	24.38	0.7757	5.7004	97.52	503.77	5.57
HaronyTransformer [15]	26.52	0.0783	31.14	0.8047	25.07	0.7933	5.6948	35.58	208.26	4.95
S^2CRNet-VGG16 [22]	21.70	0.0136	31.20	0.8055	25.12	0.7935	5.6689	35.51	205.96	4.90
EDVR [45]	20.50	0.3210	31.33	0.8048	25.51	0.7875	5.3632	35.79	198.79	**4.49**
Huang *et al.* [19]	55.27	0.1419	30.37	0.7342	24.29	0.7339	5.7749	76.48	379.99	4.60
CO_2Net [27]	26.77	0.0460	31.30	0.8050	25.53	0.7978	5.2904	35.26	196.50	4.64
Ours	16.30	0.0315	**31.57**	**0.8063**	**26.07**	**0.8044**	**5.1809**	**34.70**	**192.32**	4.55

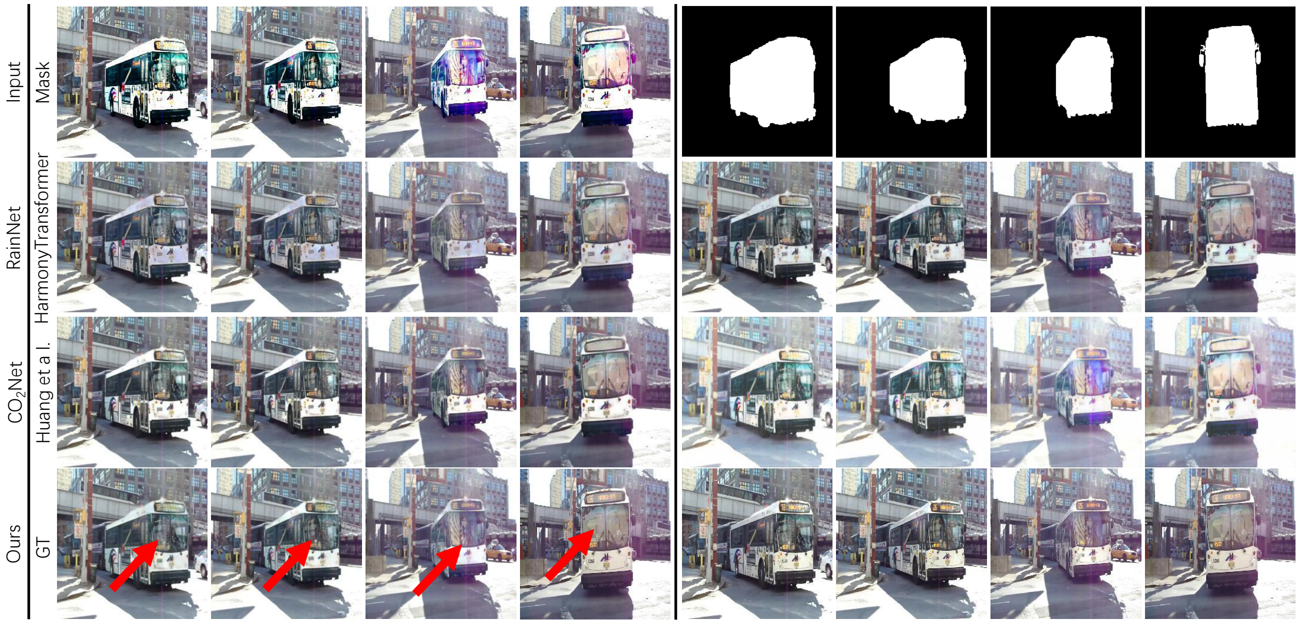


Figure 4. Visual comparisons of different representative methods on video frames from the HYouTube dataset. Red arrows indicate where our TSA^2 better harmonizes foreground regions and produces realistic appearance adjustments. Please zoom in for better visualization.

We utilize the Adam optimizer with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$ to train our method. The learning rate is initially set to $1e-4$ and is later reduced by 0.5 every 150k iterations till 600k iterations. We implement our method with the PyTorch framework and train the method using 2 NVIDIA GeForce GTX1080Ti GPUs, with each mini-batch consisting of 2 samples.

4.2. Quantitative and Qualitative Comparisons

We compare the proposed TSA^2 with two kinds of advanced methods: (1) image harmonization methods include DIH [40], S^2AM [11], DoveNet [10], FeatureModulation [18], RainNet [24], IntrinsicHarmony [15], BargainNet [8], HarmonyTransformer [15] and S^2CRNet-VGG16 [22]. We retrain all these methods using the dataset of HYouTube based on the publicly available codes released by the authors. (2) Video harmonization methods Huang *et al.* [19] and CO_2Net. We carefully implement Huang *et al.* and train the network on HYouTube. Since the CO_2Net is a framework with an image harmonization network and a refinement module, we put the refinement module behind the state-of-the-art image harmonization method HarmonyTransformer. (3) A typical video restoration method: EDVR [45]. Given that EDVR is initially designed for video super-resolution, we have omitted the final upsampling operation to adapt it for video harmonization.

Quantitative evaluations. As shown in Table 1, our TSA^2 outperforms image and video harmonization methods. For example, one can see that TSA^2 outperforms HarmonyTransformer [15], the state-of-the-art image harmonization method, by 0.43dB/0.0016/1.00dB/0.0111 on HYouTube in terms of PSNR/SSIM/fPSNR/fSSIM. Compared to Huang *et al.* [19], TSA^2 obtains 1.20dB/0.0721/1.78/0.0645 gain. In terms of PSNR/SSIM/fPSNR/fSSIM. Though our TSA^2 does not optimize for the NIQE score, it still produces

Table 2. The average rank result of our user study.

Method	Result
TSA^2 (Ours)	**1.570**
CO$_2$Net [27]	1.795
Huang *et al.* [19]	2.303

Table 3. Investigation of the TSAda and TSAgg modules.

Method	TSAda	TSAgg	fPSNR	fSSIM
Ours-Concat	✗	✗	25.21	0.7951
Ours-Baseline	✗	✗	25.67	0.8021
Ours-TSAda	✓	✗	25.85	0.8029
Ours-TSAgg	✗	✓	25.91	0.8032
Ours	✓	✓	26.07	0.8044

the lowest NIQE score, which indicates that the TSA^2 can generate harmonized results with the best visual quality.

Qualitative evaluations. Exemplar visual results of different representative methods are shown in Figure 4. It can be easily observed that our method integrates the foreground videos into the background videos, achieving better visual consistency than other methods. We mask a video for a dynamic comparison of the results obtained from various methods. Please refer to the supplementary material.

Computational efficiency. As shown in Table 1, our method achieves the highest metrics with a relatively small number of parameters and a fast runtime. For example, we achieve a 1.00dB performance gain in terms of fPSNR using only about 60% of the parameters compared to the state-of-the-art image harmonization method, HarmonyTransformer [15]. Compared with the existing video harmonization method, our runtime is about 20% of Huang *et al.* [19], but obtains a 1.78dB fPSNR improvement.

User study. We conduct a user study to subjectively compare our method against the advanced image and video harmonization methods: HarmonyTransformer [15] and Huang *et al.* [19]. 20 participants are asked to rank the results produced by comparison methods and our method for 20 randomly selected videos. For each video, the input and results from the three different methods are shown to the participants, and the participants are asked to rank the results from 1 to 3 (the lower, the better). We allow the participants to give a tie; the average rank is shown below. The user study results are summarized in Table 2. While there are different preferences across videos, it shows that our method is preferred more often by the participants.

4.3. Ablation Studies

We carry out experiments on HYouTube, employing fPSNR and fSSIM as metrics to analyze TSA^2.

Effectiveness of the TSAda and the TSAgg modules. We conduct experiments to demonstrate the contributions of two core modules in our method. We design two baselines in this part: (1) Ours-Concat: we concatenate seven inharmonious frames ($\{I^C_1, I^C_2, \ldots, I^C_7\}$) in the channel dimension at the beginning, and feed them into the same UNet structure as Huang *et al.* [19]. We set the input channel number of the first convolution layer to 21. (2) Our-Baseline: we remove the TSAda module and the TSAgg module and use several residual blocks to replace the TSAda and TSAgg modules while keeping the parameters constant. As shown in Table 3, the TSAda module (Ours-TSAda) and the TSAgg module (Ours-TSAgg) provide about 0.64 dB and 0.70 dB fPSNR gains compared with Ours-Concat, and provide about 0.18 dB and 0.24 dB fPSNR gains compared with Ours-Baseline. Using the TSAda module and the TSAgg module simultaneously provides 0.86 dB and 0.40 dB gains in terms of fPSNR compared with Ours-Concat and Ours-Baseline, respectively.

Table 4. Results of our proposed method with different number of input frames (*i.e.*, N).

Method	fPSNR	fSSIM
Ours-#Frame3 ($N = 1$)	25.81	0.8030
Ours-#Frame5 ($N = 2$)	25.98	0.8039
Ours-#Frame7 ($N = 3$)	26.07	0.8044
Ours-#Frame9 ($N = 4$)	26.14	0.8051
Ours-#Frame11 ($N = 5$)	26.04	0.8040

Table 5. Investigation of different temporal segment branches.

345	246	147	fPSNR	fSSIM
✓	✗	✗	25.81	0.8030
✗	✓	✓	25.99	0.8040
✓	✗	✓	25.98	0.8039
✓	✓	✗	25.98	0.8038
✓	✓	✓	26.07	0.8044

Table 6. Results achieved by the TSAda module and its variants.

Method	fPSNR	fSSIM
TSAda-w/o-ChannelAtt.	25.99	0.8039
TSAda-w/o-Concat	26.02	0.8042
TSAda-w/o-$\mu_{\widetilde{\boldsymbol{F}}^B_n}$&$\sigma_{\widetilde{\boldsymbol{F}}^B_n}$	26.00	0.8040
TSAda-RegionNorm.	25.96	0.8036
TSAda	26.07	0.8044

Different number of input frames. We evaluate how the number of input frames affects the harmonized performance. We enumerate the cases where $N = 1$ to 5, *i.e.*, the input frames number changes from 3 to 11. Table 4 shows that the capability of TSA^2 rises at first and drops with the increasing number of input frames. Compared with taking only three frames ($N = 1$) as the input, the nine-frame setting obtains more than 0.3dB gain on fPSNR. It demonstrates that increasing input frames leads to improvements. However, once saturated, it is unnecessary to include more input frames since more weakly correlated frames from a sequence bring unwanted noise. Considering the computational costs, we choose 7 frames as input.

Different number of temporal segment branches. We present a comprehensive analysis of the influence of varying the number of branches on our proposed approach, as illustrated in Table 5. The observed results underscore a significant trend: activating all branches leads to the attainment of notably elevated fPSNR and fSSIM values, thereby

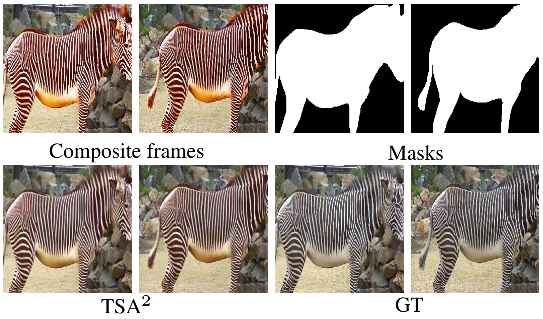
Table 7. Results achieved by the TSAgg module and its variants.

Method	fPSNR	fSSIM
TSAgg-3DConv	25.96	0.8034
TSAgg-2DConv	25.94	0.8034
TSAgg-w/o-3DConv	26.02	0.8042
TSAgg-w/o-Attention	25.89	0.8031
TSAgg	26.07	0.8044

manifesting a pronounced enhancement in overall performance. The findings substantiate the virtue of harnessing supplementary temporal information, which distinctly contributes to an augmented efficacy in the context of video harmonization. These results resonates consistently with our earlier analysis and underscores the potency of incorporating richer temporal cues to bolster video harmonization.

Comparison of the TSAda module with several variants. We further validate the effectiveness of the TSAda Module by introducing the following four variants: **(1)** TSAda-w/o-ChannelAtt.: we remove the channel attention operation in the TSAda module. **(2)** TSAda-w/o-Concat: instead of concatenating $\widetilde{\boldsymbol{F}}^{B}_{n}$ and $\widetilde{\boldsymbol{F}}^{F}_{,tn}$, we feed $\widetilde{\boldsymbol{F}}^{B}_{n}$ to the convolution layers and generate $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ directly. **(3)** TSAda-w/o-$\mu_{\widetilde{\boldsymbol{F}}^{B}_{n}}\&\sigma_{\widetilde{\boldsymbol{F}}^{B}_{n}}$: we donnot update $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ as Equations (4). **(4)** TSAda-RegionNorm.: we embed the Region-aware Adaptive Instance Normalization (RAIN) [24] into the TSAda module. As shown in Table 6, after we remove the channel attention operation, the fPSNR value drops by 0.08dB, indicating this operation can explicitly emphasize the essential background region features while suppressing redundant channels and enhancing the representation ability of the foreground region features. TSAda-w/o-Concat and TSAda-w/o-$\mu_{\widetilde{\boldsymbol{F}}^{B}_{n}}\&\sigma_{\widetilde{\boldsymbol{F}}^{B}_{n}}$ have 0.05dB and 0.07 dB drop in terms of fPSNR, which demonstrate the effectiveness of the TSAda module is to utilize and remap the distribution of foreground features to that of background features adaptively. We also compare the TSAda module with the RAIN, achieving a 0.11dB improvement in terms of fPSNR.

Comparison of the TSAgg module with several variants. We further validate the effectiveness of TSAgg Module by introducing the following four variants: **(1)** TSAgg-Conv3D: we concatenate $\left\{\boldsymbol{F}^{Ada}_{1,t}, \boldsymbol{F}^{Ada}_{2,t}, \boldsymbol{F}^{Ada}_{3,t}\right\}$ along the temporal dimension first and then feed the concatenated feature to several 3D convolution layers. **(2)** TSAgg-Conv2D: we concatenate $\left\{\boldsymbol{F}^{Ada}_{1,t}, \boldsymbol{F}^{Ada}_{2,t}, \boldsymbol{F}^{Ada}_{3,t}\right\}$ along the channel dimension first and then feed the concatenated feature to several 2D convolution layers. **(3)** TSAgg-w/o-Conv3D: we remove the branch of Conv 3D in the TSAgg module. **(4)** TSAgg-w/o-Attention: we do not calculate the multi-segment attention maps, but fuse $\left\{\boldsymbol{F}^{Ada}_{1,t}, \boldsymbol{F}^{Ada}_{2,t}, \boldsymbol{F}^{Ada}_{3,t}\right\}$ directly. As shown in Table 7, neither TSAgg-3DConv nor TSAgg-2DConv achieves good results because it is insufficient to directly aggregate cross-segment information with 2D/3D convolution layers. When we remove the multi-segment attention maps, the core part in the TSAgg module, the results drop by 0.18dB/0.0013 in terms of fPSNR/fSSIM. The results indicate that the TSAgg module can adaptively emphasize and aggregate important information across different temporal segments.


Composite frames

Masks

TSA^2

GT

Figure 5. A failure case.

4.4. Limitations

Despite the promising performance demonstrated above, the proposed method still has certain limitations in some challenging cases. For example, when the proportion of the foreground region that needs to be adjusted is too large, our method cannot adjust the inharmonious foreground regions because there is not enough background region information to assist the foreground regions to be adjusted. The failure case is shown in Figure 5. Also, due to the limited GPU memory footprint, we only experiment on frames with the spatial resolution of 256×256, which is the same as the experimental settings in CO$_2$Net. As the future work, we will explore how to adjust inharmonious videos with large-area foreground regions. Furthermore, we plan to explore the creation of efficient and lightweight video harmonization networks, specifically tailored for real-world applications. This exploration encompasses scenarios like ultra-high resolution video harmonization.

5. Conclusion

In this paper, we present TSA^2, a novel end-to-end method for video harmonization. To effectively leverage complementary information across different inharmonious frames, the input sequence is divided into several temporal segments with different frame rates. The proposed TSAda module learns the distribution difference between the background and foreground regions and remaps the distribution of foreground features to that of background features, followed by a TSAgg module to emphasize and aggregate cross-segment information adaptively. Our method achieves superior results both quantitatively and qualitatively on representative video harmonization dataset.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grants 62131003 and 62021001.

References

[1] Shan An, Si Liu, Zhibiao Huang, Guangfu Che, Qian Bao, Zhaoqi Zhu, Yu Chen, and Dennis Z. Weng. Rotateview: A video composition system for interactive product display. *IEEE Trans. Multimedia*, 21(12):3095–3105, 2019. 1

[2] Anand Bhattad and David A Forsyth. Cut-and-paste neural rendering. *arXiv preprint arXiv:2010.05907*, 2020. 2

[3] Xun Cai, Qingjie Shi, Yanbo Gao, Shuai Li, Wei Hua, and Tian Xie. A structure-preserving and illumination-consistent cycle framework for image harmonization. *IEEE Trans. Multimedia*, 2023. 2

[4] Junyan Cao, Wenyan Cong, Li Niu, Jianfu Zhang, and Liqing Zhang. Deep image harmonization by bridging the reality gap. In *Brit. Mach. Vis. Conf.*, 2021. 2

[5] Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S Yu, and Lichao Sun. A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt. *arXiv preprint arXiv:2303.04226*, 2023. 1

[6] Bor-Chun Chen and Andrew Kae. Toward realistic image compositing with adversarial learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8415–8424, 2019. 2

[7] Daniel Cohen-Or, Olga Sorkine, Ran Gal, Tommer Leyvand, and Ying-Qing Xu. Color harmonization. In *ACM Trans. Graph.*, volume 25, pages 624–630. 2006. 2

[8] Wenyan Cong, Li Niu, Jianfu Zhang, Jing Liang, and Liqing Zhang. Bargainnet: Background-guided domain translation for image harmonization. In *Int. Conf. Multimedia and Expo*, pages 1–6, 2021. 1, 2, 6

[9] Wenyan Cong, Xinhao Tao, Li Niu, Jing Liang, Xuesong Gao, Qihao Sun, and Liqing Zhang. High-resolution image harmonization via collaborative dual transformations. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 18470–18479, 2022. 2

[10] Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. Dovenet: Deep image harmonization via domain verification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8394–8403, 2020. 1, 2, 6

[11] Xiaodong Cun and Chi-Man Pun. Improving the harmony of the composite image by spatial-separated attention module. *IEEE Trans. Image Process.*, 29:4759–4771, 2020. 6

[12] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Int. Conf. Comput. Vis.*, pages 6202–6211, 2019. 1, 3

[13] Julian Jorge Andrade Guerreiro, Mitsuru Nakazawa, and Björn Stenger. Pct-net: Full resolution image harmonization using pixel-wise color transformations. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5917–5926, 2023. 2

[14] Zonghui Guo, Zhaorui Gu, Bing Zheng, Junyu Dong, and Haiyong Zheng. Transformer for image harmonization and beyond. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022. 1, 2

[15] Zonghui Guo, Dongsheng Guo, Haiyong Zheng, Zhaorui Gu, Bing Zheng, and Junyu Dong. Image harmonization with transformer. In *Int. Conf. Comput. Vis.*, pages 14870–14879, 2021. 1, 2, 5, 6, 7

[16] Zonghui Guo, Haiyong Zheng, Yufeng Jiang, Zhaorui Gu, and Bing Zheng. Intrinsic image harmonization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 16367–16376, 2021. 1, 2, 5, 6

[17] Yucheng Hang, Bin Xia, Wenming Yang, and Qingmin Liao. Scs-co: Self-consistent style contrastive learning for image harmonization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 19710–19719, 2022. 2

[18] Guoqing Hao, Satoshi Iizuka, and Kazuhiro Fukui. Image harmonization with attention-based deep feature modulation. In *Brit. Mach. Vis. Conf.*, 2020. 1, 2, 6

[19] Haozhi Huang, Senzhe Xu, Junxiong Cai, Wei Liu, and Shimin Hu. Temporally coherent video harmonization using adversarial networks. *IEEE Trans. Image Process.*, 29:214–224, 2020. 1, 2, 3, 5, 6, 7

[20] Jiaya Jia, Jian Sun, Chi-Keung Tang, and Heung-Yeung Shum. Drag-and-drop pasting. *ACM Trans. Graph.*, 25(3):631–637, 2006. 2

[21] 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 *Int. Conf. Comput. Vis.*, pages 4832–4841, 2021. 2

[22] Zhanghan Ke, Chunyi Sun, Lei Zhu, Ke Xu, and Rynson WH Lau. Harmonizer: Learning to perform white-box image and video harmonization. In *Eur. Conf. Comput. Vis.* Springer, 2022. 2, 6

[23] Jingtang Liang, Xiaodong Cun, Chi-Man Pun, and Jue Wang. Spatial-separated curve rendering network for efficient and high-resolution image harmonization. In *Eur. Conf. Comput. Vis.*, pages 334–349, 2022. 2

[24] Jun Ling, Han Xue, Li Song, Rong Xie, and Xiao Gu. Region-aware adaptive instance normalization for image harmonization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9361–9370, 2021. 2, 4, 5, 6, 8

[25] Sheng Liu, Cong Phuoc Huynh, Cong Chen, Maxim Arap, and Raffay Hamid. Lemart: Label-efficient masked region transform for image harmonization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 18290–18299, 2023. 2

[26] Yuhao Liu, Jiake Xie, Yu Qiao, Yong Tang, and Xin Yang. Prior-induced information alignment for image matting. *IEEE Trans. Multimedia*, 24:2727–2738, 2022. 1

[27] Xinyuan Lu, Shengyuan Huang, Li Niu, Wenyan Cong, and Liqing Zhang. Deep video harmonization with color mapping consistency. *Int. Joint Conf. Artif. Intell.*, 2022. 1, 3, 4, 5, 6, 7

[28] Bingbing Ni, Mengdi Xu, Bin Cheng, Meng Wang, Shuicheng Yan, and Qi Tian. Learning to photograph: A compositional perspective. *IEEE Trans. Multimedia*, 15(5):1138–1151, 2013. 1

[29] Li Niu, Wenyan Cong, Liu Liu, Yan Hong, Bo Zhang, Jing Liang, and Liqing Zhang. Making images real again: A comprehensive survey on deep image composition. *arXiv preprint arXiv:2106.14490*, 2021. 1, 2

[30] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 4, 5

[31] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. volume 22, pages 313–318. 2003. 1, 2

[32] François Pitié, Anil C. Kokaram, and Rozenn Dahyot. Automated colour grading using colour distribution transfer. *Comput. Vis. Image Underst.*, 107(1-2):123–137, 2007. 1, 2

[33] Anyi Rao, Linning Xu, Zhizhong Li, Qingqiu Huang, Zhanghui Kuang, Wayne Zhang, and Dahua Lin. A coarse-to-fine framework for automatic video unscreen. *IEEE Trans. Multimedia*, 2022. 1

[34] Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *Computer Graphics and Applications*, 21(5):34–41, 2001. 1, 2

[35] Xuqian Ren and Yifan Liu. Semantic-guided multi-mask image harmonization. In *Eur. Conf. Comput. Vis.*, pages 564–579, 2022. 2

[36] Konstantin Sofiiuk, Polina Popenova, and Anton Konushin. Foreground-aware semantic representations for image harmonization. In *Winter Conf. on Applications of Comput. Vis.*, pages 1620–1629, 2021. 1, 2

[37] Jian Sun, Jiaya Jia, Chi-Keung Tang, and Heung-Yeung Shum. Poisson matting. *ACM Trans. Graph.*, 23(3):315–321, 2004. 2

[38] Kalyan Sunkavalli, Micah K Johnson, Wojciech Matusik, and Hanspeter Pfister. Multi-scale image harmonization. *ACM Trans. Graph.*, 29(4):1–10, 2010. 1, 2

[39] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Eur. Conf. Comput. Vis.*, pages 402–419. Springer, 2020. 5

[40] Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. Deep image harmonization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3789–3797, 2017. 1, 2, 6

[41] Jeya Maria Jose Valanarasu, He Zhang, Jianming Zhang, Yilin Wang, Zhe Lin, Jose Echevarria, Yinglan Ma, Zijun Wei, Kalyan Sunkavalli, and Vishal M Patel. Interactive portrait harmonization. In *Int. Conf. Learn. Represent.*, 2023. 2

[42] Jue Wang, Maneesh Agrawala, and Michael F. Cohen. Soft scissors: an interactive tool for realtime high quality matting. *ACM Trans. Graph.*, 26(3):9, 2007. 2

[43] Ke Wang, Michaël Gharbi, He Zhang, Zhihao Xia, and Eli Shechtman. Semi-supervised parametric real-world image harmonization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5927–5936, 2023. 2

[44] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *Eur. Conf. Comput. Vis.*, pages 20–36, 2016. 1, 3

[45] Xintao Wang, Kelvin C.K. Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, 2019. 5, 6

[46] Zeyu Xiao, Xueyang Fu, Jie Huang, Zhen Cheng, and Zhiwei Xiong. Space-time distillation for video super-resolution. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2113–2122, 2021. 5

[47] Zeyu Xiao, Wenming Weng, Yueyi Zhang, and Zhiwei Xiong. Eva2: Event-assisted video frame interpolation via cross-modal alignment and aggregation. *IEEE Trans. Computational Imaging*, 8:1145–1158, 2022. 5

[48] Yazhou Xing, Yu Li, Xintao Wang, Ye Zhu, and Qifeng Chen. Composite photograph harmonization with complete background cues. In *ACM Int. Conf. Multimedia*, pages 2296–2304, 2022. 2

[49] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018. 5

[50] Ben Xue, Shenghui Ran, Quan Chen, Rongfei Jia, Binqiang Zhao, and Xing Tang. Dccf: Deep comprehensible color filter learning framework for high-resolution image harmonization. In *Eur. Conf. Comput. Vis.*, pages 300–316, 2022. 2