Video Harmonization with Triplet Spatio-Temporal Variation Patterns

Zonghui Guo[1] Xinyu Han[2] Jie Zhang[1,3] Shiguang Shan[1,3,*] Haiyong Zheng[2,*]
[1]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[2]College of Electronic Engineering, Ocean University of China, Qingdao, China
[3]University of Chinese Academy of Sciences, Beijing, China
{guozonghui,zhangjie,sgshan}@ict.ac.cn, hanxinyu8585@stu.ouc.edu.cn, zhenghaiyong@ouc.edu.cn

Abstract

Video harmonization is an important and challenging task that aims to obtain visually realistic composite videos by automatically adjusting the foreground's appearance to harmonize with the background. Inspired by the short-term and long-term gradual adjustment process of manual harmonization, we present a Video Triplet Transformer framework to model three spatio-temporal variation patterns within videos, i.e.*, short-term spatial as well as long-term global and dynamic, for video-to-video tasks like video harmonization. Specifically, for short-term harmonization, we adjust foreground appearance to consist with background in spatial dimension based on the neighbor frames; for long-term harmonization, we not only explore global appearance variations to enhance temporal consistency but also alleviate motion offset constraints to align similar contextual appearances dynamically. Extensive experiments and ablation studies demonstrate the effectiveness of our method, achieving state-of-the-art performance in video harmonization, video enhancement, and video demoiréing tasks. We also propose a temporal consistency metric to better evaluate the harmonized videos. Code is available at https://github.com/zhenglab/VideoTripletTransformer.*

1. Introduction

Video compositing is a typical operation that involves extracting a desired region (as foreground) from one video clip and pasting it into another video (as background) to create unique visual effects. However, composite video inevitably suffers from visual inconsistencies due to differences in appearance between foreground and background, such as color, brightness, and contrast [5, 13]. The manual creation of a visually natural composite video is a labor-intensive and expert-level work that demands careful adjustment of pixel intensities frame-by-frame. Thus, **video harmonization (VH)**, aiming to automatically align the foreground appearance with the background in composite videos, has emerged as a critical and challenging task [17, 31].

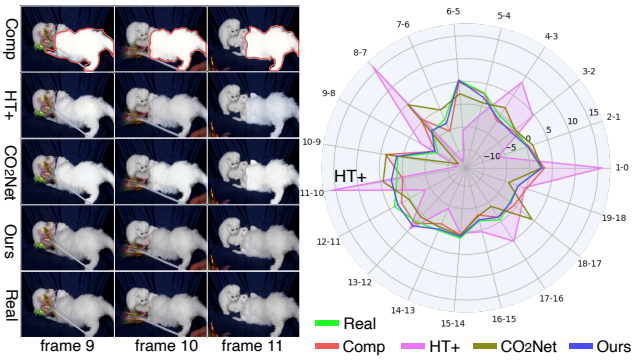

Figure 1. We present harmonized results from image-based (HT+ [14]) and video-based (CO_2Net [31] and Ours) methods (left), along with the inter-frame foreground brightness differences in the video (right). In the radar chart, values that differ greatly with "Real" indicate potential flickering, and the closer the overlap with "Real", the better the visual effect. HT+ and CO_2Net exhibit flickering, while our method closely resembles the real video.

Applying image harmonization methods [5, 13, 14, 41] to composite videos leads to undesirable inter-frame flickering, as evident in Figure 1, where HT+ [14] exhibits significant brightness differences in the harmonized video (pink area marked by "HT+" in the radar chart). Indeed, videos capture object motion and appearance changes within the scene, and these continuous spatio-temporal variations provide crucial guidance and constraints for most video tasks, e.g. video action recognition [10, 45] and inpainting [21]. Therefore, modeling the spatio-temporal variation patterns within videos is fundamental and reasonable for VH.

Similar to the local and global properties observed in images, videos also exhibit short-term and long-term temporal characteristics. Various video processing techniques (*e.g.*, SlowFast [10], TDN [45], and TSN [44]) have demon-

*Correspondence to Shiguang Shan and Haiyong Zheng.

This work was supported by the National Natural Science Foundation of China (No. 62276249 and No. 62171421), the TaiShan Scholars Youth Expert Program of Shandong Province (No. tsqn202306096) and the Youth Innovation Promotion Association CAS.

strated the advantages of considering these different multi-frame changing motions in the temporal dimension. However, unlike high-level tasks such as video classification and action recognition, which rely on detecting changes in motion, video-to-video tasks like video harmonization primarily concentrate on appearance changes while keeping their semantic features constant [13].

Actually, human tackles video harmonization as a gradual optimization process that handles short-term differences based on neighbor frames and progressively extends to long-term frames. This iterative process involves adjusting the foreground appearance from coarse to fine to achieve overall spatio-temporal consistency across the video. Inspired by this intuition, we describe the iterative adjustment in manual harmonization as a mechanism of triplet joint harmonization, which internally captures spatio-temporal variation patterns within different numbers or locations of frames, gradually optimizing the composited video.

Technically, we leverage Transformer [29, 42] to construct an innovative framework, *i.e*., *Video Triplet Transformer (VTT)*. Triplet Transformer consists of short-term spatial, long-term global, and long-term dynamic Transformer modules, each module aims to capture and process spatio-temporal variation patterns across different frame counts or locations within videos. Specifically, in the short-term spatial module, we leverage both spatial global features and temporal subtle changes between neighbor frames to improve the spatial consistency of the video; in the long-term global module, we explore spatio-temporal appearance variation trends to enhance the global temporal consistency within videos, besides, inspired by the powerful representation capabilities of BERT [9] and MAE [16] in capturing intrinsic relationships within sequential data, we introduce a masked prediction strategy to stimulate its potential for modeling long-term variation patterns; in the long-term dynamic module, we utilize dynamic spatial feature matching to alleviate motion offset effects, ensuring that appearances of similar contextual elements (e.g., objects and textures) align across different spatial positions and frames.

Furthermore, human visual system is highly sensitive to the flickering phenomenon in videos [4], which can be caused by sudden alterations in pixel intensities of individual frames. However, previous temporal consistency metrics often fail to capture these abrupt changes, as they rely on averaging results over all frames [8, 18, 23]. Hence, we present a temporal consistency metric tailored for video-to-video tasks, particularly video harmonization, which can detect and magnify the impact of outlier values on final evaluation results using an anchor value.

Our contributions include: (1) We build a Video Triplet Transformer framework that can effectively explore spatio-temporal variation patterns across frames with different lengths and locations; (2) We propose a temporal consistency metric that is suitable for video-to-video tasks. (3) We present comprehensive experiments to demonstrate the effectiveness of our framework, achieving state-of-the-art performance on video harmonization and two related tasks, *i.e*., video enhancement and video demoiréing.

2. Related Work

Image and Video Harmonization. Existing learning-based image harmonization methods can be categorized as semantic guidance [38, 41], domain verification [5, 6], intrinsic images [12, 13], style transfer [27], color space transformation [7, 11, 20, 24, 48], and self-supervised pre-training [19, 28, 33]. The lack of temporal coherence in image harmonization leads to flickering in harmonized videos. Recently, Huang *et al*. [17] made an alignment between harmonized consecutive frames by optical flow to constrain temporal consistency. Lu *et al*. [31] utilized assumption of color mapping consistency of neighboring frames to refine current frame of the videos. However, these methods rely on specific training data or prior assumptions and harmonize videos frame-by-frame without considering long-term relationships, yielding a slight boost on spatio-temporal consistency of harmonized videos. Different from these, we devote to solving video harmonization from a novel perspective of modeling triplet spatio-temporal variation patterns.

Video Temporal Consistency (TC). Previous methods mainly relied on optical flow to align objects across frames, improving temporal consistency in tasks like video inpainting [21], video denoising [49, 50], and video super-resolution [37]. Recently, Deformable DETR [51] extended into some video tasks, such as classification [43] and restoration [25], by using motion displacements or optical flow between consecutive frames to sample relevant points. But the computation of motion displacements or optical flow pose significant challenges and introduce potential inaccuracies. And these methods handle fixed-term TC of few frames, while ours implement iterative harmonization of short- and long-term. Especially, long-term global and long-term dynamic modules adjust spatio-temporal appearance at global and similar contexts cooperatively.

Video TC Metrics. Relation-based TC metrics [8, 18] are commonly utilized for video quality assessment, primarily calculating differences in temporal relations between generated and ground-truth videos. However, we observe existing metrics failing to capture flickering due to the average over all frames, thus we seek to provide a better metric for evaluating TC especially flickering. Besides, flow-based TC metrics [23, 31] calculate pixel differences of aligned frames with estimated optical flow and deem smaller value as better (≈ 0). We obtained a value of 527 on 636 real videos, so we omitted these metrics considering their contradiction with the natural scene changes and their reliance on optical flow accuracy.

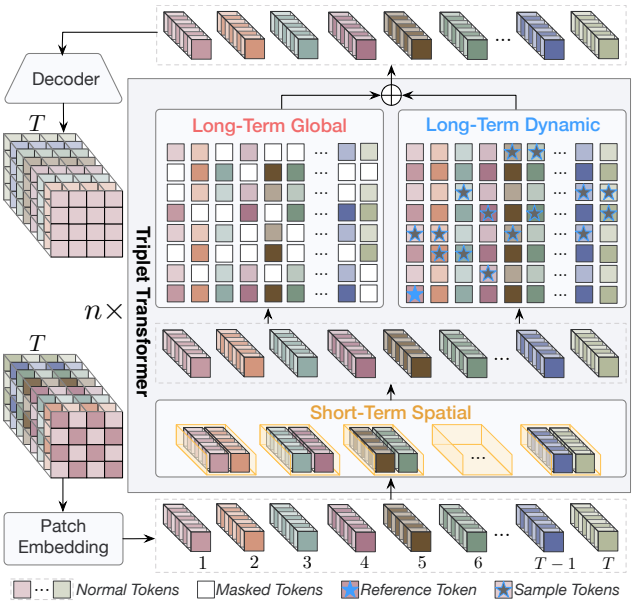

Figure 2. Our Video Triplet Transformer (VTT) framework consists of patch embedding, multi-layer Triplet Transformer with Short-Term Spatial Transformer (ST-ST), Long-Term Global Transformer (LT-GT), and Long-Term Dynamic Transformer (LT-DT) modules, and Decoder. The three Transformer modules aim to model three spatio-temporal variation patterns in videos: *spatial*, *global*, and *dynamic*. The LT-GT improves its ability to enhance global appearance consistency through our masked prediction strategy, and the LT-DT aligns appearance in dynamic contexts by using a reference token and sampled tokens of the context.

3. Framework and Method

We strive to exploit the spatio-temporal variation patterns of varying durations for video-to-video tasks (*e.g.*, video harmonization), which receive a source video and produce the target video that closely resembles the real video. As shown in Figure 2, we present our Video Triplet Transformer (VTT) framework, which comprises a patch embedding, a multi-layer Triplet Transformer, and a decoder. Our approach starts by tokenizing long-term frames from the source video into token sequences using the patch embedding. These token sequences are then fed into the multi-layer Triplet Transformer for iterative adjustments guided by three spatio-temporal variation patterns. Finally, these refined token sequences are utilized by the decoder to reconstruct the target video. Next, we will describe our Triplet Transformer and its application in video harmonization.

3.1. Video Triplet Transformer

The spatio-temporal consistency of video primarily depends on three aspects: spatial content and temporal appearance on global and object motion trajectories, which are crucial for humans to perceive integrality, coherence, and continuity in scenes, respectively. Specifically, each frame in a video captures specific momentary scene information, and the integrity of its content is the foundation of high-quality videos. Moreover, global temporal appearance determines visual coherence, mainly influenced by how objects interact with lighting conditions [13]. Meanwhile, object motion trajectories track the movement and interaction of elements within the scene, providing a sense of dynamic continuity. These factors fundamentally impact the viewer's experience and perception of the video's realism and naturalness.

Evidently, spatial and temporal consistency manifest themselves at both short-term and long-term temporal scales. Therefore, we aim to enhance the spatio-temporal consistency of videos within long sequences by considering three key aspects: short-term spatial appearance, long-term global appearance, and long-term dynamic context. We collectively refer to the three aspects of *spatial*, *global*, and *dynamic* as the triplet spatio-temporal variation patterns in videos, with each corresponding to short-term spatial Transformer, long-term global Transformer, and long-term dynamic Transformer in our Triplet Transformer.

Short-Term Spatial Transformer (ST-ST). We aim to build a specialized module that focuses on adjusting the spatial features to promote visual effects. Meanwhile, we intend to leverage the subtle variations from the neighbor frames as a reference to ensure their appearance is consistent within the neighbor frames. Moreover, Transformer architecture [2, 29, 42] has emerged as a fundamental paradigm for most computer vision tasks with promising performance. Thus, we leverage the powerful contextual capabilities of Transformer to explore and concurrently adjust the spatial features within neighbor frames.

Specifically, our ST-ST receives an input token sequence $z \in \mathbb{R}^{T\times N\times C'}$ and partitions it into independent short-term groups as $z' \in \mathbb{R}^{T//T_{st}\times(T_{st}\times N\times C')}$. These groups are then flattened into 1D token sequences and fed into Transformer to produce adjusted spatial features as z^{st}, which subsequently are inversely reshaped to their original dimensions as $z^{st} \in \mathbb{R}^{T\times N\times C'}$. Here, z is obtained from long-term frames using the patch embedding, T and T_{st} represent the total number of input frames and the number of frames in each short-term group, respectively, P is the patch size, C' is the token dimensions, and N represents the number of patches within each frame calculated by $N = \frac{H}{P} \times \frac{W}{P}$, where H and W are height and width of the source video.

Long-Term Global Transformer (LT-GT). Indeed, motion variations across long-term frames provide valuable temporal information for understanding object structure and shape, yielding discriminative features for high-level video tasks, such as video action recognition [10, 45] and video semantic segmentation [44]. These established and effective methods further support the idea exploring global appearance variation patterns in videos over long temporal sequences will be highly appreciated for video-to-video tasks.

However, existing video-to-video methods often lack the utilization of long-term scene change information, typically concentrating on only a few consecutive frames [17, 31, 49]. Hence, we seek to enhance the global consistency of videos by capturing long-term appearance variations.

Many video-to-video tasks, such as video harmonization and enhancement, aim to address appearance inconsistencies and degradation problems caused by imaging conditions. Essentially, these tasks revolve around exploring and adjusting the low-level features of the video to improve visual quality. Meanwhile, unlike the semantic differences introduced by motion offset, the temporal and spatial appearance variations in real videos are typically consistent and gradual, influenced by lighting conditions with a smooth transition [12, 13]. Therefore, to enhance the target video's temporal consistency, we leverage the long-context capabilities of Transformer to explore the spatio-temporal global appearance variation patterns across long-term frames.

Specifically, our LT-GT receives $z^{st} \in \mathbb{R}^{T\times N\times C'}$ from ST-ST and segments it into independent windows in the spatial dimensions to obtain $z^{st}_{gt} \in \mathbb{R}^{T\times(N//M)\times M\times C'}$ (where M represents the product of the window's width and height). Then, we reshape and flatten z^{st}_{gt} across the temporal dimensions to obtain $z'^{st}_{gt} \in \mathbb{R}^{(N//M)\times T\cdot M\times C'}$, and feed it into Transformer to produce adjusted features, which subsequently are inversely reshaped to their original shape $z^{gt} \in \mathbb{R}^{T\times N\times C'}$.

Furthermore, we consider that the abundance of redundant similar features within windows at the same spatial positions hinders the learning of temporal appearance variation patterns across long-term frames. To alleviate this issue and unlock the full potential of the LT-GT, we investigate the performance of two self-supervised learning mechanisms in video-to-video tasks: autoregressive (e.g., GPT [36]) and masked prediction (e.g., BERT [9], MAE [16]). Refer to Section 5.4 for empirical study and analysis.

Long-Term Dynamic Transformer (LT-DT). Then, we direct our attention towards the issue of ensuring dynamic contextual continuity within the target video. Since the target video maintains the same content information as the source video, we are mainly concerned with the appearance continuity of similar objects between long-term frames. However, the spatial location misalignment of objects between frames is an inherent characteristic of videos, and it presents a substantial challenge to enhance the temporal appearance continuity of target video.

To mitigate the impact of motion offset, we expand the use of deformable attention [51] to dynamically locate and align the temporal appearance of similar contexts within videos. This dynamic is achieved by employing relative position offsets and attention weights to sample tokens across long-term frames that share similarities with a reference token and assigning them varying levels of attention.

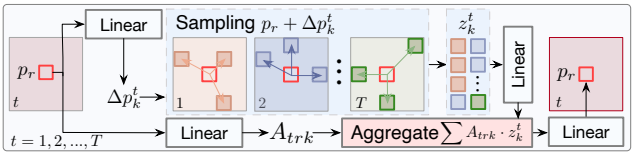

Figure 3. Implementation process of our LT-DT module.

Figure 3 illustrates LT-DT's process, involving sampling tokens by position offsets Δp^t_k and aggregating them using attention weights A_{trk} to align the appearance of moving objects. Specifically, LT-DT receives $z^{st} \in \mathbb{R}^{T\times N\times C'}$ from ST-ST and divides it into iterative groups, each consisting of T_{dt} frames, denoted as $z^{st}_{dt} \in \mathbb{R}^{T//T_{dt}\times(T_{dt}\times N\times C')}$. These groups are then flattened into 1D token sequences and fed into Transformer with the deformable attention to produce aligned features, which subsequently are inversely reshaped to their original dimensions as $z^{dt} \in \mathbb{R}^{T\times N\times C'}$. For simplicity, we present a formalization of the single-header deformable attention operation as follows:

$$DeAttn(z_r) = W\left[\sum_{t=1}^{T_{dt}}\sum_{k=1}^{k} A_{trk}\cdot W' z^t_{p_r+\Delta p^t_k}\right], \quad (1)$$

where z_r and p_r represent the r-index reference token and normalized coordinates of the 2D reference token, respectively, Δp^t_k denotes the sampling offset of the k-th sampling token in the t-th frame, W and W' represent the weights of linear projections, and the attention weight A_{trk} normalized by $\sum_{t=1}^{T_{dt}}\sum_{k=1}^{k} A_{trk} = 1$. Besides, A_{trk} and Δp^t_k are learnable through linear projection over z_r.

Overall, we sum the outputs z^{gt} and z^{dt} from LT-GT and LT-DT as the outputs of our Triplet Transformer.

3.2. Video Harmonization Triplet Transformer

Given a composite video $\tilde{\mathbf{V}}$ and a foreground mask $\mathbf{M}$ which indicates the inharmonious region, our goal is to learn a model that takes $\tilde{\mathbf{V}}$ and $\mathbf{M}$ as inputs and produces a harmonized video $\hat{\mathbf{V}}$ as output, where $\hat{\mathbf{V}}$ is expected to be as harmonious as the real video $\mathbf{V}$.

Based on our VTT framework, we devise a Video Harmonization Triplet Transformer (VHTT) method, which aims to harmonize $\tilde{\mathbf{V}}$ to $\hat{\mathbf{V}}$ by exploiting spatio-temporal variation patterns within triplet groups of varying temporal lengths and locations. In VHTT, we first employ a multi-layer ST-ST to harmonize the foreground, making it coarse harmony with the background. Then, we apply a multi-layer Triplet Transformer to refine foreground appearance and enhance temporal consistency iteratively.

Formally, we first concatenate $\tilde{\mathbf{V}}$ and $\mathbf{M}$, and embed them into token sequence $z \in \mathbb{R}^{T\times N\times C'}$ through a linear projection, then feed it into m layers of ST-ST for producing spatial harmonized tokens z_{st}. Further, we employ n layers of Triplet Transformer, consisting of ST-ST (TR_{ST}), LT-GT (TR_{GT}), and LT-DT (TR_{DT}), taking z_{st} as input for

producing spatio-temporal consistency tokens z_{tt}. Finally, we inversely reshape z_{tt} back to 3D feature maps and feed it into 2D-CNN decoder for yielding harmonized video $\hat{\mathbf{V}}$. The Triplet Transformer process is formulated as follows:

$$z'_{st} = TR_{ST}(z_{st}), \tag{2}$$

$$z_{tt} = TR_{GT}(z'_{st}) + TR_{DT}(z'_{st}). \tag{3}$$

The loss function $\mathcal{L}$ of our VHTT model comprises a reconstruction error $\mathcal{L}_1$ and a relation-based penalty [8] $\mathcal{L}_R$, which constrain the spatial and temporal consistency of the harmonized video, respectively.

$$\mathcal{L}_R = \frac{1}{T-1}\sum_{t=1}^{T-1}\|(\mathbf{V_{t+1}} - \mathbf{V_t}) - (\hat{\mathbf{V}}_{\mathbf{t+1}} - \hat{\mathbf{V}}_{\mathbf{t}})\|_2^2, \tag{4}$$

$$\mathcal{L} = \frac{1}{T}\sum_{t=1}^{T}\|\mathbf{V_t} - \hat{\mathbf{V}}_{\mathbf{t}}\|_1 + \lambda\mathcal{L}_R, \tag{5}$$

where λ is the weight to control the contribution of $\mathcal{L}_R$.

4. Temporal Consistency Metric

Relation-based Temporal Consistency (RTC) metrics [8, 18] establish the temporal relation within a video by calculating the mean of pixel intensity gradients from frames $(t+1)$-th to t-th, then measuring the differences in temporal relations between the target video and the paired real video, which we denote RTC_t. The RTC is calculated by averaging all time steps:

$$\text{RTC}_t = \|\frac{1}{K}\sum_{k=1}^{K}(\mathbf{V}_{t+1}^k - \mathbf{V}_t^k) - \frac{1}{K}\sum_{k=1}^{K}(\hat{\mathbf{V}}_{t+1}^k - \hat{\mathbf{V}}_t^k)\|_2,$$

$$\text{RTC} = \frac{1}{T-1}\sum_{t=1}^{T-1}\text{RTC}_t^2, \quad t \in (1, 2, ..., T-1). \tag{6}$$

where K is the pixel number of t-th frame.

Indeed, flickering in videos is a particularly troublesome issue that often stems from various distortions [4]. This phenomenon can be reflected by the value of RTC_t, which involves rapid and frequently irregular fluctuations in brightness or color. Human visual system is naturally adapted to consistent lighting and color conditions, and such rapid fluctuations disrupt the smooth visual processing that the human eye is accustomed to, leading to a noticeable interruption in the viewing experience and causing discomfort to viewers [1, 32, 39]. Therefore, considering flickering in videos is crucial for temporal consistency evaluation.

As indicated by Equation 6, the abrupt increase in RTC_t, which reflects video flickering, will be significantly reduced when averaged with the other intensity differences over $T-1$ frames. Based on this clue, we seek to address this limitation in RTC by introducing variance to identify and amplify signals associated with abrupt changes in RTC_t. We then propose a Refined RTC (R-RTC) metric, which has a robust ability to capture the appearance of flickering in target videos without compromising its original evaluation capabilities. The calculation process is as follows:

$$\mu = \frac{1}{T-1}\sum_{t=1}^{T-1}\text{RTC}_t, \quad \text{RTC}'_t = max(\text{RTC}_t - \mu, 0) \tag{7}$$

$$\text{R-RTC} = \frac{1}{T-1}\sum_{t=1}^{T-1}(\text{RTC}_t^2 + {\text{RTC}'_t}^2). \tag{8}$$

Due to our visual system's high sensitivity to changes in brightness, using the brightness may be more suitable for evaluating temporal consistency [15, 26, 40]. Moreover, the HSV color space aligns more closely with human visual perception than the RGB color space [34]. Thus, we conduct further investigations into R-RTC's performance in different color spaces, *i.e.*, Value in the HSV, RGB, and Gray. Experimental results indicate that Value channel yields the best for R-RTC. Refer to Section 5.5 for empirical analysis.

Similar to analysis of MSE and fMSE [13], metrics for temporal consistency in harmonized videos should focus on the changes in foreground region between frames. However, the inter-frame foreground regions may shift positions, making it impossible to calculate the RTC_t within the foreground region directly pixel-by-pixel using Equation 6.

In fact, human visual system typically perceiving temporal appearance changes at the region level. Therefore, we extend the calculation of RTC_t from pixel intensity to the mean region pixel intensity, without being constrained by position offset. Thus, we provide a temporal consistency metric in the foreground, named fR-RTC. First, we calculate the mean v_t^f within foreground region of each frame. Then, we compute the RTC_t^f values between harmonized video and real video, formulated as:

$$v_t^f = \frac{1}{K^f}\sum_{k=1}^{K}\mathbf{V}_t^k\mathbf{M}_t^k \tag{9}$$

$$\text{RTC}_t^f = \|(v_{t+1}^f - v_t^f) - (\hat{v}_{t+1}^f - \hat{v}_t^f)\|_2, \tag{10}$$

where K^f is the foreground pixel number of t-th frame, and $\mathbf{M}$ denotes the foreground mask. Finally, we compute the R-RTC using Equations 7 and 8 by providing RTC_t^f as RTC_t, yielding our fR-RTC (the computed R-RTC).

5. Experiments on Video Harmonization

5.1. Dataset and Metrics

Dataset. We conduct experiments on the public synthesized HYouTube dataset [31], created by adjusting the appearance of foreground regions in videos from the YouTube-VOS dataset [47]. HYouTube consists of 3194 pairs of synthetic and real videos, with 2558 pairs used for training and 636 pairs for testing [31]. Each video is composed of 20 frames along with their corresponding foreground masks.

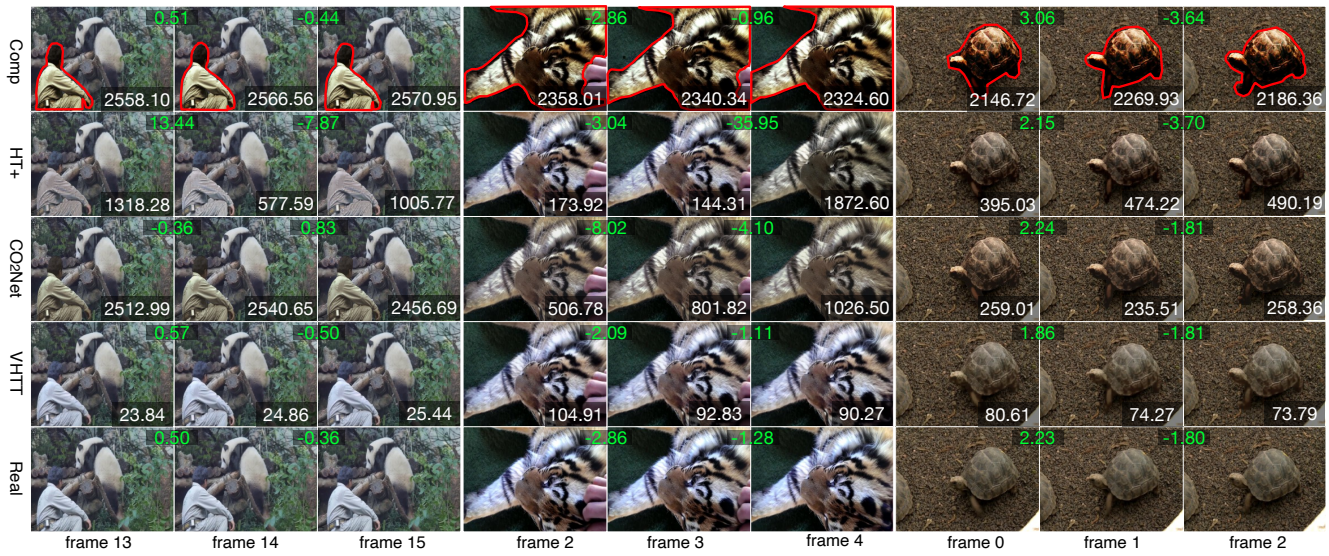

Figure 4. Qualitative comparison of different harmonization methods on HYouTube dataset [31]. The white and green numbers represent fMSE↓ and inter-frame brightness difference (the closer to "Real", the better). Red boxes in composite frames mark foreground.

Evaluation Metrics. We evaluate the quality of harmonized videos from both spatial and temporal dimensions. For spatial dimensions, we use PSNR and foreground region metrics (fMSE and fPSNR) for a better indication of evaluating harmonization ability of the method. For temporal dimensions, we employ our fR-RTC metric to evaluate the temporal consistency of harmonized videos. Additionally, we consider fMSE and fR-RTC as the primary metrics.

5.2. Implementation Details

We train our model using Adam optimizer [22] with parameters of $\beta_1 = 0.5$, $\beta_2 = 0.999$ for 400 epochs. The initial learning rate is 10^{-4} for the first 200 epochs and linearly decayed to zero over the next 200 epochs. We resize source videos as 256×256 for training and testing, and our model produces harmonized videos of the same size. During the training phase, we randomly sample 5 frames with varying frame rates from a video for each batch, while during the inference phase, we use sequences of 20 or more consecutive frames for each batch. We empirically set $\lambda = 5$ for training our model. More details are in *supplementary file*.

5.3. Comparison with State-of-the-arts

We compare our VHTT method with state-of-the-art video harmonization methods: TCVHAN [17] and CO_2Net [31], as well as image harmonization methods: IIH [13], RainNet [27], BargainNet [6] and HT+ [14].

Table 1 shows the quantitative comparison of video harmonization on HYouTube dataset [31]. Columns 2 to 4 and columns 5 to 6 represent the evaluation results in spatial and temporal dimensions, respectively. As we can see, our VHTT model achieves state-of-the-art performance across all spatial and temporal metrics, demonstrating the advanced capabilities and effectiveness of our VTT framework, which adeptly extracts and adjusts triplet spatio-temporal variation patterns in videos. Furthermore, the low R-RTC values across all methods can be attributed to numerous pixels having zero changes, significantly impacting the overall average. This phenomenon occurs because the harmonized video shares the same background as the real video. In contrast, fR-RTC exclusively considers the foreground region, providing a more accurate reflection of the harmonization effect. Besides, column 7 indicates that HT+ has lower costs but disregards the indispensable temporal consistency in video tasks, while VHTT outperforms CO_2Net in cost-efficiency and speed.

Method	Spatial			Temporal		Inference
	PSNR↑	fPSNR↑	fMSE↓	R-RTC↓	fR-RTC↓	FLOPs/Times
Composite	30.14	19.92	1029.50	0.52	11.3	–
RainNet [27]	35.47	25.22	330.50	0.38	11.13	$379G/2.15s$
IIH [13]	35.59	25.85	296.24	0.39	7.60	$3778G/0.50s$
BargainNet [6]	35.41	26.00	293.09	0.30	8.23	$385G/2.68s$
HT+ [14]	38.55	29.44	154.20	0.6	12.51	$212G/0.42s$
TCVHAN [17]	37.44	27.34	199.89	0.59	15.23	$301G/0.22s$
CO_2Net [31]	37.61	27.56	186.72	0.24	6.07	$5190G/1.65s$
VHTT	**40.03**	**31.23**	**90.35**	**0.03**	**1.26**	$1727G/1.24s$

Table 1. Quantitative comparisons of different harmonization methods on HYouTube dataset [31], with Inference FLOPs/Times measured on a 20-frame video using one 3090 GPU.

Figure 4 illustrates the qualitative comparison results of video harmonization on HYouTube dataset [31]. It demonstrates that our VHTT method, benefiting from the VTT framework's capacity to handle triplet variation patterns, achieves the best visual effect comparable to real videos across spatial and temporal dimensions.

Method	ST-ST	ST-ST &LT-GT w/o masked	ST-ST &LT-GT	ST-ST &LT-DT	VHTT w/o masked	VHTT
fMSE↓	111.59	99.97	95.64	105.24	91.97	**90.35**
fR-RTC↓	2.48	1.39	1.30	2.10	1.36	**1.26**

Table 2. Quantitative comparison of using our Triplet Transformer with different modules. "w/o masked" means vanilla self-attention in LT-GT is used without the masked prediction strategy.

Metric	TR	Auto-regressive F-D	Bi-D	FB-D	Masked prediction MS	MS&50	M50	M75	M90
fMSE↓	99.97	106.57	104.05	98.83	98.54	**94.16**	94.76	95.64	96.29
fR-RTC↓	1.39	1.80	1.58	1.45	1.46	1.41	1.35	**1.30**	1.46

Table 3. Quantitative comparison of using different mechanisms for LT-GT. TR denotes the vanilla self-attention mechanism.

5.4. Ablation Studies

Analysis of Triplet Transformer. We then conduct experiments to analyze the efficacy of our Triplet Transformer with ST-ST, LT-GT, and LT-DT as follows: (1) Triplet Transformer with only ST-ST as baseline (ST-ST), (2) Triplet Transformer with ST-ST and LT-GT, where LT-GT uses vanilla self-attention (ST-ST&LT-GT w/o masked), (3) Triplet Transformer with ST-ST and LT-GT (ST-ST&LT-GT), (4) Triplet Transformer with ST-ST and LT-DT (ST-ST&LT-DT), (5) our VHTT model with LT-GT using vanilla self-attention (VHTT w/o masked), and we also present the results of our VHTT model as reference.

The quantitative comparison in Table 2 shows that: (1) both ST-ST&LT-GT and ST-ST&LT-DT outperform ST-ST in terms of spatial and temporal consistency, demonstrating the effectiveness of maintaining consistency in long-term global and dynamic contexts, especially concerning the long-term global appearance, although achieving further performance improvements becomes increasingly challenging when the similarity between harmonized videos and real videos reaches a certain limit, (2) ST-ST&LT-GT outperforms ST-ST&LT-GT w/o masked, as well as VHTT outperforms VHTT w/o masked, indicating that our masked prediction strategy can enhance the ability of LT-GT to capture long-term global variations, (3) VHTT w/o masked outperforms ST-ST&LT-GT w/o masked and ST-ST&LT-DT, as well as VHTT outperforms ST-ST&LT-GT and ST-ST&LT-DT, demonstrating the effectiveness of our processing triplet spatio-temporal variation patterns in videos, *i.e.*, *spatial*, *global*, and *dynamic*. Moreover, we also analyze the boundaries between ST and LT in *supplementary file*.

Analysis of LT-GT. To explore LT-GT's ability to capture long-term temporal global appearance variation patterns, we introduce three mechanisms: (1) vanilla self-attention (TR), (2) auto-regressive with different directionalities, including forward (F-D), bidirectional (Bi-D), and separate forward and backward (FB-D), (3) masked prediction with different masking strategies, including self-window masking (MS), self-window along with random 50% masking (MS&50), random 50% masking (M50), random 75% masking (M75), and random 90% masking (M90) within the current temporal windows.

The quantitative comparison in Table 3 shows that, compared to the TR, auto-regressive strategies (F-D, Bi-D, and FB-D) perform poorly both on fMSE↓ and fR-RTC↓, possibly due to insufficient data. In contrast, M75 and M50 show significant improvements, benefiting from masking a substantial amount of redundant information across the temporal dimension and providing diverse data for model training.

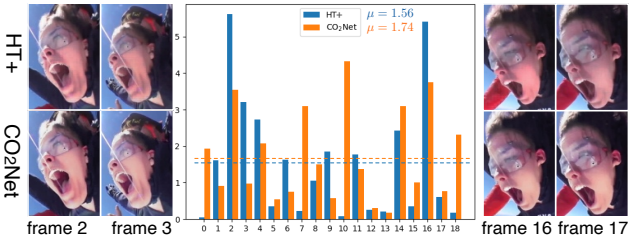

Figure 5. Comparison of the differences RTC_t^f between the harmonized video's and real video's inter-frame relations, as well as examples of harmonized frames. μ is the mean of RTC_t^f.

Method	Value (HSV) fRTC↓	fR-RTC↓	RGB fR-RTC↓	Gray fR-RTC↓
HT+ [14]	5.16	7.07	3.96	3.47
CO_2Net [31]	4.56	5.51	3.60	3.30

Table 4. Comparison of different temporal consistency metrics on the video shown in Figure 5, as well as different color spaces.

5.5. Analysis of R-RTC and fR-RTC

We delve into the temporal consistency metrics to analyze the effectiveness of our R-RTC and fR-RTC, comparing them with the RTC metrics [8, 18]. As discussed in Section 5.3, fR-RTC better reflects the temporal consistency of the harmonized video. Here, we use fR-RTC to analyze its effectiveness corresponding to R-RTC, and similarly, we calculate RTC using only the foreground, named fRTC.

The chart in Figure 5 shows the differences RTC_t^f between the inter-frame relations (from $t+1$ to t) of the harmonized video and those of the paired real video. Compared to CO_2Net [31], HT+ [14] exhibits abrupt changes in RTC_2^f and RTC_{16}^f, indicating flickering occurring between frames 2 and 3 as well as frames 16 and 17. The frames in Figure 5 highlight brightness changes within the facial region in the HT+ method, whereas CO_2Net exhibits subtle variations. Meanwhile, Table 4 shows that HT+ and CO_2Net yield close fRTC results (5.16 vs. 4.56), indicating that fRTC is not sensitive to flickering. In contrast, our fR-RTC metric captures flickering effectively (7.07 vs. 5.51).

We further explore by calculating RTC_t^f in HSV, RGB, and Gray color spaces. Results in Table 4 show that fR-

Method	HYouTube				Real Composite	
	Times↑	Degree↓	fRTC↓	fR-RTC↓	Times↑	Degree↓
HT+ [14]	25.0%	2.54	7.62	10.23	21.6%	2.48
CO_2Net [31]	13.5%	1.79	4.21	6.03	24.1%	1.92
VHTT	**61.5%**	**0.85**	**0.73**	**1.27**	**54.3%**	**1.01**

Table 5. User study comparison on HYouTube dataset and Real Composite Videos [31]. "Times" and "Degree" represent the fraction of times the video was selected as the best and the degree of video flickering rated on a 5-point Likert scale [35], respectively.

Method	PSNR↑	SSIM↑	LPIPS↓	RTC↓	R-RTC↓
Source	19.12	0.4724	0.4724	**0.03**	**0.03**
VEN-Retinex [46]	26.10	0.7450	0.0832	0.37	0.46
Ours	**26.61**	**0.7507**	**0.0747**	0.06	0.08

Table 6. Quantitative comparison of video enhancement on SDSD dataset [46].

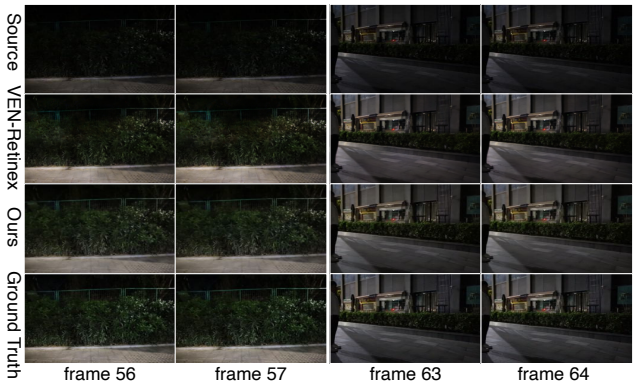

Figure 6. Visual comparison of video enhancement on SDSD dataset [46].

RTC calculated in the Value channel of HSV performs better, aligning with human sensitivity to brightness changes.

5.6. User Study

We finally conduct user study to evaluate our VHTT method alongside CO_2Net[31] and HT+[14] on both HYouTube dataset and Real Composite Videos [31]. The results listed in Table 5 illustrate that our VHTT model achieves the best performance in terms of visual quality and flickering degree. A higher flickering degree indicates more significant flickering, and the degree results also confirm the superiority of our fR-RTC metric over fRTC in capturing flickering within videos. More details are in *supplementary file*.

6. Beyond Video Harmonization

6.1. Video Enhancement

We apply our VHTT method to the video enhancement task on SDSD dataset [46], compared to state-of-the-art VEN-Retinex [46]. Insufficient lighting can lead to video degradation. Table 6 demonstrates our model's superior performance, and Figure 6 further validates the effectiveness of our model in enhancing contrast and color. Besides, all the methods listed may destroy the temporal consistency of source videos, as they are captured in real-world scenarios with inherent consistency characteristics. In contrast, our model excels in maintaining temporal consistency.

Method	PSNR↑	SSIM↑	LPIPS↓	RTC↓	R-RTC↓
Source	16.21	0.6720	0.2073	1.23	1.73
VDRTC [8]	23.71	0.8104	0.0696	1.02	1.32
Ours	**28.56**	**0.8387**	**0.0494**	**0.86**	**1.24**

Table 7. Quantitative comparison of video demoiréing on video demoiréing dataset [8].

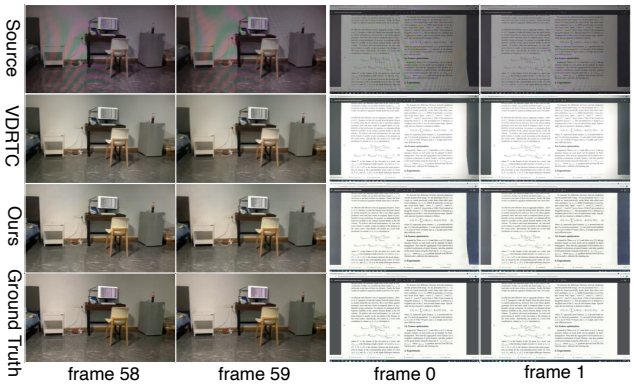

Figure 7. Visual comparison of video demoiréing on video demoiréing dataset [8].

6.2. Video Demoiréing

We further employ our VHTT method to video demoiréing task on the video demoiréing dataset [8], compared to state-of-the-art VDRTC [8]. Video demoiréing aims to remove undesirable moiré patterns in videos, which is caused by frequency aliasing in photographs. Table 7 and Figure 7 demonstrate our model's superior performance in detail recovery and consistency enhancement, thanks to our framework's robust spatio-temporal context capabilities.

7. Conclusion

In this paper, we build a novel framework for video harmonization modeling triplet spatio-temporal variation patterns to address both spatial inharmonies and temporal inconsistencies. We conduct comprehensive experiments to demonstrate the effectiveness of our Video Triplet Transformer framework and employ our method on video harmonization, video enhancement, and video demoiréing tasks, achieving state-of-the-art performance. Besides, we propose a new temporal consistency metric that aligns better with human visual perception. We hope that our work opens up new avenues for further study of video-to-video tasks. Additionally, our work's limitation and societal impact are discussed in *supplementary file*.

References

[1] Nicolai Behmann, Sousa Weddige, and Holger Blume. Psychophysical study of human visual perception of flicker artifacts in automotive digital mirror replacement systems. *Electronic Imaging*, 2021(11):10401–1, 2021. 5

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229, 2020. 3

[3] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *ICML*, pages 1691–1703, 2020. 1

[4] Lark Kwon Choi and Alan C Bovik. Perceptual flicker visibility prediction model. *Electronic Imaging*, 28:1–6, 2016. 2, 5

[5] Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. DoveNet: Deep image harmonization via domain verification. In *CVPR*, pages 8394–8403, 2020. 1, 2

[6] Wenyan Cong, Li Niu, Jianfu Zhang, Jing Liang, and Liqing Zhang. BargainNet: Background-guided domain translation for image harmonization. In *ICME*, 2021. 2, 6

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

[8] Peng Dai, Xin Yu, Lan Ma, Baoheng Zhang, Jia Li, Wenbo Li, Jiajun Shen, and Xiaojuan Qi. Video demoireing with relation-based temporal consistency. In *CVPR*, pages 17622–17631, 2022. 2, 5, 7, 8

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2, 4, 1

[10] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, pages 6202–6211, 2019. 1, 3

[11] Julian Jorge Andrade Guerreiro, Mitsuru Nakazawa, and Björn Stenger. Pct-net: Full resolution image harmonization using pixel-wise color transformations. In *CVPR*, pages 5917–5926, 2023. 2

[12] Zonghui Guo, Dongsheng Guo, Haiyong Zheng, Zhaorui Gu, Bing Zheng, and Junyu Dong. Image harmonization with transformer. In *CVPR*, pages 14870–14879, 2021. 2, 4

[13] Zonghui Guo, Haiyong Zheng, Yufeng Jiang, Zhaorui Gu, and Bing Zheng. Intrinsic image harmonization. In *CVPR*, pages 16367–16376, 2021. 1, 2, 3, 4, 5, 6

[14] Zonghui Guo, Zhaorui Gu, Bing Zheng, Junyu Dong, and Haiyong Zheng. Transformer for image harmonization and beyond. *IEEE TPAMI*, 2022. 1, 6, 7, 8

[15] Lewis O Harvey. Flicker sensitivity and apparent brightness as a function of surround luminance. *JOSA*, 60(6):860–864, 1970. 5

[16] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022. 2, 4, 1

[17] Haozhi Huang, Senzhe Xu, Junxiong Cai, Wei Liu, and Shimin Hu. Temporally coherent video harmonization using adversarial networks. *IEEE TIP*, 29:4267–4278, 2019. 1, 2, 4, 6

[18] Haiyang Jiang and Yinqiang Zheng. Learning to see moving objects in the dark. In *ICCV*, pages 7324–7333, 2019. 2, 5, 7

[19] 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 *ICCV*, pages 4832–4841, 2021. 2

[20] Zhanghan Ke, Chunyi Sun, Lei Zhu, Ke Xu, and Rynson WH Lau. Harmonizer: Learning to perform white-box image and video harmonization. In *ECCV*, pages 690–706. Springer, 2022. 2

[21] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Deep video inpainting. In *CVPR*, pages 5792–5801, 2019. 1, 2

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[23] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *ECCV*, pages 170–185, 2018. 2

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

[25] Jingyun Liang, Yuchen Fan, Xiaoyu Xiang, Rakesh Ranjan, Eddy Ilg, Simon Green, Jiezhang Cao, Kai Zhang, Radu Timofte, and Luc V Gool. Recurrent video restoration transformer with guided deformable attention. In *NeurIPS*, pages 378–393, 2022. 2

[26] Liqun Lin, Shiqi Yu, Liping Zhou, Weiling Chen, Tiesong Zhao, and Zhou Wang. Pea265: Perceptual assessment of video compression artifacts. *IEEE TCSVT*, 30(11):3898–3910, 2020. 5

[27] Jun Ling, Han Xue, Li Song, Rong Xie, and Xiao Gu. Region-aware adaptive instance normalization for image harmonization. In *CVPR*, pages 9361–9370, 2021. 2, 6

[28] Sheng Liu, Cong Phuoc Huynh, Cong Chen, Maxim Arap, and Raffay Hamid. Lemart: Label-efficient masked region transform for image harmonization. In *CVPR*, pages 18290–18299, 2023. 2

[29] 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 *ICCV*, pages 10012–10022, 2021. 2, 3, 1

[30] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *CVPR*, pages 3202–3211, 2022. 1

[31] Xinyuan Lu, Shengyuan Huang, Li Niu, Wenyan Cong, and Liqing Zhang. Deep video harmonization with color mapping consistency. In *IJCAI*, pages 1232–1238, 2022. 1, 2, 4, 5, 6, 7, 8, 3

[32] AI Mozhaeva, IV Vlasuyk, AM Potashnikov, Michael J Cree, and Lee Streeter. The method and devices for research

the parameters of the human visual system to video quality assessment. In *2021 Systems of Signals Generating and Processing in the Field of on Board Communications*, pages 1–5. IEEE, 2021. 5

[33] Li Niu, Junyan Cao, Wenyan Cong, and Liqing Zhang. Deep image harmonization with learnable augmentation. In *ICCV*, pages 7482–7491, 2023. 2

[34] T Ojala, M Rautiainen, E Matinmikko, and M Aittola. Semantic image retrieval with hsv correlograms. In *Proceedings of the Scandinavian conference on Image Analysis*, pages 621–627, 2001. 5

[35] Martin Pernuš, Vitomir Štruc, and Simon Dobrišek. Maskfacegan: High resolution face editing with masked gan latent code optimization. *IEEE TIP*, 2023. 8, 1

[36] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. *OpenAI*, 2018. 4

[37] Mehdi SM Sajjadi, Raviteja Vemulapalli, and Matthew Brown. Frame-recurrent video super-resolution. In *CVPR*, pages 6626–6634, 2018. 2

[38] Konstantin Sofiiuk, Polina Popenova, and Anton Konushin. Foreground-aware semantic representations for image harmonization. In *WACV*, pages 1620–1629, 2021. 2

[39] Jakob Suchan and Mehul Bhatt. Semantic question-answering with video and eye-tracking data: Ai foundations for human visual perception driven cognitive film studies. In *IJCAI*, pages 2633–2639, 2016. 5

[40] Roger B Tootell, John B Reppas, Kenneth K Kwong, Rafael Malach, Richard T Born, Thomas J Brady, Bruce R Rosen, and John W Belliveau. Functional analysis of human mt and related visual cortical areas using magnetic resonance imaging. *Journal of Neuroscience*, 15(4):3215–3230, 1995. 5

[41] Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. Deep image harmonization. In *CVPR*, pages 3789–3797, 2017. 1, 2

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017. 2, 3

[43] Jue Wang and Lorenzo Torresani. Deformable video transformer. In *CVPR*, pages 14053–14062, 2022. 2

[44] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE TPAMI*, 41 (11):2740–2755, 2018. 1, 3

[45] Limin Wang, Zhan Tong, Bin Ji, and Gangshan Wu. Tdn: Temporal difference networks for efficient action recognition. In *CVPR*, pages 1895–1904, 2021. 1, 3

[46] Ruixing Wang, Xiaogang Xu, Chi-Wing Fu, Jiangbo Lu, Bei Yu, and Jiaya Jia. Seeing dynamic scene in the dark: A high-quality video dataset with mechatronic alignment. In *ICCV*, pages 9700–9709, 2021. 8

[47] 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

[48] 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 *ECCV*. Springer, 2022. 2

[49] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *IJCV*, 127:1106–1125, 2019. 2, 4

[50] Songhyun Yu, Bumjun Park, Junwoo Park, and Jechang Jeong. Joint learning of blind video denoising and optical flow estimation. In *CVPRW*, pages 500–501, 2020. 2

[51] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *ICLR*, 2021. 2, 4