

Supplementary Materials: Harmony Everything! Masked Autoencoders for Video Harmonization

Anonymous Authors

This supplementary includes the following items:

- Section 1 elaborates on the architectural details of our model and our end-to-end training configuration.
- Section 2 delves into a comprehensive analysis of the hyperparameters in our VHMAE, specifically investigating the impacts of varying degrees of the α and β in the loss function.
- Section 3 visualizes the qualitative results of our VHAME and other advanced methods, showcasing the large-scale results and analyzing frame consistency.
- Section 4 displays some samples and their corresponding sources of our RCVH dataset.

Our source code and the RCVH dataset will be publicly available for research purposes.

1 IMPLEMENTATION DETAILS

1.1 Model Architectures

Our masked video modeling framework adopts an asymmetric encoder-decoder architecture for video reconstruction, similar to VideoMAE [7]. Notably, our VHMAE operates as an end-to-end network, generating frames directly through the projection head without requiring a fine-tuning stage. We employ an 8-frame vanilla ViT-based model, with detailed architectural specifications for the encoder and decoder provided in Table 1. To enhance the capture of spatio-temporal information within frames, we utilize joint space-time attention [1, 3]. Furthermore, we introduce the Pattern Alignment Module (PAM), comprised of a series of MLPs, designed to align the pattern style between the masked foreground and visible background in the feature space, thereby providing initial information for the masked tokens.

1.2 Training Configuration

Our training setting is depicted in Table 2. Following VideoMAE [7], we do not use color jittering, drop path, or gradient clip.

2 HYPERPARAMETER ANALYSIS

As indicated in Eq. (6) in the manuscript, α and β serve as the factors governing the weights assigned to the two distinct losses in our model, *i.e.*, the Pattern Alignment Loss (L_{align}) and the Patch Balancing Loss ($L_{balance}$), which are crucial for the video harmonization performance. We conducted comprehensive comparison experiments to determine the significance of each term, and the results are presented in Table 3. We found that reducing the weights of these factors during model optimization leads to degraded results, underscoring their critical role in the effectiveness of our model.

3 VISUALIZATION ANALYSIS

3.1 Frames Consistency

In video harmonization, ensuring frame-to-frame coherence is critical. To evaluate this, we analyzed the color consistency between

Table 1: Architectures details of our VHMAE. $MHA(\cdot)$ denotes the joint space-time self-attention, $MLP(\cdot)$ indicates the multi-layer perceptions, and N_{p_m} represents the number of masked patches. The output sizes are denoted by $\{C \times T \times S\}$ for channel, temporal, and spatial sizes.

Term	Layer	Output Size
input video	resize to 256×256	$3 \times 8 \times (256 \times 256)$
patch embedding	$MLP(768)$ stride $(2 \times 16 \times 16)$	$768 \times 4 \times 256$
masked tokens	foreground masking patch index = p_m	$768 \times 4 \times N_{p_m}$
PAM	$MLP(768, 384)$	$384 \times 4 \times N_{p_m}$
visible tokens	remaining patches patch index = $P - p_m$	$768 \times 4 \times (256 - N_{p_m})$
Transformer encoder	$\begin{bmatrix} MHA(768) \\ MLP(3072) \end{bmatrix} \times 12$	$768 \times 4 \times (256 - N_{p_m})$
encoder projector	$MLP(384)$	$384 \times 4 \times (256 - N_{p_m})$
concatenation	merge all tokens	$384 \times 4 \times 256$
Transformer decoder	$\begin{bmatrix} MHA(384) \\ MLP(1538) \end{bmatrix} \times 4$	$384 \times 4 \times 256$
decoder projector	$MLP(1536)$	$1536 \times 4 \times 256$
output video	reshape 1536 to $3 \times 2 \times 16 \times 16$	$3 \times 8 \times (256 \times 256)$

Table 2: Our end-to-end training setting.

Config	Value
optimizer	AdamW [5]
base learning rate	0.001
weight decay	0.05
optimizer momentum	0.9
learning rate schedule	cosine decay [4]
batch size	32
training epochs	100
warmup epochs	10
augmentation	rotation, flipping [6]

adjacent frames by comparing the RGB values of the same pixel, as shown in Figure 1, which demonstrates the temporal consistency. It is evident that the pixels in the same position across adjacent frames in Huang *et al.*'s results display significant color variation, indicating inconsistent harmonization across frames. This inconsistency may lead to flickering artifacts in the video, compromising the overall smoothness. Similarly, CO₂Net also shows inconsistencies between neighboring frames, while our method consistently maintains temporal coherence, improving the overall visual continuity and video quality.

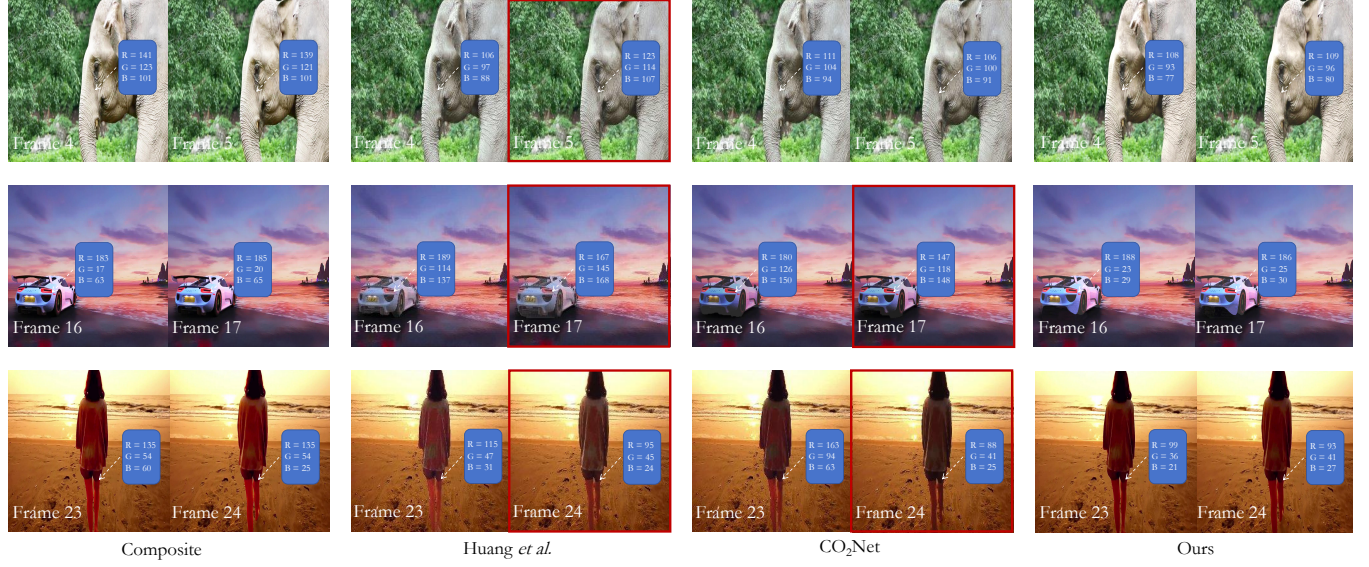


Figure 1: Qualitative result of comparison between Ours, Huang *et al.* and CO₂Net on RCVH dataset. The frames with red borders are not consistent with their neighboring frames, which may cause flickering artifacts. We also show the RGB values of temporally identical pixels in two adjacent frames.

Table 3: Comparison results of different value of L_{align} and $L_{balance}$. The best results are in bold and the worst results are in *italics*.

$\alpha(L_{align})$	$\beta(L_{balance})$	MSE ↓	fMSE ↓	PSNR ↑	fSSIM ↑
0.0	0.0	26.39	188.01	36.53	0.8821
0.3	0.3	26.13	185.23	36.61	0.8823
0.5	0.3	26.07	183.16	36.89	0.8827
0.7	0.3	26.10	179.04	36.73	0.8822
0.3	0.5	26.09	184.48	36.71	0.8824
0.5	0.5	25.94	182.93	36.96	0.8824
0.7	0.5	25.88	180.42	37.26	0.8826
0.3	0.7	26.02	184.19	36.99	0.8823
0.5	0.7	25.76	181.83	36.92	0.8828
0.7	0.7	25.64	175.92	37.41	0.8830
1.0	1.0	25.47	173.65	37.59	0.8832

3.2 Large-scale Results

To highlight our method’s effectiveness in handling large-scale inharmonious foregrounds, we conducted extensive visual comparisons using the RCVH dataset. As illustrated in Figure 2, our results showcase superior color recovery and overall coherence. Examining the first two rows of Figure 2, the input composite video features a cool-toned foreground against a warm-toned background. In contrast, the results from Huang *et al.* [2] retain a cool tone in the harmonized foreground, making it starkly different from the background. Meanwhile, CO₂Net’s [6] results, though closer in tone between the foreground and background, show a lack of uniformity in the foreground’s color, likely due to inadequate background information, resulting in an overall lack of smoothness. In contrast,

our method provides a more authentic color representation and a naturally integrated appearance.

4 DATASET VISUALIZATION

To supplement our proposed dataset, RCVH, in Section 3.4 of the manuscript, we employ some example pairs of composite and corresponding real video samples from the RCVH dataset. As shown in Figure 3, the RCVH dataset comprises entirely real composite videos, where the foregrounds and backgrounds of each video are sourced from different original videos and manually combined to create new composite videos. Given that the foregrounds and backgrounds originate from diverse sources, they inherently display variations in semantic context and photometric information, presenting a greater challenge than synthetic video data created simply by altering foreground colors.

REFERENCES

- [1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. 2021. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*. 6836–6846.
- [2] Hao-Zhi Huang, Sen-Zhe Xu, Jun-Xiong Cai, Wei Liu, and Shi-Min Hu. 2019. Temporally coherent video harmonization using adversarial networks. *IEEE Transactions on Image Processing* 29 (2019), 214–224.
- [3] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. 2022. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3202–3211.
- [4] Ilya Loshchilov and Frank Hutter. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983* (2016).
- [5] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=BkksMN0cF7>
- [6] Xinyuan Lu, Shengyuan Huang, Li Niu, Wenyan Cong, and Liqing Zhang. 2022. Deep video harmonization with color mapping consistency. *IJCAI* (2022).
- [7] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems* 35 (2022), 10078–10093.



Figure 2: Harmonious results of comparing with the large-scale foreground on RCVH dataset, our results are more superior.

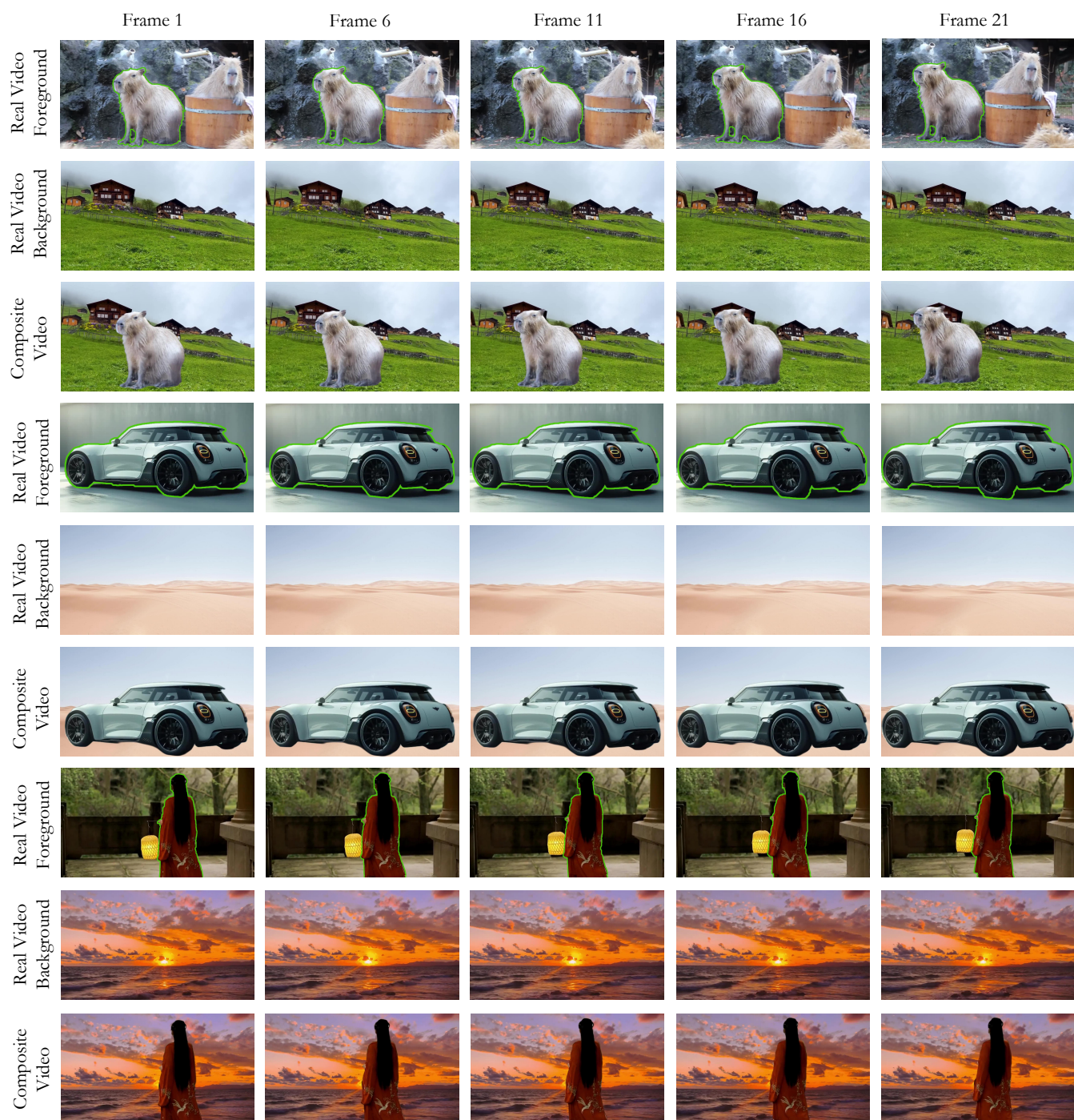


Figure 3: Some example pairs of composite video samples and their corresponding real video samples. The foregrounds are highlighted with green outlines.