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

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

This supplementary document is organized as follows:
Section 1 provides detailed structures of the feature encoder and the feature decoder.
Section 2 provides detailed structure of the TSAgg module.
Section 3 analyzes different divided temporal segments in TSA^2.
Section 4 provides more visual comparison results.
Section 5 provides more analysis of the TSAda module.

1. Details of the Feature Encoder and the Feature Decoder

Here we show the detailed structure of the feature encoder and the feature decoder in Figure 1.

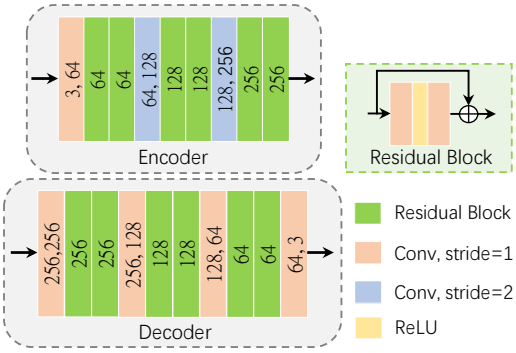


Figure 1. The feature encoder and the feature decoder. The numbers indicate the number of channels used in the TSA^2.

*Corresponding author.

2. Details of the TSAgg Module

Here we show the detailed structure of the TSAgg module.

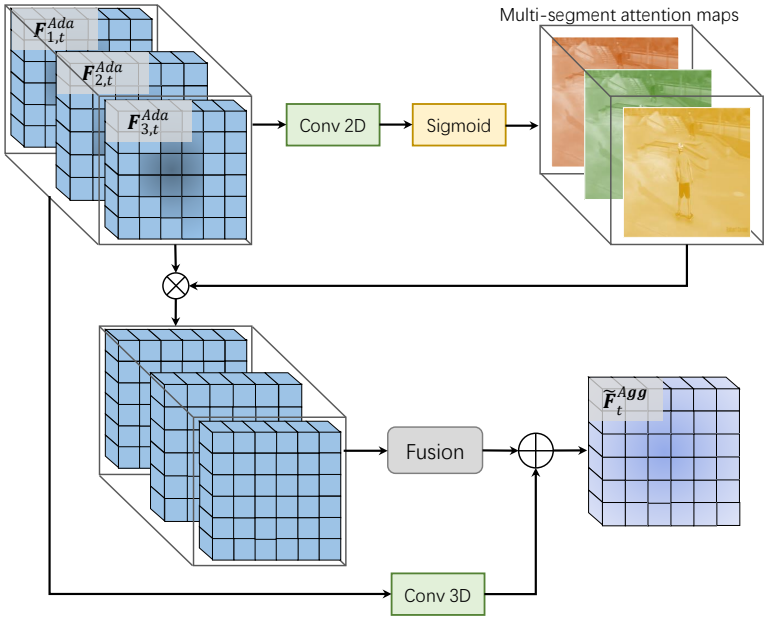


Figure 2. Structure of the TSAgg module.

3. Different Divided Temporal Segments

Apart from the temporal segments $\{\{I_3^C, I_4^C, I_5^C\}, \{I_2^C, I_4^C, I_6^C\}, \{I_1^C, I_4^C, I_7^C\}\}$ divided in our TSA^2, we explore four typical ways to divide 7-frame inputs into different temporal segments as shown in Table 1. Temporal segments in Method-(1) achieve the lowest fPSNR result, demonstrating the advantage of adding the central frame in each temporal segment. Temporal segments in Method-(4) achieve better results than the segments in Method-(2) and Method-(3) because the temporal segments in Method-(4) have a longer temporal context, which is crucial for the video harmonization task when transfer the background region information to the foreground region information.

Table 1. Analysis on different divided segments in the TSAda module.

Method	Temporal segments	fPSNR	fSSIM
(1)	$\{\{I_1^C, I_2^C, I_3^C\}, \{I_3^C, I_4^C, I_5^C\}, \{I_5^C, I_6^C, I_7^C\}\}$	25.99	0.8040
(2)	$\{\{I_1^C, I_2^C, I_4^C\}, \{I_3^C, I_4^C, I_5^C\}, \{I_4^C, I_6^C, I_7^C\}\}$	26.01	0.8040
(3)	$\{\{I_1^C, I_3^C, I_4^C\}, \{I_4^C, I_5^C, I_6^C\}, \{I_2^C, I_4^C, I_6^C\}\}$	26.05	0.8042
(4)	$\{\{I_2^C, I_3^C, I_4^C\}, \{I_4^C, I_5^C, I_6^C\}, \{I_1^C, I_4^C, I_7^C\}\}$	26.06	0.8044
Ours	$\{\{I_3^C, I_4^C, I_5^C\}, \{I_2^C, I_4^C, I_6^C\}, \{I_1^C, I_4^C, I_7^C\}\}$	26.07	0.8044

4. More Experimental Results

Here we show more visual comparisons between our TSA^2 and other methods in Figure 3.

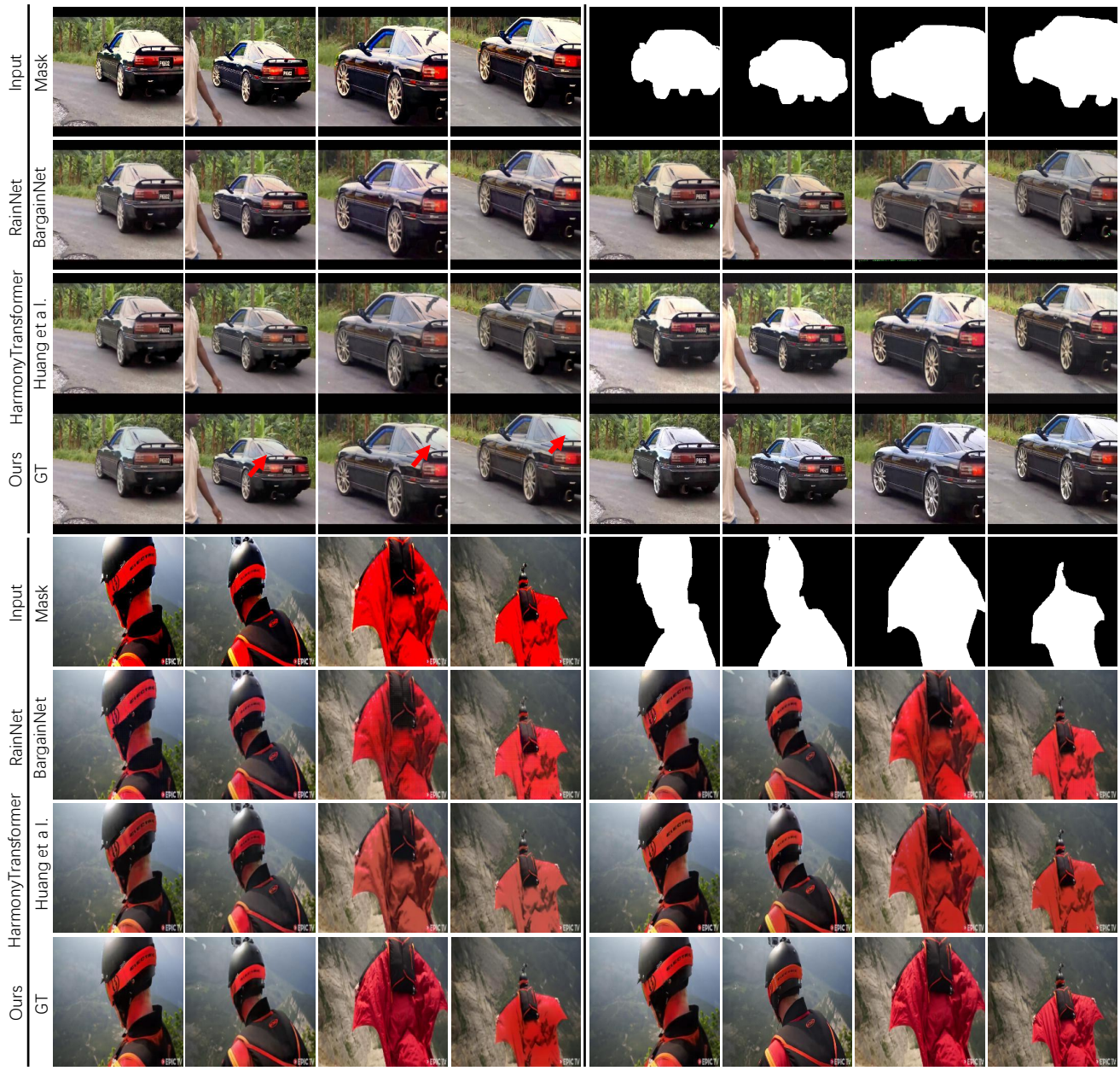
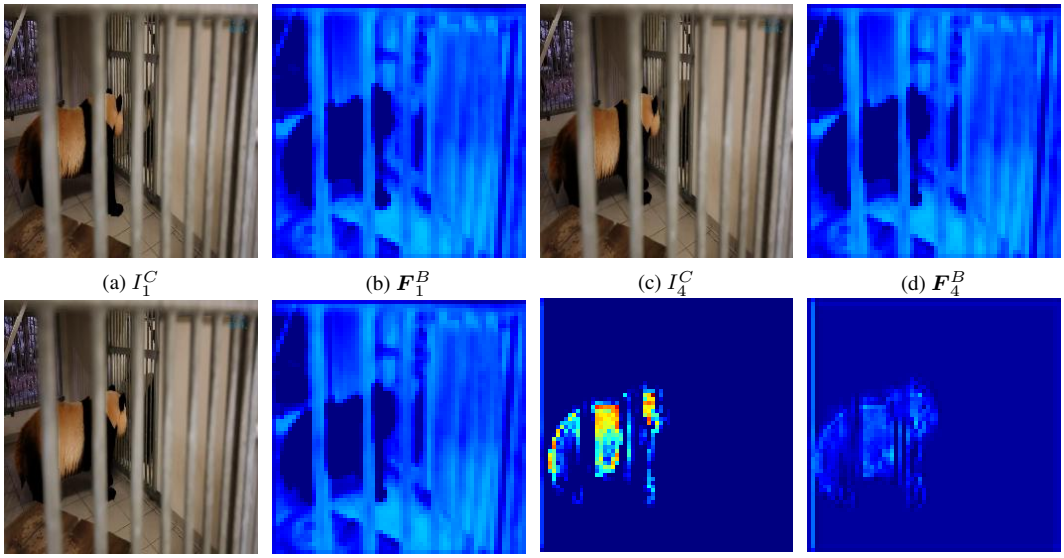


Figure 3. Visual comparisons of different representative methods on video frames from the HYouTube dataset. Red arrows indicate where our TSA^2 better harmonizes foreground regions and produces realistic appearance adjustments. Please zoom in for better visualization.

5. A Deeper Look at the TSAda Module

We visualize the intermediate features extracted from the temporal segment (*i.e.*, $\{I_1^C, I_4^C, I_7^C\}$). We show the background region features (*i.e.*, $\boldsymbol{F}_1^B$, $\boldsymbol{F}_4^B$ and $\boldsymbol{F}_7^B$), the foreground region feature (*i.e.*, $\boldsymbol{F}_{3,4}^F$) and the output feature from the TSAda module (*i.e.*, $\boldsymbol{F}_{3,4}^{F,Ada}$) in Figure 4. As shown in Figure 4(b), (d) and (f), features of harmonious background regions share similar feature representations, and the feature of the inharmonious region in the central frame (see Figure 4(g)) has significantly different representation (*i.e.*, the response values in the feature space of the extracted intermediate features). After we utilize the TSAda module to transfer the background information to the inharmonious foreground region, we find that the adjusted foreground and harmonious background regions have similar representation, indicating the effectiveness of this module.



(a) I_1^C (b) $\boldsymbol{F}_1^B$ (c) I_4^C (d) $\boldsymbol{F}_4^B$

(e) I_7^C (f) $\boldsymbol{F}_7^B$ (g) $\boldsymbol{F}_{3,4}^F$ (h) $\boldsymbol{F}_{3,4}^{F,Ada}$

Figure 4. Effectiveness of the TSAda module. (a), (c) and (e) are inharmonious input frames, and (b), (d), (f) are extracted features of the harmonious background regions. (g) and (h) are selected feature maps before and after adjusting using the TSAda module.