

# Supplementary Materials: TVPR:Text-to-Video Person Retrieval and a New Benchmark

Anonymous Authors

## 1 FURTHER DISCUSSION ABOUT THE NUMBER OF INPUT FRAMES TO THE VISUAL ENCODER

In the paper, we show the test results of the proposed MFGF on TVPReid-Dataset, in which the visual encoder of MFGF only uses 4 discrete video frames as input. However, the number of input video frames will affect the quality of the video features extracted by the visual encoder. Here we will discuss the case where the visual encoder uses 1, 4, and 8 video frames as input respectively. And we show the test results in Table 1.

As can be seen from Table 1, using 8 frames as input can improve the accuracy of retrieval, because more video frames can provide more detailed features. Therefore, more frames can deepen the model’s understanding of the video. Especially when a person is continuously occluded, or he has more changes in movement and clothing, more video frames will provide more information to make up for the missing dynamic details in the video features. This is a positive improvement for retrieval accuracy, but it will also bring higher computational costs. As we mentioned in Section 3.2 of the submitted manuscript, we chose the fragments learning strategy precisely to reduce computational costs while also ensuring good retrieval performance. Therefore, we choose 4 frames as the input of the visual encoder, and as shown in the results in Table 1, when 4 frames are used as the input, MFGF still has powerful retrieval performance. But when we reduce the input of the visual encoder to 1 frame, the retrieval performance of MFGF drops significantly. This is because when we only use one video frame, the visual encoder will lose the ability to solve the occlusion problem. When the pedestrians in the extracted video frames are occluded, the visual encoder will not be able to obtain complete appearance features, so that the appearance information of the pedestrians in the video cannot match the appearance information provided in the text.

As shown in Figure 2, when a video frame is extracted from the video as a sample, the pedestrian in it is obscured by a yellow car, thus losing appearance information such as coat, pants, shoes, and backpack. However, the provided text contains a complete description of the pedestrian’s appearance, which causes failure of match between video and text. In another case, we extract 4 video frames as input. As can be seen in Figure 2, two of the 4 extracted video frames have occlusion problems. But by extracting the visual information provided by other video frames, we can still complete the pedestrian’s appearance features to match the text, which is something that a single video frame cannot do.

To sum up, our use of 4 video frames as the input of the visual encoder is a choice that comprehensively considers retrieval performance and computational cost.

## 2 DETAILED MODIFICATION OF SPACE-TIME ATTENTION BLOCK IN VISUAL ENCODER

The visual encoder uses an improved ViT model, in which the detailed structure of the space-time attention block is shown in Figure 1.

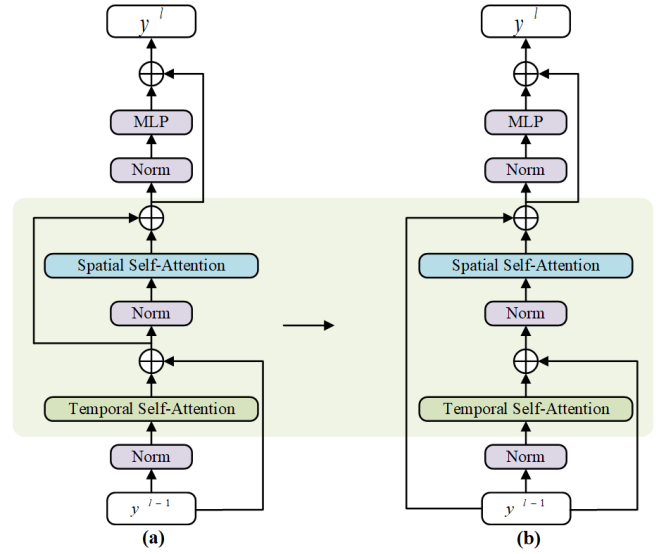


Figure 1: Detailed structure of the original space-time attention block (a), and minor modification of the residual connection between the temporal and spatial attention layers (b).

We employ divided space-time attention blocks and aggregate information from different attention layers within each block by using residual connections. Figure 1 (a) shows the original attention block in Timesformer [1], while Figure 1 (b) shows a minor modification of the residual connection between the temporal and spatial attention layers, this modification makes the training of the model faster and more stable.

## REFERENCES

- [1] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is space-time attention all you need for video understanding?. In *ICML*, Vol. 2. 4.

**Table 1: The retrieval results of MFGF on three sub-datasets use 1, 4 and 8 frames as input of the visual encoder respectively.**

Num of frames	TVPreid-PRID				TVPreid-iLIDs				TVPreid-Duke			
	R@1	R@5	R@10	R@50	R@1	R@5	R@10	R@50	R@1	R@5	R@10	R@50
1	23.3	60.0	78.3	88.3	24.7	64.7	85.0	90.0	28.7	51.5	62.0	84.0
4	32.3	76.3	85.1	100.0	30.0	71.7	91.7	100.0	35.2	62.2	84.0	90.4
8	35.5	80.7	89.9	100.0	34.3	78.6	96.0	100.0	39.0	66.7	85.3	95.9

**Figure 2: The occlusion problem in a single video frame will cause the failure of the matching between text and video, but multiple video frames can solve the occlusion problem.**