1. First Appendix

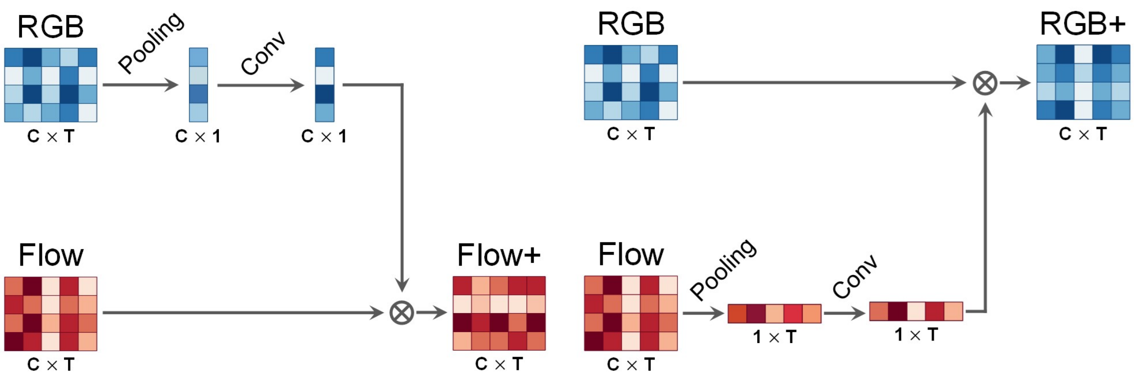

Figure 1: Diagram of the Dual-Modality Fusion (DMF), which consists of two parts: the left is RGB-generated channel attention for Flow enhancement, and the right is Flow-generated temporal attention for RGB enhancement.

1.1. Detail of Dual-Modality Fusion

When the DMF operation is stacked and used in subsequent fusion steps, it may lead to over-strengthening a small number of features and lead to over-fitting. Therefore, we set a hyper-parameter α to adjust the degree of DMF enhancement.

For the RGB and optical flow modalities of the video, we define two modality fusion strategies: *Concatenate Fusion* and *Mutual Fusion* (for simplicity, hereafter referred to as *concat* and *mutual*). *concat* can perform fusion between different modalities without information loss, and it has been used in most of the mainstream WSTAL works at present Huang et al. (2022); Yang et al. (2022); Hong et al. (2021); He et al. (2022). While *mutual* enhances the implicit information shared between modalities before fusion, performing specific tasks based on mutual fusion features can be regarded as double-checking between two modalities. This fusion strategy is referred to in a recent work Hong et al. (2021).

The *concat* fusion is denoted as:

$$concat = \mathcal{C}_{con}(cat(X_{rgb}, X_{flow})) \tag{1}$$

where $\mathcal{C}_{con}$ is a 1×1 convolution for information interaction between channels, and *cat* represents to concatenate X_{rgb} and X_{flow} along the channel dimension to achieve information fusion of the two modalities. And the *mutual* fusion is denoted as:

$$mutual = h_r(\mathcal{C}_r(X_{rgb})) + h_f(\mathcal{C}_f(X_{flow})) \tag{2}$$

where h_r and h_f denote to be the projectors for preserving the differences between modalities and avoiding their convergence to be completely the same.

For *mutual* fusion, we introduce contrastive learning for reinforcing the mutual information between modalities, i.e., to optimize the loss:

$$\mathcal{L}_{ml} = MSE(\mathcal{C}_r(X_{rgb}), \mathcal{C}_f(X_{flow})) \tag{3}$$

Combining the DMF operation and the fusion method, we produce various options for cross-modality fusion strategies. After experiments, our optimal fusion strategy is to use *mutual* fusion strategy with DMF feature enhancement as the input of T-RPN learning (Stage-1) and *concat* fusion without DMF enhancement as the input of action classification (Stage-2). We will make a detailed discussion of the reason why the heterogeneous feature fusion strategies in different stages of the Two-Stage Detection framework in Section 1.2.

Fusion Strategy	Stage-1				Stage-2				mAP@AVG
	RGB	Flow	Concat	Mutual	RGB	Flow	Concat	Mutual	
w/o Fusion	✓				✓				37.9
		✓				✓			41.4
w/ Fusion			✓				✓		44.4
				✓			✓		**45.5**
			✓					✓	42.6
				✓				✓	44.3

Table 1: Performance comparison of using different fusion strategies on different stages.

Stage-1		Stage-2		mAP@AVG
Mutual	Mutual+	Concat	Concat+	
✓		✓		45.5
	✓	✓		**47.0**
✓			✓	45.9
	✓		✓	46.9

Table 2: Performance comparison with and without Cross-Modality Attention Network. + indicates that the input features are enhanced by DMF.

1.2. Fusion Stragety

Table 1 and Table 2 show the effect of the fusion mechanism of RGB and optical flow on the performance of WSTAL. According to the tabular results, first, we can observe that the model performance is much lower purely using single-modal information than the scheme using cross-modal fusion, indicating that both RGB and optical flow have irreplaceable value for WSTAL tasks. Besides, the motion information in the optical flow modality is relatively more important than the appearance information in the RGB modality.

Another important finding is that for Stage-1, the *mutual* fusion strategy is always better than the *concat* fusion strategy, while Stage-2 is just the opposite. We speculate that the *mutual* fusion strategy can enhance the mutual information between modalities, which can be regarded as the implicit expression of the action to be detected in the two modalities. These mutual constraints between modalities make T-RPN screen out more stringent foreground fragments, that is, higher-quality temporal region proposals. The *concat* fusion strategy can be regarded as a union of two modalities, which provide sufficient semantic information for action classification and benefit the recognition of Stage-2.

References

Bo He, Xitong Yang, Le Kang, Zhiyu Cheng, Xin Zhou, and Abhinav Shrivastava. Asm-loc: action-aware segment modeling for weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13925–13935, 2022.

Fa-Ting Hong, Jia-Chang Feng, Dan Xu, Ying Shan, and Wei-Shi Zheng. Cross-modal consensus network for weakly supervised temporal action localization. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1591–1599, 2021.

Linjiang Huang, Liang Wang, and Hongsheng Li. Weakly supervised temporal action localization via representative snippet knowledge propagation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3272–3281, 2022.

Zichen Yang, Jie Qin, and Di Huang. Acgnet: Action complement graph network for weakly-supervised temporal action localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3090–3098, 2022.