Supplementary Material for “DDG-Net: Discriminability-Driven Graph Network for Weakly-supervised Temporal Action Localization”

1. Pre-classification Performance

As mentioned in this paper (Section 5.2), the performance of pre-classification for pseudo-action and pseudo-background snippets on ActivityNet1.2 is worse than THOMOS14 in Table 1, which leads to less improvement by our method on ActivityNet1.2.

dataset	$\mathrm{proportion}_{\mathrm{am}}$	precision	recall
THUMOS14	18.9%	85.1%	69.0%
ActivityNet1.2	14.9%	76.1%	64.8%

Table 1. Pre-classification results on THUMOS14 and ActivityNet1.2. Pre-classification denotes the classification of snippets in this paper (Section 4.2). $\mathrm{Proportion}_{\mathrm{am}}$ denotes the proportion of ambiguous snippets. Precision and recall are computed for both action and background snippets.

2. Extra Experiments

To demonstrate the powerful capabilities of our method, we design extra two sets of experiments where the complementary learning loss [1] is abandoned or the cross-modal consensus module [2] is applied. As shown in Table 2, the performance of three baselines is improved with DDG-Net consistently. Without the complementary learning loss, the increase is even more. As mentioned in this paper, both DDG-Net and the cross-modal consensus module are modules for feature enhancement. No matter whether combined with the cross-modal consensus module, our method brings significant improvements.

3. Quantitative Analysis

We make several analyses of the changes in attention weights before and after applying DDG-Net to demonstrate the function of our method.

Classification performance. We record the classification results of snippets on THOMOS14 dataset based on the pre-classification method (classify snippets as pseudo-action, pseudo-background, and ambiguous snippets) before and after applying DDG-Net. The number of ambiguous snippets decreases from 13431 to 10429, which means the snippet-level representations are more discriminative. We count the number of correct, wrong, and missed samples for classification results of action snippets and background snippets shown in Figure1, 2 respectively. We can find the number of missed snippets declines after applying DDG-Net, which indicates more snippets are regarded as discriminative. In Figure 3, we evaluate the classification results with different criteria. After applying DDG-Net, the precision decreases slightly but recall is raised obviously, which means more snippet-level representations are considered to be discriminative. And the improvement of the F1-score shows a better balance, which indicates better quality for discrimination.

Distribution. We statistic the distribution of the attention weights for pseudo-action, pseudo-background and ambiguous snippets separately. As shown in Figure 4, 5, the attention weights of pseudo-action and pseudo-background snippets tend to be more marginal, which explains stronger discriminability of snippet-level representations after applying DDG-Net. In Figure 6, the number of the attention weights of ambiguous snippets in the central regions is significantly reduced, which indicates the discriminability of ambiguous snippets is enhanced through DDG-Net. In summary, DDG-Net enhances the discriminability of snippet-level representations, especially ambiguous snippets.

References

[1] Jia-Run Du, Jia-Chang Feng, Kun-Yu Lin, Fa-Ting Hong, Xiao-Ming Wu, Zhongang Qi, Ying Shan, and Wei-Shi Zheng. Weakly-supervised temporal action localization by progressive complementary learning. *arXiv e-prints*, pages arXiv–2206, 2022.

[2] 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.

Method	mAP(%)@IoU							
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	Avg
baseline*	70.4	65.1	55.5	46.1	38.2	24.8	12.8	44.7
+DDG-Net	71.8$^{+1.4}$	66.7$^{+1.6}$	57.5$^{+2.0}$	48.4$^{+2.3}$	40.8$^{+2.6}$	**27.6**$^{+2.8}$	14.3$^{+1.5}$	46.7$^{+2.0}$
baseline	71.5	65.9	56.5	47.3	39.1	26.0	14.3	45.8
+DDG-Net	72.5$^{+1.0}$	**67.7**$^{+1.8}$	58.2$^{+1.7}$	49.0$^{+1.7}$	**41.4**$^{+2.3}$	**27.6**$^{+1.6}$	14.8$^{+0.5}$	47.3$^{+1.5}$
baseline†	72.2	66.4	56.9	47.9	39.6	26.5	13.7	46.2
+DDG-Net	**73.0**$^{+0.8}$	67.5$^{+1.1}$	**58.3**$^{+1.4}$	**49.1**$^{+1.2}$	40.6$^{+1.0}$	**27.6**$^{+1.1}$	**15.4**$^{+1.7}$	**47.4**$^{+1.2}$

Table 2. Comparison results on THUMOS14 dataset. * denotes the model without complementary learning loss. † denotes the model combined with the cross-modal consensus module.

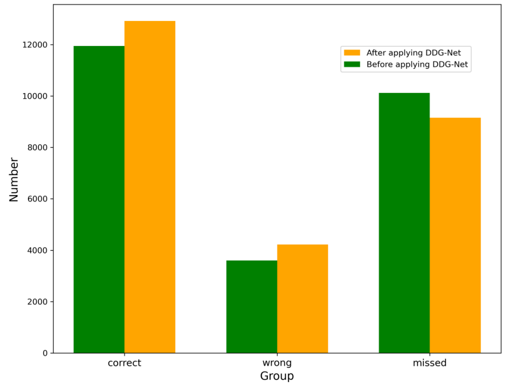

Figure 1. Classification results for action snippets. Correct samples denote the overlap between action (ground truth) and pseudo-action snippets. Wrong samples denote the overlap between background (ground truth) and pseudo-action snippets. Missed samples denote the action snippets which are not pseudo-action snippets.

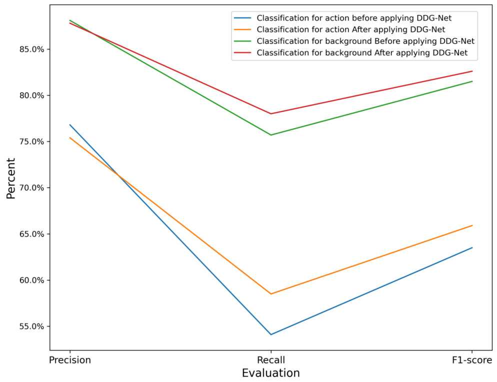

Figure 3. Evaluation of classification results with different criteria.

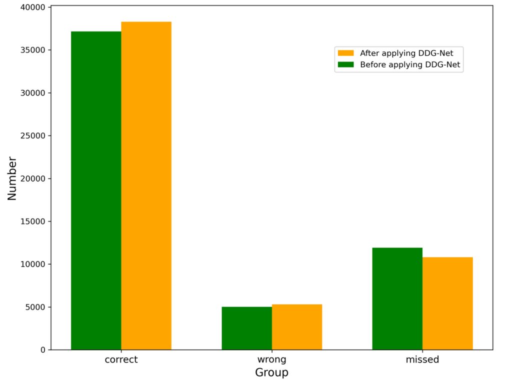

Figure 2. Classification results for background snippets. The definitions of correct, wrong, and missed samples are similar to Figure 1.

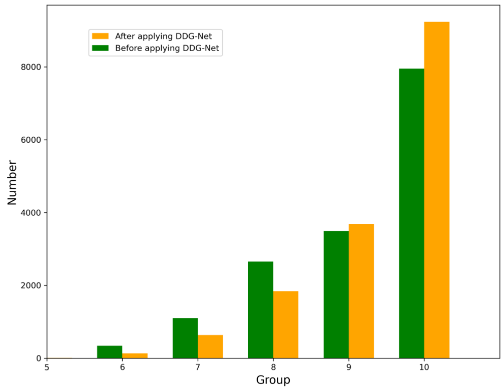

Figure 4. Distributions of attention weights of pseudo-action snippets. The 6th group denotes the region between 0.5 and 0.6, and Others are defined like this.

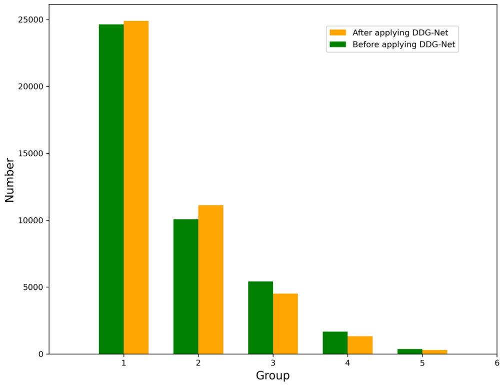

Figure 5. Distributions of attention weights of pseudo-background snippets. The 1st group denotes the region between 0.0 and 0.1, and Others are defined like this.

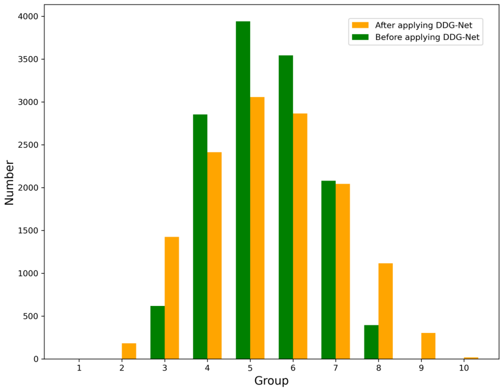

Figure 6. Distributions of attention weights of ambiguous snippets. The 1st group denotes the region between 0.0 and 0.1, and Others are defined like this.