Learning to Refactor Action and Co-occurrence Features for Temporal Action Localization
Supplementary Material

Method	mAP@tIoU (%)					
	0.1	0.2	0.3	0.4	0.5	Avg.
P-GCN	69.5	67.8	63.6	57.8	49.1	61.6
P-GCN*	71.2	69.0	63.7	58.1	49.0	62.2
Ours + P-GCN*	**73.7**	**71.1**	**65.3**	**59.8**	**50.6**	**64.1**

Table 1. Ablation studies of P-GCN on THUMOS14. "∗" indicates the reproduced results based on publicly available code.

1. Additional Implementation Details

Choice of coupling samples. High-quality coupling samples can encourage the encoder ϕ_{C} to effectively decouple co-occurring features of the action. In the experiment, we select five high-quality coupling samples with high cosine similarity for each action sample. They are 2048-dimensional feature vectors obtained by the two-stream network. It is worth mentioning that most of high-quality coupling samples come from video snippets in the neighborhood of action boundaries. Particularly, the action samples of the same class within each video often share coupling samples, which can help alleviate unbalanced data.

Network structure. Encoders ϕ_{A} and ϕ_{C} are implemented by three 1D temporal convolutional layers respectively and they have no weight sharing. The input dimension and output dimension of each encoder are 2048 and 1024, respectively.

2. Impact on Temporal Relation Modeling

Short- or long-term temporal relation modeling between actions and context is a promising strategy to refine imperfect action proposals. However, the overwhelming co-occurrence component often dominates subtle actions and causes inefficient temporal relation modeling. We take P-GCN [2] as an example to model the relationship between action proposals for action localization, and quantitatively analyze the effectiveness of our method. As shown in Table 1, comparison results between P-GCN* and Ours + P-GCN* demonstrate our method can provide more salient supportive information for temporal relation modeling and proposal refinement. This can also demonstrate our model is model-agnostic.

Method	AR@100	AUC
BMN	75.01	67.10
BMN*	75.40	67.40
Ours + BMN*	**76.54**	**68.11**

Table 2. Temporal proposal generation comparisons in terms of AR@AN (%) on ActivityNet v1.3. "∗" indicates the reproduced results based on publicly available code.

3. Impact on Temporal Action Proposal Generation

High-quality action proposals can cover action instances with high recall and high temporal overlap. If a TAL model overly relies on the co-occurrence component, it will confuse action boundary locations. To verify that our method can help generate high-quality action proposals with high recall, we take BMN [1] as an example for in-depth analysis. Particularly, we adopt average recall (AR) under different average numbers of proposals (AN) and the area under the AR vs. AN curve (AUC) as evaluation metrics on ActivityNet v1.3. Table 2 demonstrates that BMN combined with our method can achieve more accurate action boundary detection and improve the average recall.

References

[1] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. In *ICCV*, pages 3889–3898, 2019. 1

[2] Runhao Zeng, Wenbing Huang, Mingkui Tan, Yu Rong, Peilin Zhao, Junzhou Huang, and Chuang Gan. Graph convolutional networks for temporal action localization. In *ICCV*, pages 7094–7103, 2019. 1