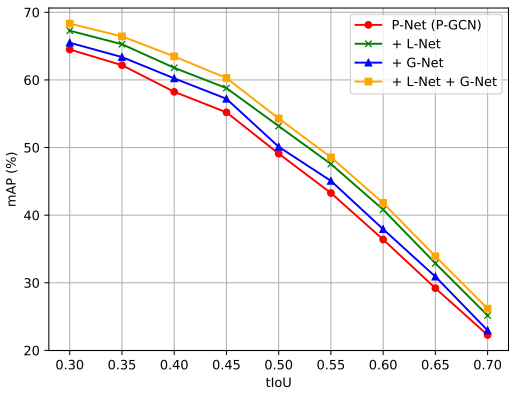
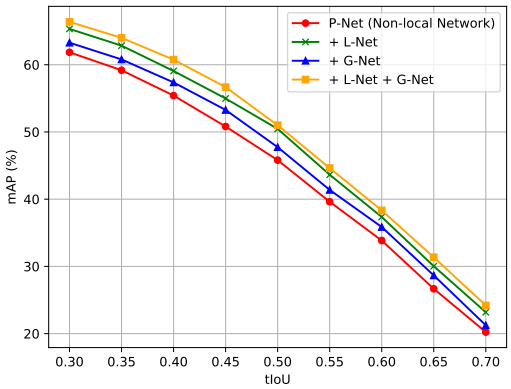
Supplementary Material for
Enriching Local and Global Contexts for Temporal Action Localization

1. More Details of Ablation Study



(a) P-GCN is used as P-Net.



(b) Non-local network is used as P-Net.

Figure 1. The curves show the effectiveness of L-Net and G-Net. P-Net is taken as a strong baseline. The mAPs (%) at different tIoU thresholds from 0.30 to 0.70 are reported on the THUMOS14 test set, and the step size is 0.05.

We take P-Net as a strong baseline to validate the effectiveness of L-Net (local context) and G-Net (global context), which are the main contribution of this paper. As shown in Figure 1, either the local context modeled via L-Net or the global context modeled via G-Net constantly improves the performance at different tIoU thresholds. L-Net, G-Net and P-Net are complementary, and their combination leads to the best performance. We can also see that the effectiveness of LNet, G-Net and ContextLoc does not depend on any specific instantiation of P-Net.

2. ContextLoc* on ActivityNet v1.3

Network. ContextLoc* is our augmented model on ActivityNet v1.3. We augment a new branch for the video-level classification, which predicts the action category of the video. The formula is

$$\boldsymbol{s}_v = \mathrm{FC}(\boldsymbol{z}), \tag{1}$$

where FC is a fully-connected (FC) layer, $\boldsymbol{z}$ is the video-level representation and $\boldsymbol{s}_v$ is the predicted video-level scores of each action category.

Using the network architecture described in Section 3.1 of the paper, we could obtain the classification scores $\boldsymbol{s}_p$ of a proposal and $\boldsymbol{s}_{ep}$ of the corresponding extended proposal. In addition, following the setting in P-GCN, we could obtain the classification scores $\boldsymbol{s}_{bsn}$ of the proposal via BSN. The final scores $\boldsymbol{s}_{fin}$ of each proposal are defined as

$$\boldsymbol{s}_{fin} = \boldsymbol{s}_p \times \boldsymbol{s}_{ep} \times \boldsymbol{s}_v \times \boldsymbol{s}_{bsn}, \tag{2}$$

where $\times$ denotes element-wise multiplication.

Loss Function. We train the video-level classification branch in Equation (1) via a set of binary cross-entropy losses. They are formulated as

$$\mathcal{L} = -\sum_{c=1}^{C} y^c \log\left(\sigma\left(s_v^c\right)\right) + \left(1 - y^c\right) \log\left(1 - \sigma\left(s_v^c\right)\right), \tag{3}$$

where σ is the sigmoid function, s_v^c is the cth element of $\boldsymbol{s}_v$, and $y^c \in \{0, 1\}$ denotes whether the cth action category occurs in the video, C is the number of action categories, respectively. We adopt the stochastic gradient descent (SGD) solver for optimization, and the initial learning rate is 0.0001.

3. Network Structures

We compare our ContextLoc model with deep and deeper P-GCNs in Table 4 of the paper to show that the

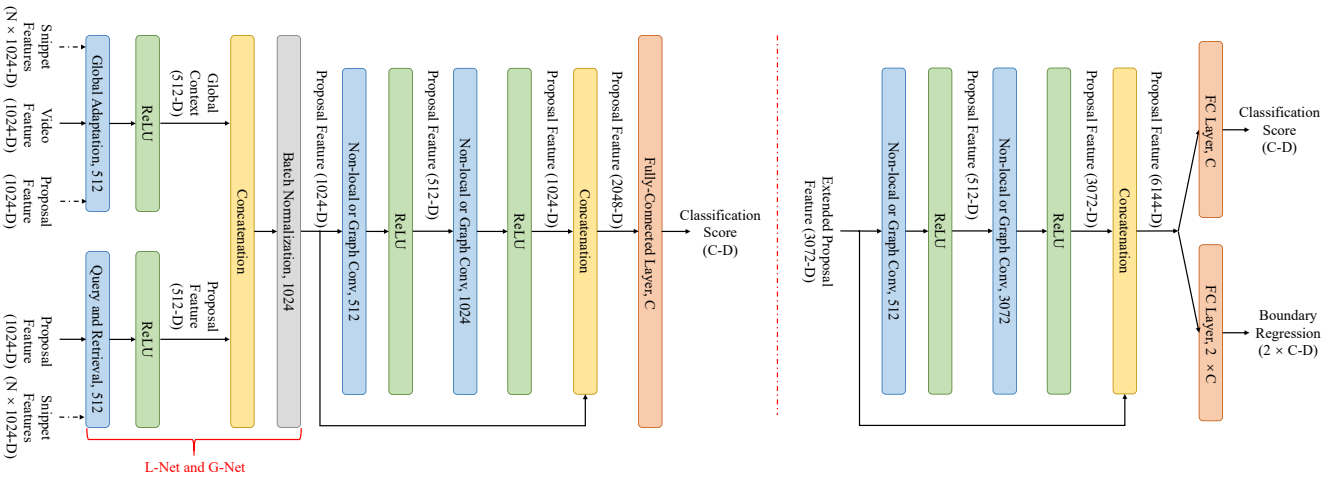


Figure 2. The network structure of our ContextLoc. The extended proposal features have already been processed by L-Net and G-Net, as described in the paper. C denotes the number of action categories.

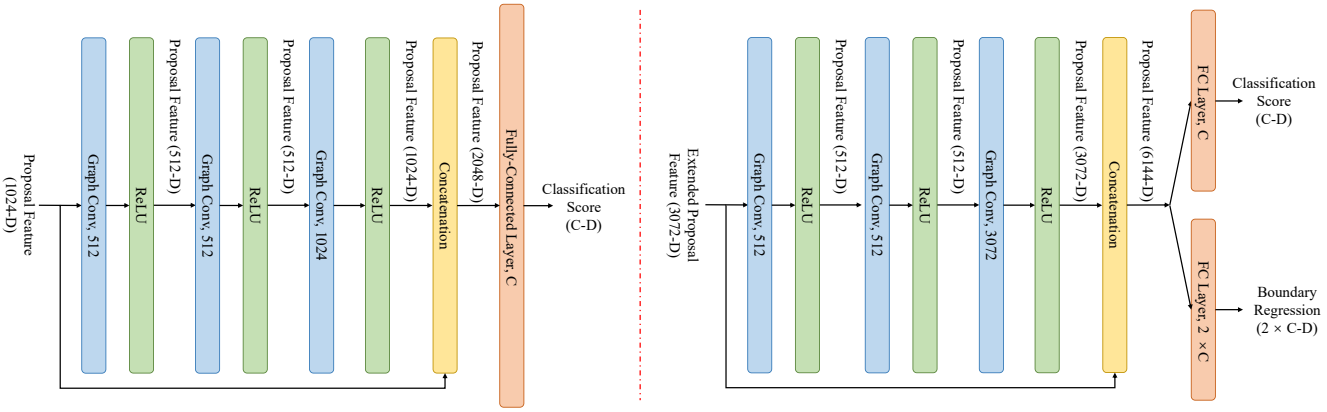


Figure 3. The network structure of the deep P-GCN. The extended proposal features are obtained by max pooling.

performance gain is not caused by increasing the network depth or the number of parameters. Here we draw their network structures.

ContextLoc. Our ContextLoc model is shown in Figure 2, where C is the number of action categories. We could change the feature dimension of the L-Net and G-Net from 512 to 256 to get the light model.

Deep P-GCN. The deep P-GCN model is shown in Fig-

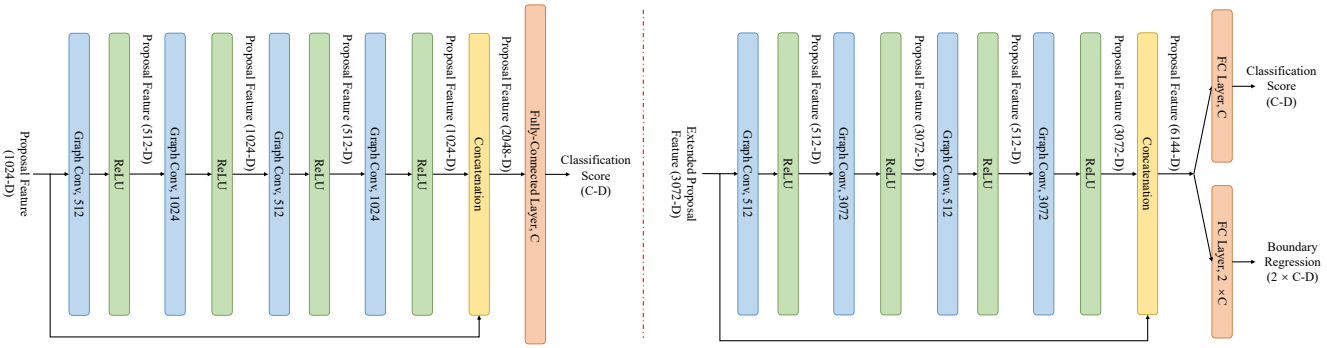


Figure 4. The network structure of the deeper P-GCN. The extended proposal features are obtained by max pooling.

Figure 5. Illustration of inaccurate annotations on ActivityNet v1.3. The durations of some annotations are longer than those of the true action instances, which causes the low performance at tIoU 0.95. The ground truth action instances and results obtained by ContextLoc are respectively illustrated using blue and orange bars. The inaccurate annotations are denoted using red fonts and green bars.

ure 3. In order to make its number of parameters similar as ours, we add one graph convolutional layer to P-GCN.

Deeper P-GCN. The deeper P-GCN model is shown in Figure 4. In order to make its number of flops similar as ours, we add two graph convolutional layers to P-GCN.

4. More Experimental Analysis

On ActivityNet v1.3, the performance of our network on the extremely high tIoU@0.95 is not as good as those of a few previous methods and only marginally outperforms that of P-GCN. After digging more deeply, we find that, in addition to the frame frequencies mentioned in Section 4.2 of the paper, another cause of the low performance on tIoU@0.95 is the inaccurate annotations on this dataset.

As shown in Figure 5, the durations of some annotations are longer than those provided in the annotations. Even if the prediction of ContextLoc can be very close to the ground truth action boundaries (tIoU reaches 0.95), the inaccurate annotation leads to a much lower tIoU at evaluation. In some other scenarios, different but nearby action instances are annotated as a continuous action instance. In sum, those inaccurate annotations make the evaluation at a high tIoU threshold very unreliable.

Method	RGB	Flow	Fusion	tIoU@0.5
P-Net (P-GCN)	22.37	22.92	26.99	42.90
+ L-Net	25.49	25.73	29.87	49.14
+ G-Net	24.47	24.55	28.10	46.82
+ L-Net + G-Net	**26.00**	**26.39**	**30.59**	**51.24**

Table 1. Ablation study on the ActivityNet v1.3 validation set.

5. Ablation study on ActivityNet v1.3

Table 1 reports the the ablation study on ActivityNet v1.3 and obtains the same conclusion on THUMOS14. When P-GCN is taken as P-Net, adding L-Net alone before it improves 2.88% on average mAP and 6.24% on tIoU 0.5. Adding G-Net alone before P-Net improves 1.11% on average mAP and 3.92% on tIoU 0.5. The complete ContextLoc (L-Net+G-Net+P-Net) further improves the performance.

6. The query-and-retrieval procedure.

We use the query-and-retrieval procedure in L-Net. In order to support its effectiveness, we replace the query-and-retrieval procedure with pooling and conduct the ablation study on THUMOS14 test set. Table 2 indicates the query-

Method	RGB	Flow	Fusion	tIoU@0.5
Avg	36.66	41.32	44.53	52.71
Max	35.69	41.67	44.04	51.95
Attention	**37.23**	**42.52**	**45.70**	**54.30**

Table 2. Ablation study on the effectiveness of the query-and-retrieval procedure (attention).

and-retrieval procedure outperforms pooling. This is because attention allows both L-Net and G-Net to dynamically retrieve relevant context and ignore irrelevant noise.