We thank the reviewers for their general positive feedbacks. We address concerns below.

To Reviewer 1: We have clarified the “number of layer” and “feature concatenation” issues in the text. The corresponding model adopted by the previous work is also added to the result table. Although it is true that the best performance is due in part to a more advanced model (e.g. Deep19), it is impressive that DAG-CNN could still achieve a considerable amount of gain given less room for improvement. This suggests that, under our architecture, both model complexity and multi-scale features are two parallel paths for improving performance. It is unclear if the performance gain for other approaches will match DAG-CNN when switching to an advanced model.

To Reviewer 2: We agree with the reviewer about evaluations on object recognition datasets. We are currently conducting experiments on large-scale object recognition tasks (ImageNet) and fine-grain classifications. We are in hope of including more results in the Camera-ready, if accepted.

To Reviewer 3: We have reported the results for various pooling and normalization strategies to combine intermediate layers in the text. The resulting table demonstrates that average pooling with L2 normalization performance the best.

We first fix the pool-window to be the entire spatial location and experiment on various pooling and normalization approaches.

|  |  |  |  |
| --- | --- | --- | --- |
|  | L2 norm | Min-max norm | Unit variance |
| Sum-pool |  |  |  |
| Average-pool |  |  |  |
| Max-pool |  |  |  |

We then use the best performing pooling and normalization approaches and vary the pooling-window size. Note the trade-off that the feature dimension increases as the pooling window becomes smaller.

|  |  |  |
| --- | --- | --- |
| Pooling-size | Feature dimension | Accuracy |
| 1x1 | x |  |
| 2x2 | 4x |  |
| 4x4 | 16x |  |
| 8x8 | 64x |  |