# Supplementary Material : Switching Convolutional Neural Network for Crowd Counting

Deepak Babu Sam* Shiv Surya* R. Venkatesh Babu
Indian Institute of Science
Bangalore, INDIA 560012
bsdeepak@grads.cds.iisc.ac.in, shiv.surya314@gmail.com, venky@cds.iisc.ac.in

## 1. Analysis of Switch-CNN

### 1.1. Effect of Coupled Training

Differential training on the CNN regressors $R_1$ through $R_3$ generates a multichotomy that minimizes the predicted count by choosing the best regressor for a given crowd scene patch. However, the trained switch is not ideal and the manifold separating the space of patches is complex to learn (see Section 5.2 of the main paper). To mitigate the effect of switch inaccuracy and inherent complexity of task, we perform coupled training of switch and CNN regressors. We ablate the effect of coupled training by training the switch classifier in a stand-alone fashion. For training the switch in a stand-alone fashion, the labels from differential training are held fixed throughout the switch classifier training.

The results of the ablation are reported in Table 1. We see that training the switch classifier in a stand-alone fashion results in a deterioration of Switch-CNN crowd counting performance. While Switch-CNN with the switch trained in a stand-alone manner performs better than MCNN, it performs significantly worse than Switch-CNN with coupled training. This is reflected in the 13 point higher count MAE. Coupled training allows the patch labels to change in order to adapt to the ability of the switch classifier to relay a patch to the optimal regressor $R_k$ correctly. This co-adaption is absent when training switch alone leading to deterioration of crowd counting performance.

| Method | MAE |
|---|---|
| MCNN [5] | 110.2 |
| Switch-CNN without Coupled Training | 103.26 |
| **Switch-CNN with Coupled Training** | **90.41** |

Table 1. Comparison of MAE for Switch-CNN trained with and without *Coupled Training* on Part A of the ShanghaiTech dataset [5].

*Equal contribution

### 1.2. Ablations on UCF_CC_50 dataset

We perform ablations referenced in Section 5.1 and 5.3 of the main paper on the UCF_CC_50 dataset [3]. The results of these ablations are tabulated in Table 2. The results follow the trend on ShanghaiTech dataset and reinforce the superiority of Switch-CNN (See Section 5.1 and 5.3 of the main paper for more details).

| Method | MAE |
|---|---|
| Cluster by count | 319.16 |
| Cluster by mean inter-head distance | 358.78 |
| Switch-CNN($R_1$,$R_3$) | 369.58 |
| Switch-CNN($R_1$,$R_2$) | 362.22 |
| Switch-CNN($R_3$,$R_2$) | 334.66 |
| **Switch-CNN** | **318.07** |

Table 2. Additional results for ablations referenced in Section 5.1 and 5.3 of the main paper for UCF_CC_50 dataset[3].

### 1.3. Choice of Switch Classifier

The switch classifier is used to infer the multichotomy of crowd patches learnt from differential training. The accuracy of the predicted count in Switch-CNN is critically dependent on the choice of the switch classifier. We repurpose different classifier architectures, from shallow CNN classifiers to state-of-the art object classifiers to choose the best classifier that strikes a balance between classification accuracy and computational complexity.

Figure 1 shows the different architectures of switch classifier that we evaluate. CNN-small is a shallow classifier derived from VGG-16 [4]. We retain the first three convolutional layers from VGG-16 and add a 512 dimensional fully-connected layer along with a 3-way classifier. The convolutional layers in CNN-small are initialized from VGG-16. We also repurpose VGG-16 and VGG-19 [4] by global average pooling the Conv 5 features and using a 512 dimensional fully-connected layer along with a 3-way classifier. All the convolutional layers in VGG-16 and VGG-19 are initialized from VGG models trained on Imagenet [1].

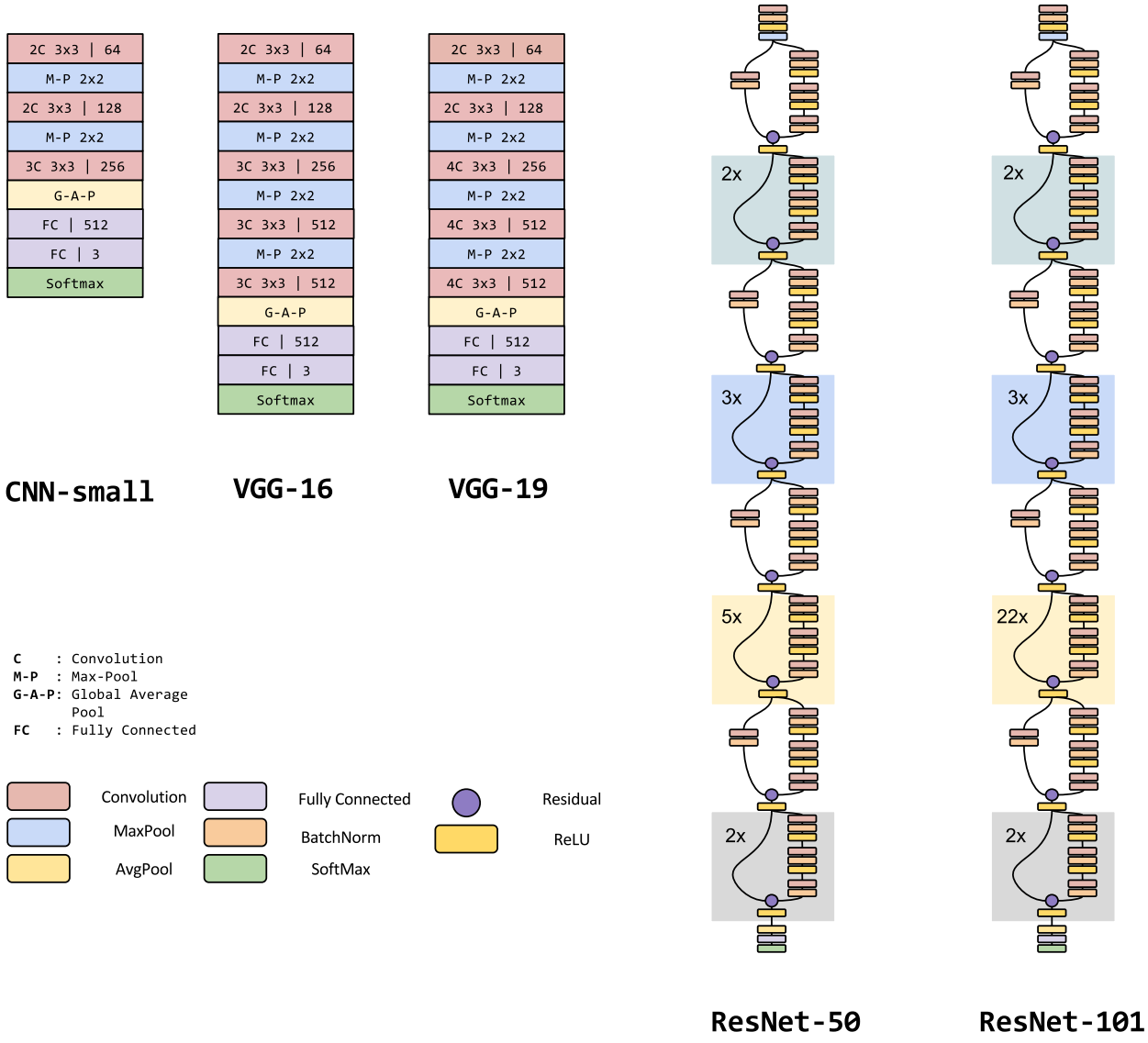


Figure 1. The architecture of different switch classifiers evaluated in **Switch-CNN**.

| Method | Acc |
|---|---|
| CNN-small | 64.39 |
| VGG-16 | 73.75 |
| VGG-19 | 74.3 |
| ResNet-50 | 75.03 |
| ResNet-101 | 74.95 |

Table 3. Comparison of classification accuracy for different switch architectures on Part A of the ShanghaiTech dataset [5]. The final switch-classifier selected for all Switch-CNN experiments is highlighted in red.

The state-of-the-art object recognition classifiers, Resnet-50 and Resnet-101 [2] are also evaluated. We replace the final 1000-way classifier layer with a 3-way classifier. For ResNet training, we do not update the Batch Normalization (BN) layers. The BN statistics from ResNet model trained for ILSCVRC challenge [1] are retained during fine-tuning for crowd-counting. The BN layers behave as a linear activation function with constant scaling and offset. We do not update the BN layers as we use a batch size of 1 during SGD and the BN parameter update becomes noisy.

We train each of the classifier on image patch-label pairs, with labels generated from the differential training stage (see Section 3.3 of the main paper). The classifiers are trained using SGD in a stand-alone manner similar to Section 1.1. Table 3 shows the performance of the different switch classifiers on Part A of the ShanghaiTech dataset [5]. CNN-small shows a 10% drop in classification accuracy over the other classifiers as it is unable to model the com-

plex multichotomy inferred from differential training. We observe that the performance plateaus for the other classifiers despite using more powerful classifiers like ResNet. This can be attributed to complexity of manifold inferred from differential training. Hence, we choose the repurposed VGG-16 model for all our Switch-CNN experiments as it gives classification accuracy competitive with deeper models like ResNet, but with a lower computational cost. A lower computational cost is critical as it allows faster training during coupled training of the switch-classifier and CNN regressors $R_{1-3}$.

## References


[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 1.3

[2] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. 1.3

[3] H. Idrees, I. Saleemi, C. Seibert, and M. Shah. Multi-source multi-scale counting in extremely dense crowd images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2547–2554, 2013. 1.2, 2

[4] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1.3

[5] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma. Single-image crowd counting via multi-column convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 589–597, 2016. 1.1, 1, 3, 1.3