Supplementary Materials for Neural Architecture Search for Joint Human Parsing and Pose Estimation

Dan Zeng[1] Yuhang Huang[1] Qian Bao[2] Junjie Zhang[1] Chi Su[3] Wu Liu[2]
[1]Shanghai University [2]JD AI Research [3]Kingsoft Cloud
dzeng@shu.edu.cn, huangyuhang@shu.edu.cn, baoqian@jd.com
junjie_zhang@shu.edu.cn, suchi@kingsoft.com, liuwu1@jd.com

1. Loss Function

We use the Mean Square Error as the loss function for pose branch while the Cross-Entropy loss is employed for parsing branch. Therefore, the loss function of each branch can be calculated by:

$$\begin{aligned} L_{pos} &= L_{MSE}(P_{pos}, G_{pos}) + L_{MSE}(P_{pos\_aux}, G_{pos\_aux}) \\ L_{par} &= L_{CE}(P_{par}, G_{par}) + L_{MSE}(P_{edg}, G_{edg}) \end{aligned} \tag{1}$$

where P denotes prediction and G denotes ground truth. Since the two loss values are quite different, referring to [1], we adopt the uncertainty loss to learn the weights for the two branches.

$$L = e^{-\sigma_1} \cdot L_{pos} + \sigma_1 + e^{-\sigma_2} \cdot L_{par} + \sigma_2 \tag{2}$$

where σ_1 and σ_2 are learnable parameters that balance the two tasks.

Since we use supervision in both two stages of forward, in order to distinguish between the two losses, we denote the losses as L_1 and L_2 respectively, so the total loss L_{total} can be formulated as:

$$L_{total} = L_1 + L_2 \tag{3}$$

2. Pruning Schemes

After searching, we combine two pruning schemes. First of all, since the search of encoder-decoder and feature fusion are both cell-based, we apply the pruning scheme in DARTS [3]. However, the search for the multi-scale feature interaction is not cell-based, and the connections are much denser. More valuable connections should be retained to ensure the efficiency of the network after pruning. Therefore, we design a new pruning scheme and its pseudo-code is in Algorithm 1.

Algorithm 1 Pruning algorithm used in multi-scale feature interaction search.

```
Input: The proportion of the operation $o(\cdot)$ in connection
    from node $j$ to node $i$, $\gamma^o_{i,j}$, $j \leq i$;
Output: The remaining operations list, $remain_o$;
Initialize $s = 0, n = 0, remain_o = [\,]$;
while $s \leq 0.7$ and $n \leq 4$ do
    $p = max(\gamma^o_{i,j})$;
    $o', i', j' = p.index()$;
    $remain_o$.append($o'$ from $j'$ to $i'$)
    $s = s + p, n = n + 1$
end while
return $remain_o$;
```

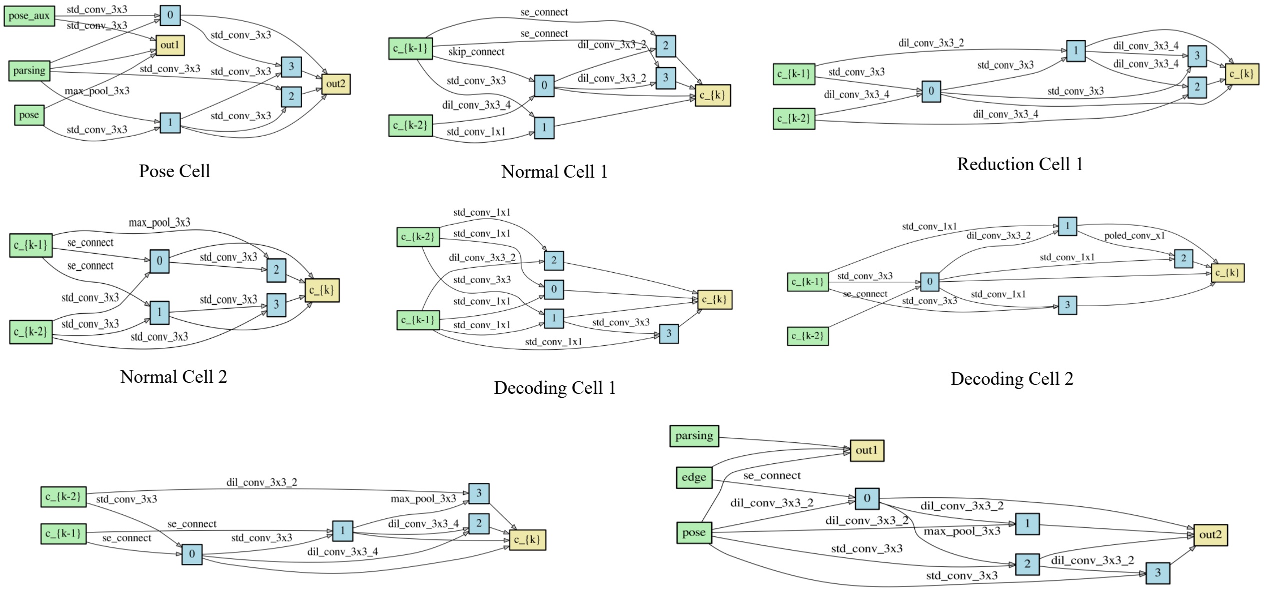
3. Architecture Detail

We design three search spaces, where encoder-decoder search and high-level feature fusion search are cell-based, and multi-scale feature interaction search is not. We show the searched architecture in Fig. 1 and Fig. 2 respectively. In Fig. 1, the Normal Cell 1, Reduction Cell 1, and Decoding Cell 1 are used in the pose branch while the Normal Cell 2, Reduction Cell 2, and Decoding Cell 2 are used in the parsing branch. The encoder-decoder of the pose branch has more dilated convolutions while the parsing branch prefers to use standard convolution. Moreover, the cells in the parsing branch are deeper. These all show there are differences between the pose branch and the parsing branch which suggests using different architecture to extract task-specific features for the two tasks. For feature fusion, the parsing feature is fused with both the pose feature and the pose auxiliary feature, which means the parsing branch assists the pose branch more. On the other hand, the pose feature is fused with the edge feature in the parsing cell indicates pose information may help enhance the boundary of each area.

Fig. 2 shows the connections of intermediate features between the two tasks. The pose branch absorbs more connections from the parsing branch while the parsing branch has fewer connections from another branch. This is consistent with our experiments that parsing information helps all types of key points while the promotion of human parsing is concentrated in a few special classes.

4. Computational cost analysis

To demonstrate the high efficiency of the proposed NPP-Net, we compare the computational cost with other state-of-the-art methods. As shown in Tab. 1, our model achieves the best performances on both tasks with a comparable compu-



Pose Cell

Normal Cell 1

Reduction Cell 1

Normal Cell 2

Decoding Cell 1

Decoding Cell 2

Reduction Cell 2

Parsing Cell

Figure 1. The searched cell architecture.

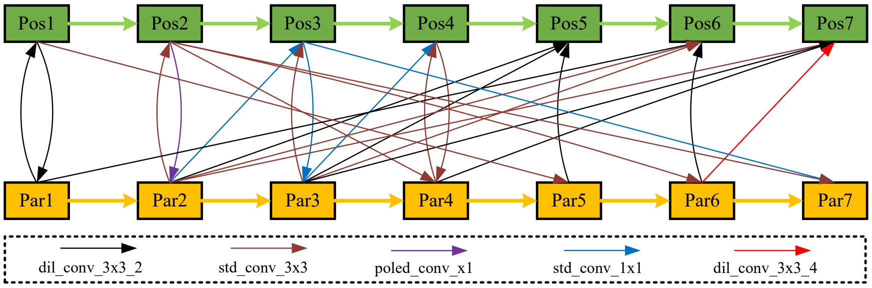


Figure 2. The multi-scale feature interaction architecture.

tational cost which shows the great potential of NPPNet to explore the interaction of the two branches. Note that the pose estimation networks are usually smaller than human parsing networks, and the MuLA [4] is mainly designed for pose estimation and its performance on human parsing is less satisfactory.

Table 1. Model size and computational cost on LIP dataset.

Method	Param	FLOPs	mIOU	PCK	Task
CNIF [6]	83.6M	300.0G	57.74	-	parsing only
HRNet [5]	60.6M	44.0G	-	88.0	pose only
JPPNet [2]	89.1M	263.5G	51.37	82.7	both tasks
MuLA [4]	44.4M	43.4G	49.30	87.5	both tasks
NPPNet(ours)	73.4M	113.7G	58.56	88.9	both tasks

References

[1] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *CVPR*, 2018. 1

[2] Xiaodan Liang, Ke Gong, Xiaohui Shen, and Liang Lin. Look into person: Joint body parsing & pose estimation network and a new benchmark. *IEEE transactions on pattern analysis and machine intelligence*, 2018. 2

[3] Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. In *ICLR*, 2019. 1

[4] Xuecheng Nie, Jiashi Feng, and Shuicheng Yan. Mutual learning to adapt for joint human parsing and pose estimation. In *ECCV*, 2018. 2

[5] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 2

[6] Wenguan Wang, Zhijie Zhang, Siyuan Qi, Jianbing Shen, Yanwei Pang, and Ling Shao. Learning compositional neural information fusion for human parsing. In *ICCV*, pages 5703–5713, 2019. 2