Appendix

Zhanzhou Feng$^1$ Shiliang Zhang$^{1,2}$

$^1$National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University
$^2$Peng Cheng Laboratory

fengzz@stu.pku.edu.cn, slzhang.jd1@pku.edu.cn

1. Masking strategy

The proposed method makes the masks evolve with the training process and combines the effects of grid-wise and part-wise masking by weighted adding the corresponding probability values. We elaborate method in Sec. 3. Further, this section provides a pseudocode implementation in Algorithm 1.

Algorithm 1 PyTorch pseudocode for masking strategy.

```python
def Mask(x, r, S, α):
    B, H, W, C = x.shape
    N = H * W
    # assign part-wise probability
    part_ini = Random(N)
    p_parts = torch.gather(part_noise Ini, index=S)
    # assign grid-wise probability
    grid_ini = Random(2, 2)
    p_grid = [grid_ini[(i//W)%2, i%2] for i in range(N)]
    # aggregate probabilities with a dynamic weight
    p = (1 - α) * p_grid + α * p_parts
    # decide mask or not according to the probability
    rank = torch.argsort(p)
    patch_location = torch.argsort(rank)
    ids_mask = rank[:N * r]
    x_masked = torch.gather(x, index=ids_mask)
    return x_masked, patch_location
```

Figure 1. ImageNet-1K [1] top-1 classification performance.

The random sequences part ini and grid ini correspond to the δ in the article, i.e., a set of random values to assign the probability P. For grid-wise masks, we assign identical probability values to the patches with the same relative position in the grid to ensure that these are preserved or masked simultaneously. For part-wise masks, we set the exact probability values for the patches with the same value in S (i.e., patches belonging to the same part). After that, we add the $p_{\text{parts}}$ and $p_{\text{grid}}$ together and mask out the top $[N \times r]$ patches with the highest probability values.

Figure 2. ADE20K [3] semantic segmentation mIoU curves.

2. Downstream performance

In this section we show the performance curves of the proposed method and three basic mask methods on downstream tasks, i.e. classification, segmentation and detection in 1, 2 and 3 respectively. It can be seen that the evolved method outperforms the static mask method in different pre-training epochs. Our methods can effectively converge with the limited training epochs, while with longer pre-training epochs, the method can further improve performance on various downstream tasks.

3. Compared with straightforward evolved baseline.

The partition is done with graph-cut in this paper. To validate the effectiveness of the proposed partition strategy, we replace our graph partition with random block mask-
ing as a baseline, \textit{i.e.}, the mask evolves from grid masking to block masking. It achieves 79.18\% \textit{(v.s. 79.89\% of ours)} ImageNet classification top-1 accuracy and 38.30\% \textit{(v.s. 40.42\% of ours)} ADE20K \cite{zhou2017} mIoU with 200 epochs pretraining.

References

\begin{itemize}
\item \cite{deng2009} Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In \textit{2009 IEEE conference on computer vision and pattern recognition}, pages 248–255. Ieee, 2009.
\end{itemize}