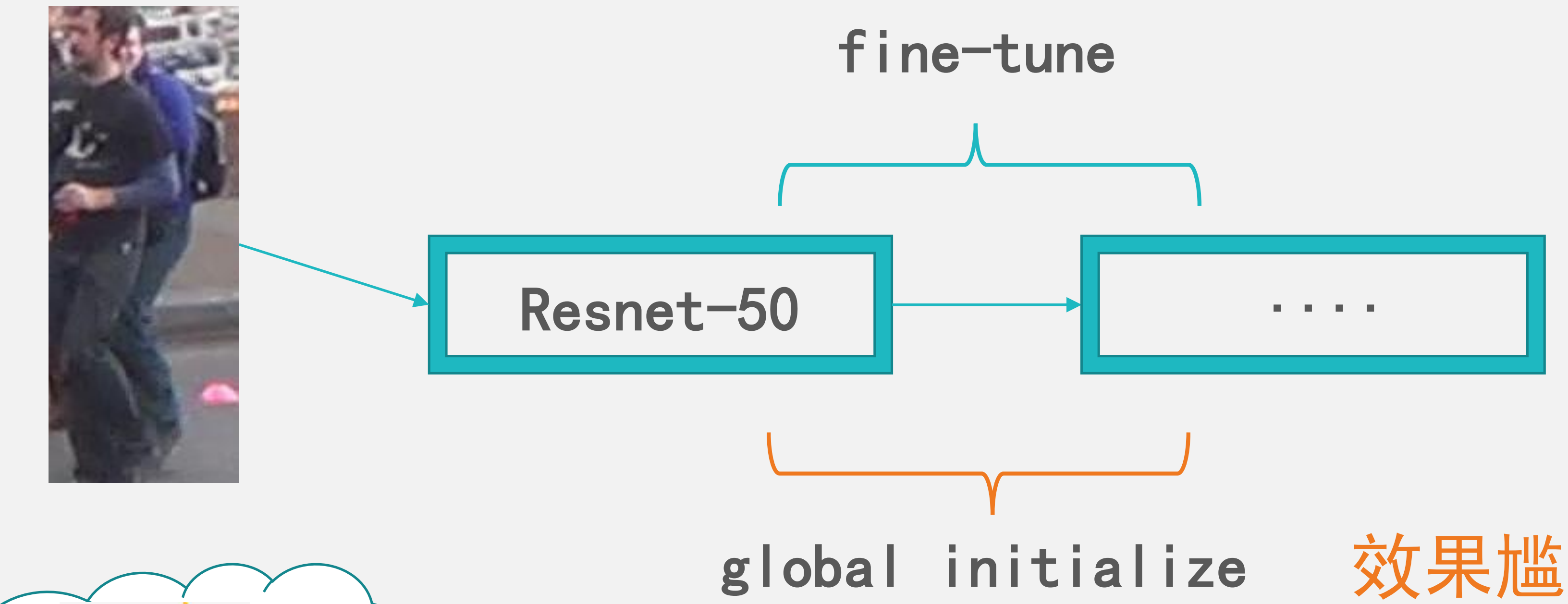
Deep Clustering for Unsupervised Learning of Visual Features

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze

Facebook AI Research

{mathilde,bojanowski,ajoulin,matthijs}@fb.com

引言（从尴尬的错误开始）



效果尴尬

1． 崩溃原因：网络深，需要好数据
2． fine-tune成功原因：ImageNet

深度聚类

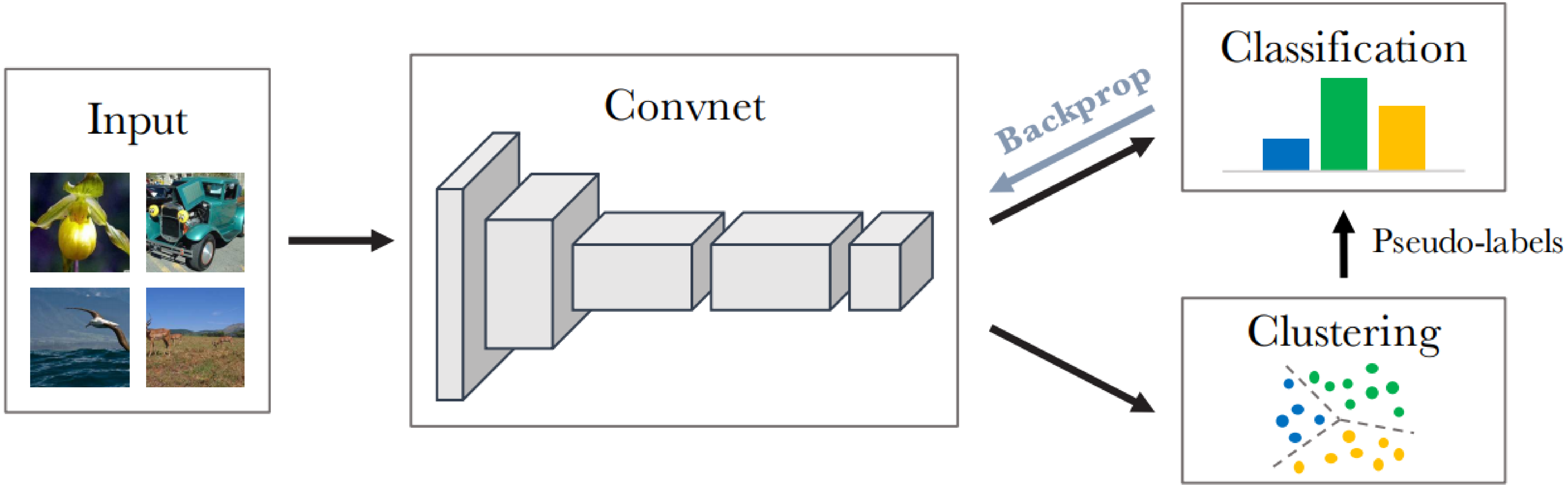


Fig. 1: Illustration of the proposed method: we iteratively cluster deep features and use the cluster assignments as pseudo-labels to learn the parameters of the convnet

理论基础是什么？

ImageNet1000类的分类问题

Random选 → 准确率为 1 / 1000, 0.1%

只初始化的AlexNet网络 → 准确率为12%

卷积的框架对于输入的信号还是有一些**先验的知识**。

较弱的先验知识能用来不断增强网络的特征学习能力

Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solvingjigsaw puzzles. In: ECCV (2016)

细节

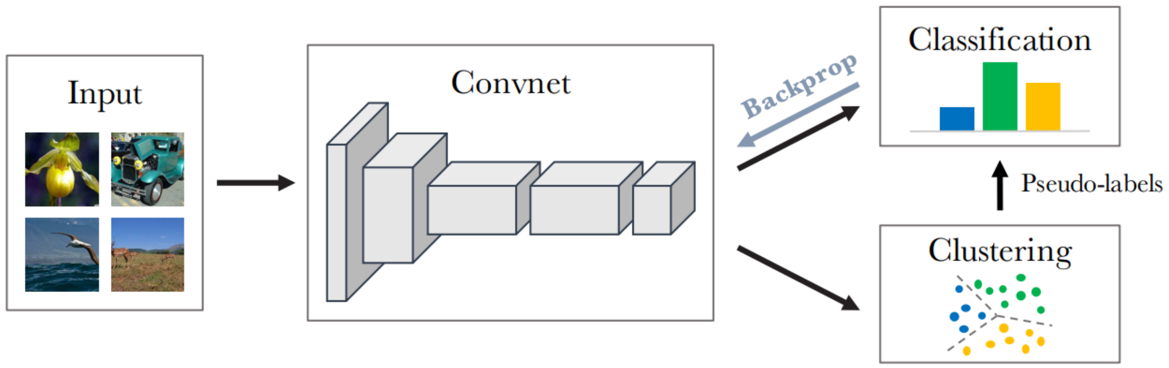


Fig. 1: Illustration of the proposed method: we iteratively cluster deep features and use the cluster assignments as pseudo-labels to learn the parameters of the convnet

$$\min_{\theta,W} \frac{1}{N}\sum_{n=1}^{N} \ell\left(g_W\left(f_\theta(x_n)\right), y_n\right)$$

$$\min_{C\in\mathbb{R}^{d\times k}} \frac{1}{N}\sum_{n=1}^{N} \min_{y_n\in\{0,1\}^k} \|f_\theta(x_n) - Cy_n\|_2^2 \quad \text{such that} \quad y_n^\top 1_k = 1.$$

但是只是这样优化不加约束的话，整个网络有可能**崩溃**。

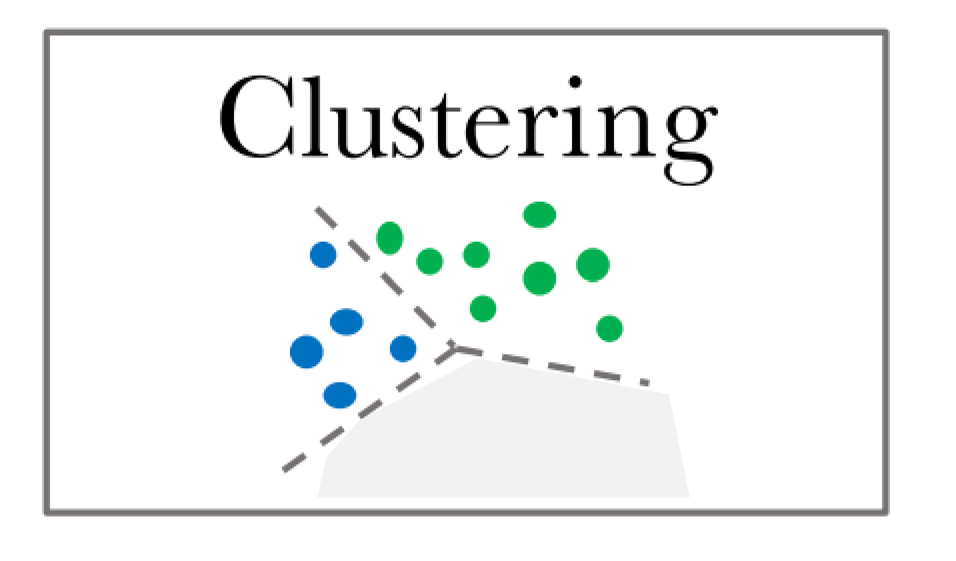
细节

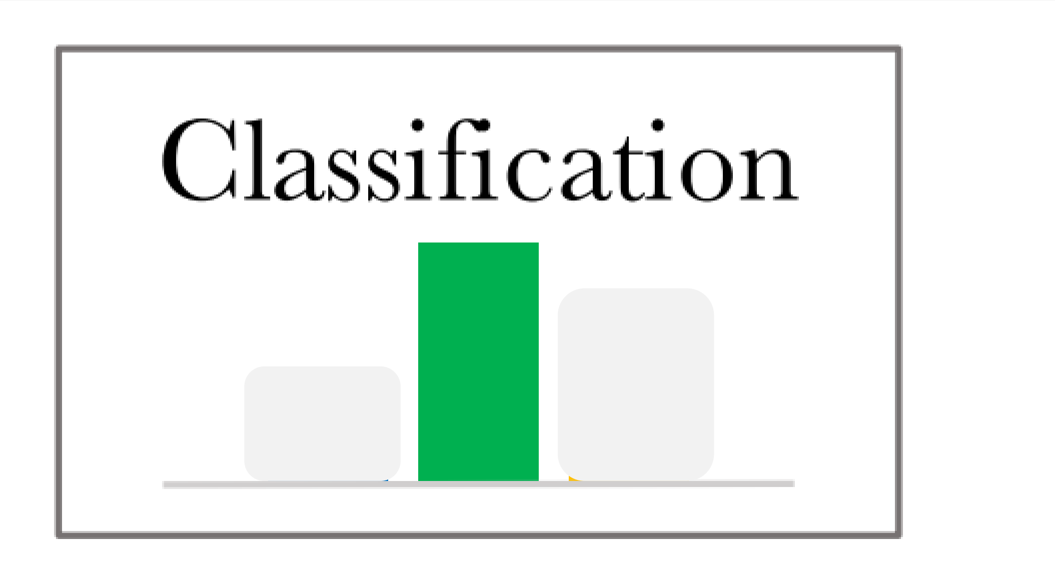
$$\min_{C\in\mathbb{R}^{d\times k}} \frac{1}{N}\sum_{n=1}^{N} \min_{y_n\in\{0,1\}^k} \|f_\theta(x_n) - C y_n\|_2^2 \quad \text{such that} \quad y_n^\top 1_k = 1.$$

$$\min_{\theta,W} \frac{1}{N}\sum_{n=1}^{N} \ell\left(g_W\left(f_\theta(x_n)\right), y_n\right)$$

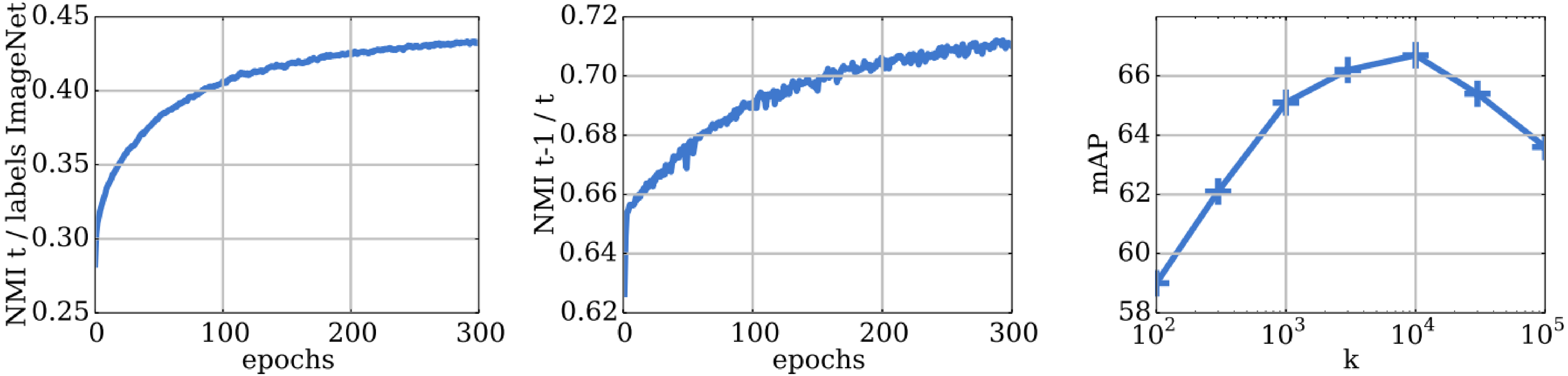
有可能特征为0，但是他们也被归为一簇，离聚类中心距离为0， 也是最小
或者某一簇为空，质心也是0， 距离也是最小。

数据不均匀，那有可能优化的时候，
直接一直输出大类，那这样的loss也是比较小了





作者就在ImageNet上无监督地训练，在一块ＧＰＵ跑了１２天。然后分析实验。



(a) Clustering quality (b) Cluster reassignment (c) Influence of k

Fig. 2: Preliminary studies. (a): evolution of the clustering quality along training epochs; (b): evolution of cluster reassignments at each clustering step; (c): validation mAP classification performance for various choices of k

性能对比

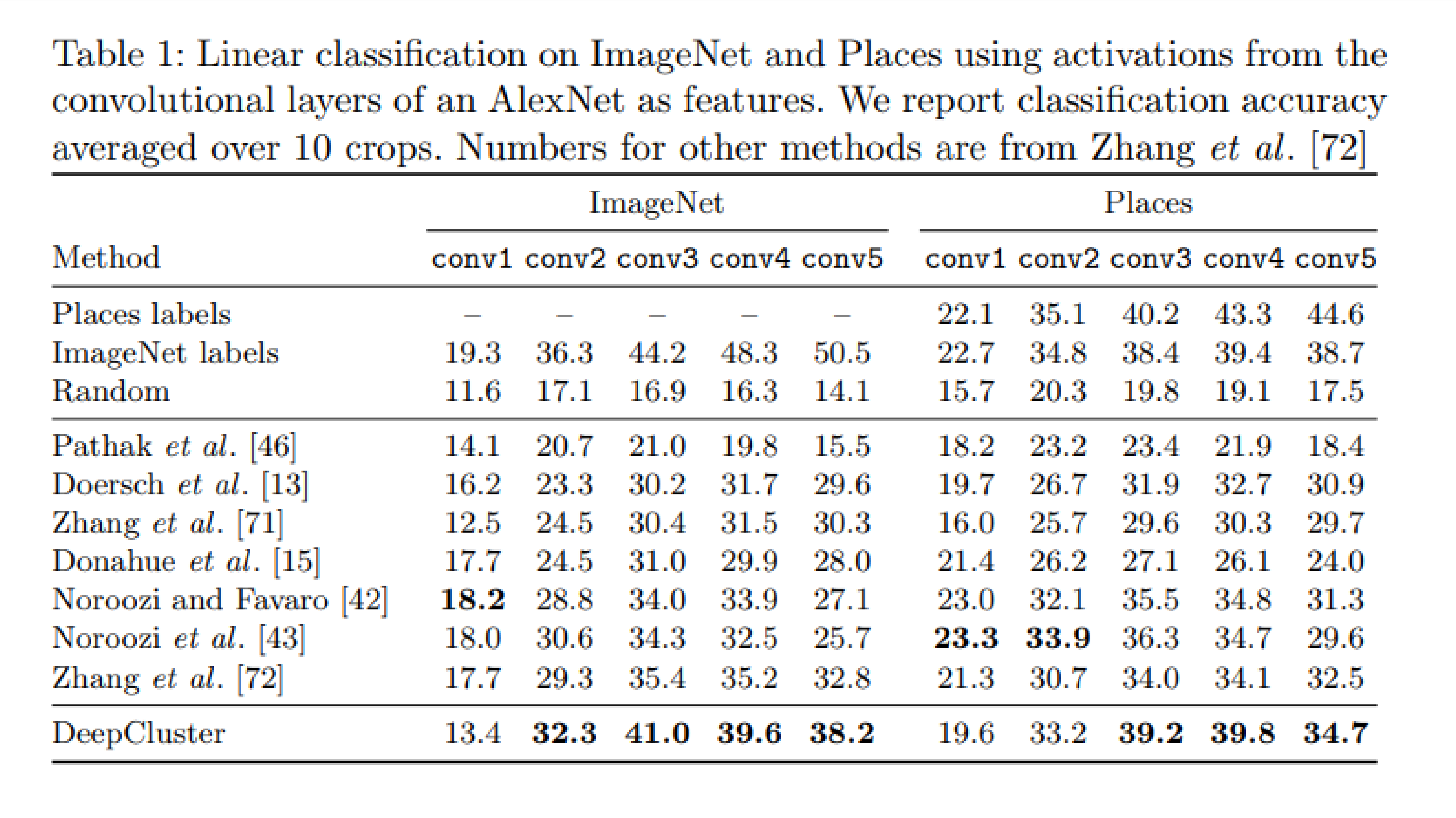


Table 1: Linear classification on ImageNet and Places using activations from the convolutional layers of an AlexNet as features. We report classification accuracy averaged over 10 crops. Numbers for other methods are from Zhang *et al.* [72]

	ImageNet					Places				
Method	conv1	conv2	conv3	conv4	conv5	conv1	conv2	conv3	conv4	conv5
Places labels	–	–	–	–	–	22.1	35.1	40.2	43.3	44.6
ImageNet labels	19.3	36.3	44.2	48.3	50.5	22.7	34.8	38.4	39.4	38.7
Random	11.6	17.1	16.9	16.3	14.1	15.7	20.3	19.8	19.1	17.5
Pathak *et al.* [46]	14.1	20.7	21.0	19.8	15.5	18.2	23.2	23.4	21.9	18.4
Doersch *et al.* [13]	16.2	23.3	30.2	31.7	29.6	19.7	26.7	31.9	32.7	30.9
Zhang *et al.* [71]	12.5	24.5	30.4	31.5	30.3	16.0	25.7	29.6	30.3	29.7
Donahue *et al.* [15]	17.7	24.5	31.0	29.9	28.0	21.4	26.2	27.1	26.1	24.0
Noroozi and Favaro [42]	**18.2**	28.8	34.0	33.9	27.1	23.0	32.1	35.5	34.8	31.3
Noroozi *et al.* [43]	18.0	30.6	34.3	32.5	25.7	**23.3**	**33.9**	36.3	34.7	29.6
Zhang *et al.* [72]	17.7	29.3	35.4	35.2	32.8	21.3	30.7	34.0	34.1	32.5
DeepCluster	13.4	**32.3**	**41.0**	**39.6**	**38.2**	19.6	33.2	**39.2**	**39.8**	**34.7**

拓展任务对比

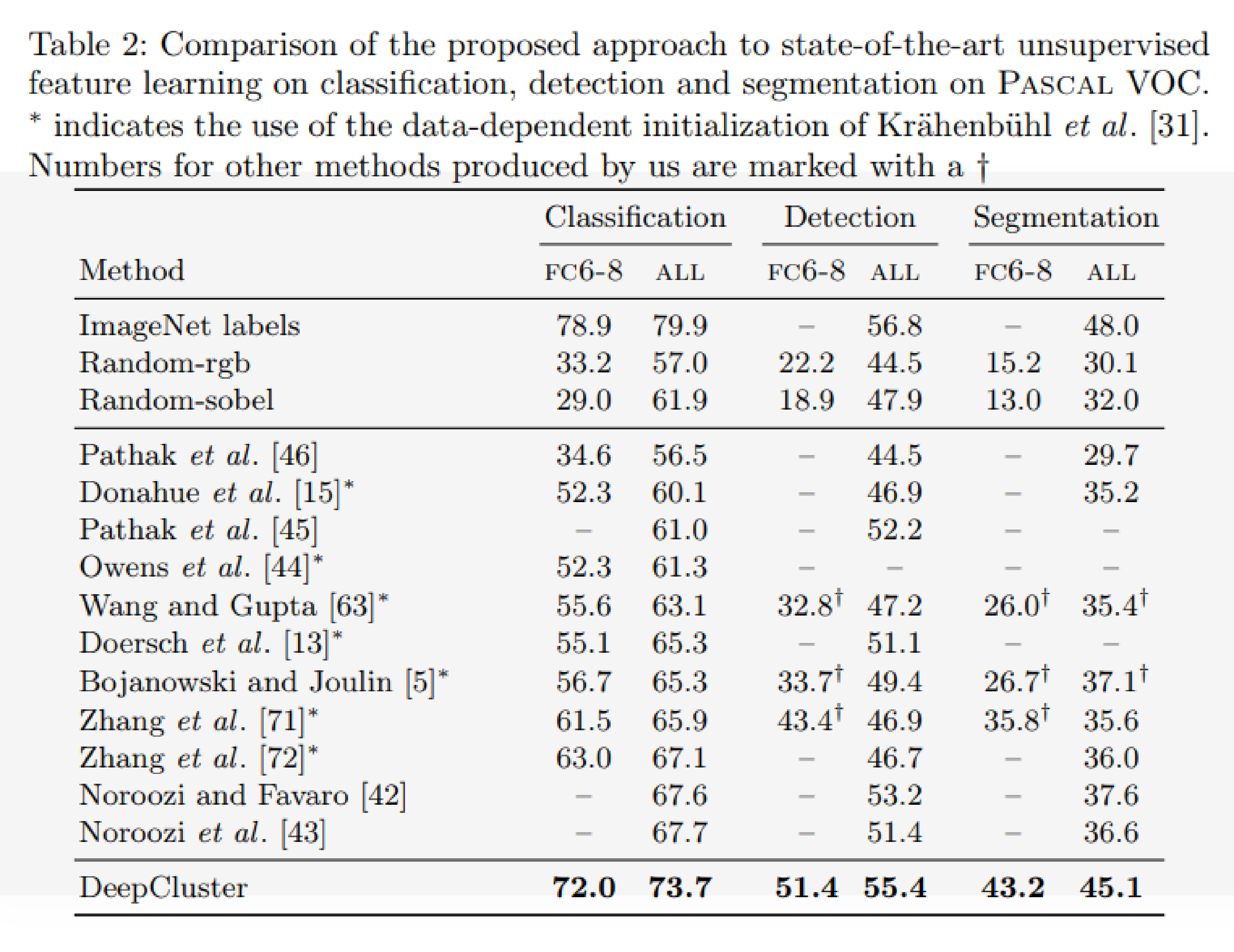


Table 2: Comparison of the proposed approach to state-of-the-art unsupervised feature learning on classification, detection and segmentation on PASCAL VOC. * indicates the use of the data-dependent initialization of Krähenbühl *et al.* [31]. Numbers for other methods produced by us are marked with a †

	Classification		Detection		Segmentation	
Method	FC6-8	ALL	FC6-8	ALL	FC6-8	ALL
ImageNet labels	78.9	79.9	–	56.8	–	48.0
Random-rgb	33.2	57.0	22.2	44.5	15.2	30.1
Random-sobel	29.0	61.9	18.9	47.9	13.0	32.0
Pathak *et al.* [46]	34.6	56.5	–	44.5	–	29.7
Donahue *et al.* [15]*	52.3	60.1	–	46.9	–	35.2
Pathak *et al.* [45]	–	61.0	–	52.2	–	–
Owens *et al.* [44]*	52.3	61.3	–	–	–	–
Wang and Gupta [63]*	55.6	63.1	32.8†	47.2	26.0†	35.4†
Doersch *et al.* [13]*	55.1	65.3	–	51.1	–	–
Bojanowski and Joulin [5]*	56.7	65.3	33.7†	49.4	26.7†	37.1†
Zhang *et al.* [71]*	61.5	65.9	43.4†	46.9	35.8†	35.6
Zhang *et al.* [72]*	63.0	67.1	–	46.7	–	36.0
Noroozi and Favaro [42]	–	67.6	–	53.2	–	37.6
Noroozi *et al.* [43]	–	67.7	–	51.4	–	36.6
DeepCluster	**72.0**	**73.7**	**51.4**	**55.4**	**43.2**	**45.1**

不同数据集上的对比试验

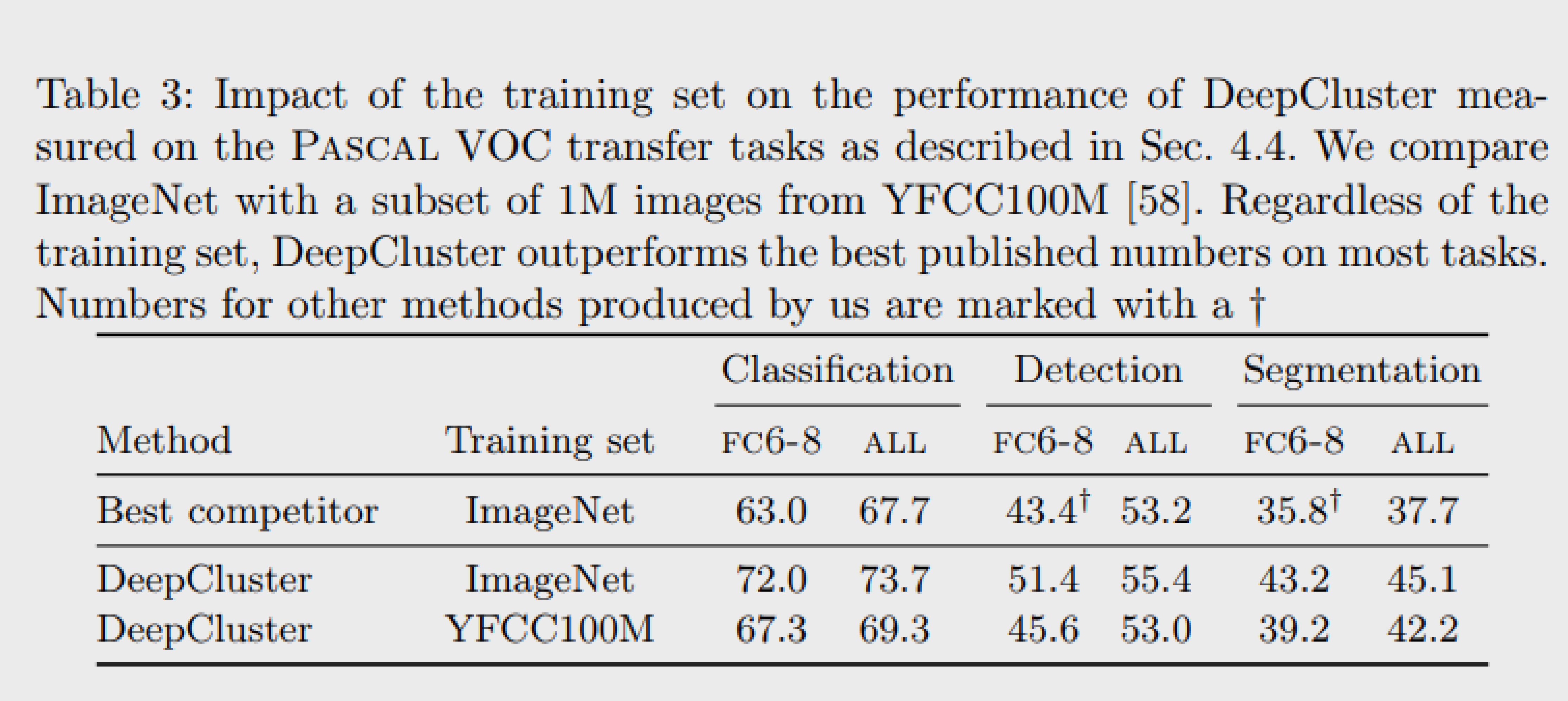


Table 3: Impact of the training set on the performance of DeepCluster measured on the PASCAL VOC transfer tasks as described in Sec. 4.4. We compare ImageNet with a subset of 1M images from YFCC100M [58]. Regardless of the training set, DeepCluster outperforms the best published numbers on most tasks. Numbers for other methods produced by us are marked with a †

		Classification		Detection		Segmentation	
Method	Training set	FC6-8	ALL	FC6-8	ALL	FC6-8	ALL
Best competitor	ImageNet	63.0	67.7	43.4†	53.2	35.8†	37.7
DeepCluster	ImageNet	72.0	73.7	51.4	55.4	43.2	45.1
DeepCluster	YFCC100M	67.3	69.3	45.6	53.0	39.2	42.2

不同网络框架对比

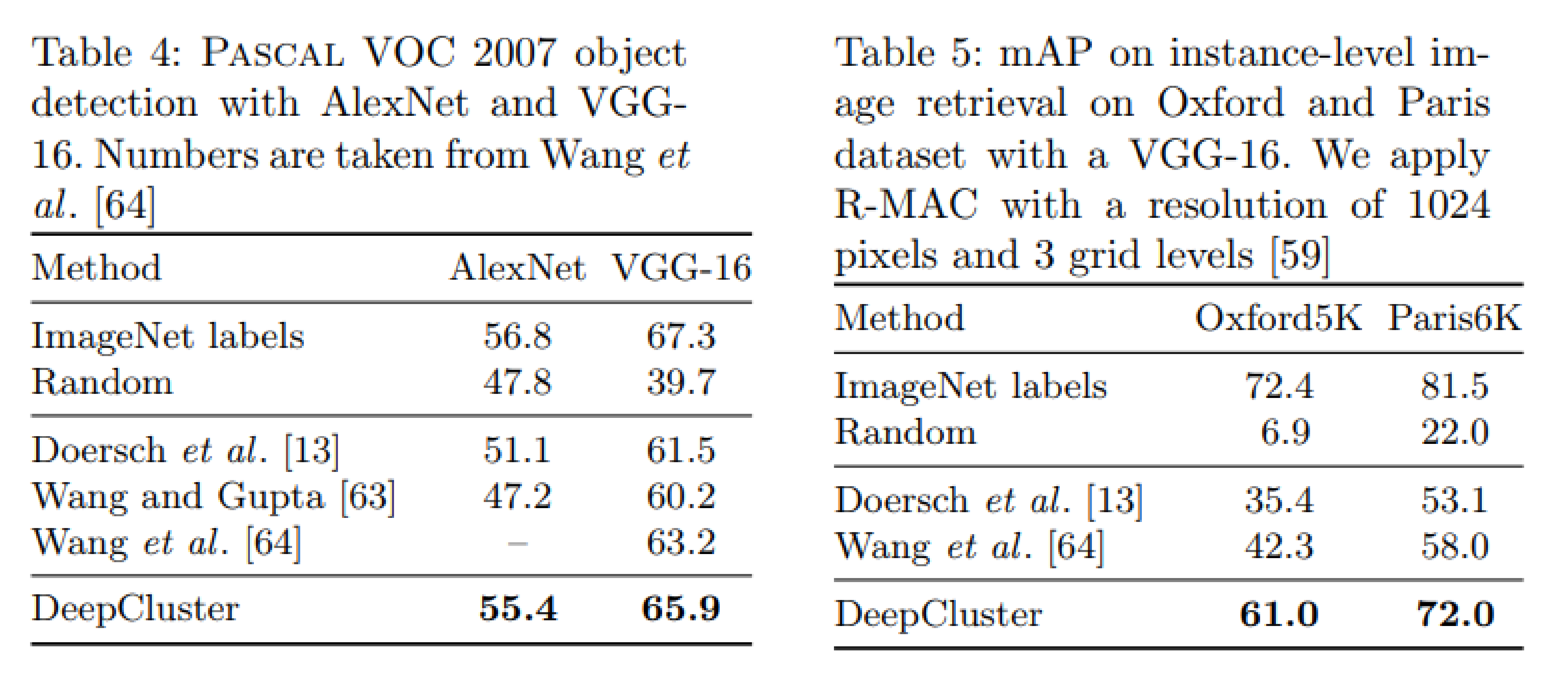


Table 4: Pascal VOC 2007 object detection with AlexNet and VGG-16. Numbers are taken from Wang *et al.* [64]

Method	AlexNet	VGG-16
ImageNet labels	56.8	67.3
Random	47.8	39.7
Doersch *et al.* [13]	51.1	61.5
Wang and Gupta [63]	47.2	60.2
Wang *et al.* [64]	–	63.2
DeepCluster	**55.4**	**65.9**

Table 5: mAP on instance-level image retrieval on Oxford and Paris dataset with a VGG-16. We apply R-MAC with a resolution of 1024 pixels and 3 grid levels [59]

Method	Oxford5K	Paris6K
ImageNet labels	72.4	81.5
Random	6.9	22.0
Doersch *et al.* [13]	35.4	53.1
Wang *et al.* [64]	42.3	58.0
DeepCluster	**61.0**	**72.0**

最后，作者做了在**不同网络结构**上的效果对比，和在图像恢复任务上的试验。按道理说，实验的效果会根据网络结构的变化而变化，网络结构越好最终的效果也越高，作者的实验也证明了，他的方法确实符合这个规则，也证明了他的模型的鲁棒性。

总结：

１．在这么多模型堆叠的现在，该网络框架是一股清流，简单框架，但是达到 state-of-the-art

２．对细节问题更加考究，比如解释为什么有这样的思路。读完让人很愉快

３．大量的实验对比，是真的超级有耐心，不清楚靠什么指标让作者守候训练12天。