Supplementary Materials

Yang Li[1] Shichao Kan[2] Jianhe Yuan[1] Wenming Cao[3] Zhihai He[1]*
[1]University of Missouri, MO, USA [2]Beijing Jiaotong University, Beijing, China
[3]Shenzhen University, Shenzhen, China
yltb5@mail.missouri.edu 16112062@bjtu.edu.cn
{yuanjia, hezhi}@missouri.edu wmcao@szu.edu.cn

In this Supplementary Materials, we provide additional experimental results and ablation studies for further understanding of the proposed algorithm and its performance.

A. Performance comparisons based on other backbone networks

In the main paper, we have evaluated the performance of our spatial assembly network (SAN) based on the GoogleNet backbone network. To demonstrate that this SAN module can be inserted into different networks to improve their performance, in this section, we evaluate another network, the BN-Inception [3] network, on the CUB dataset. We include the recent state-of-the-art methods with BN-Inception for performance comparison. These methods include: HIL (hierarchical triplet loss) [1], MS (Multi-Similarity) [9], SoftTriple [8], and XBM (Cross-Batch Memory) [10]. We use the multi-similarity loss [9] with momentum memory bank [2, 5] as the baseline system for our SAN method. From the Table 1, we can see that the Baseline with SAN has improved the Recall@1 rate by 1.3%.

Table 1. Recall@K (%) performance on the CUB dataset with **BN-Inception** in comparison with other supervised metric learning methods.

Methods	CUB			
	R@1	R@2	R@4	R@8
HTL [1] ECCV18	57.1	68.8	78.7	86.5
MS [9] CVPR19	65.7	77.0	**86.3**	91.2
SoftTriple [8] CVPR19	65.4	76.4	84.5	90.4
XBM [10] CVPR20	65.8	75.9	84.0	89.9
Baseline	66.3	76.7	85.1	90.8
Baseline with SAN	**67.6**	**77.6**	85.5	**91.3**

B. Sensitivity analysis of hyper-parameters

In (12) of Section 3.3, λ_1, λ_2, and λ_3 are the weighting parameters. In our experiments, if we change the values of λ_1 and λ_2 between 0.5 and 1, the final Recall@1 rate will vary by about 0.5%. To address the issue that feature points from the same object may be disassembled into different locations in the output feature map, we introduce the local coherence into the spatial assembly operation. In this section, we conduct experiments on different λ_3 values to study impact of local coherence. Figure 1 shows the Recall@K performance of our proposed SAN module with different local coherence weights.

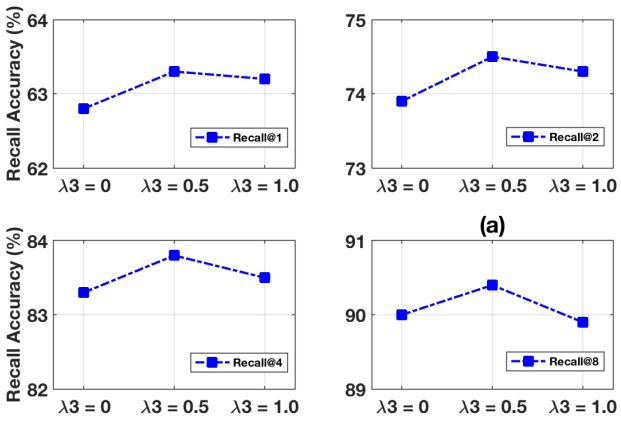

Figure 1. Recall@K (%) performance of SAN module with different local coherence (LC) weights on the CUB dataset.

C. Impact of different embedding sizes on supervised deep metric learning

In this section, we follow existing supervised metric learning methods [7, 9, 6] to evaluate the impact of different embedding sizes. Table 2 shows the impact of different embedding sizes on the CUB dataset with GoogleNet. We can see that the performance gradually improves with the dimension from 64, 128, 256, 512, to 1024.

*Corresponding author: Zhihai He, e-mail: hezhi@missouri.edu.

Figure 2. Retrieval examples by the baseline with our SAN module on the CUB, Cars, and SOP datasets from unsupervised metric leaning. The query images and the incorrect retrieved images are highlighted with *blue* and *red*.

D. Retrieval examples from unsupervised metric learning

Figure 2 shows the retrieval examples by the baseline with our SAN module on the CUB, Cars, and SOP datasets from unsupervised metric learning. Examples highlighted with blue and red boxes are query images and incorrect retrieval results. We can see that our SAN can learn discriminitive features, even image labels are not available.

Table 2. Recall@K (%) performance with **GoogleNet** on the CUB dataset in comparison with different embedding sizes.

Embedding Size	CUB			
	R@1	R@2	R@4	R@8
64	58.1	70.3	80.3	88.2
128	60.6	72.5	81.7	88.6
256	62.4	73.6	82.4	89.7
512	63.3	**74.5**	**83.8**	**90.4**
1024	**63.5**	74.3	83.3	90.2

E. Spatial assembly on the CLEVR compositional images.

CLEVR is a synthetic dataset developed by [4] for studying compositional languages and elementary visual reasoning of spatial relationship between objects. We modify the CLEVR toolkit [4] to generate synthesized examples. Figure 3 (left) shows examples from Class 1 and Class 2. We can see that, within the same class, all images have the same objects, but with different spatial layout. We perform supervised metric learning on these examples. Figure 3 (right) shows the generated features for Class 1 (blue) and Class 2 (red) being projected into a 1-D space with (bottom) and without the SAN module (top). We can see that, using the SAN module, the learned features are much more discriminative.

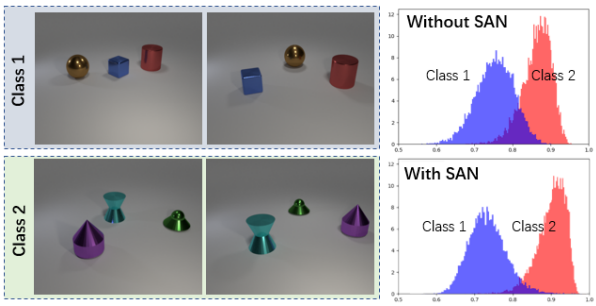

Figure 3. A synthetic two-class example generated by the CLEVR toolkit. Left: examples images from Class 1 and Class 2. Right, the learned features for Class 1 (blue) and Class 2 (red) being projected into a 1-D space with and without the SAN module.

References

[1] Weifeng Ge. Deep metric learning with hierarchical triplet loss. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 269–285, 2018. 1

[2] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019. 1

[3] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 1

[4] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910, 2017. 2

[5] Shichao Kan, Yigang Cen, Yang Li, Mladenovic Vladimir, , and Zhihai He. Contrastive bayesian analysis for supervised deep metric learning. In *Github*, 2020. 1

[6] Sungyeon Kim, Dongwon Kim, Minsu Cho, and Suha Kwak. Proxy anchor loss for deep metric learning. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3238–3247, 2020. 1

[7] Michael Opitz, Georg Waltner, Horst Possegger, and Horst Bischof. Deep metric learning with bier: Boosting independent embeddings robustly. *IEEE transactions on pattern analysis and machine intelligence*, 2018. 1

[8] Qi Qian, Lei Shang, Baigui Sun, Juhua Hu, Hao Li, and Rong Jin. Softtriple loss: Deep metric learning without triplet sampling. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6450–6458, 2019. 1

[9] Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R Scott. Multi-similarity loss with general pair weighting for deep metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5022–5030, 2019. 1

[10] Xun Wang, Haozhi Zhang, Weilin Huang, and Matthew R Scott. Cross-batch memory for embedding learning. *arXiv preprint arXiv:1912.06798*, 2019. 1