Supplementary Material: Efficient Object Embedding for Spliced Image Retrieval

Bor-Chun Chen[1,2] Zuxuan Wu[3*] Larry S. Davis[1] Ser-Nam Lim[2]

[1]University of Maryland, College Park [2]Facebook AI [3]Fudan University

{sirius, lsd}@cs.umd.edu, zxwu@fudan.edu.cn, sernamlim@fb.com

1. Feature Extraction Model

Different spatial pooling techniques [7] and post-processing steps such as dimensionality reduction [4] have been shown to greatly affect retrieval performance. We provide a detailed analysis of different parameters for selecting an effective feature extraction model.

PCA and pooling. Given a convolutional feature map from conv5_3 layer $F \in R^{W\times H\times C}$, we consider the following spatial pooling functions P: $R^{W\times H\times C} \rightarrow R^{C}$: (1) sum pooling [1] (SPoC), (2) max-pooling [10] (Max), (3) regional max-pooling [13] (R-MAC), and (4) generalized mean pooling [9] (GeM). We also perform experiments while varying the number of dimensions in PCA from 64 to 2,048 with whitening. Figure 1 shows a detailed analysis of the effect of different pooling techniques and post-processing steps. Figure 1 (a) shows retrieval performance of four benchmarks with different PCA dimensions. Even though the performance of all embeddings decreases as the feature dimension goes down, embeddings from the classification model (ResNet50) consistently perform the best for all dimensions, which further supports our observation in the paper. Figure 1 (b) shows the mAP for different pooling techniques. Here, ResNet50 embeddings again consistently achieve the best performance among embeddings from different pre-trained models on all datasets.

Embeddings from different layers. Figure 2 shows the performance with embeddings extracted from different layers in ResNet50 backbone from conv4_1 to conv5_3. Note that for lower-level embeddings, detection models and classification models share similar performance, because they represent similar low-level texture features. However, their performance diverges for embeddings from high-level layers. This is an important observation since embeddings extracted from a higher level (conv5_x) achieve better retrieval performance across all datasets. This again supports the embeddings from classification models as being better suited for image retrieval.

*Work done during author was in Facebook AI.

2. Student Networks

Figure 3 provide an illustration of the proposed student network (S_3) and two baseline student networks (S_1 and S_2) described in the paper. Figure 3 (a) shows the S_1 baseline network, which consists of five bottleneck layers and directly takes the image as input. (b) shows the S_2 baseline network, which consists of three bottleneck layers. S_2 share the low-level features from the detector and take the feature map from the second residual stage of the detector model F_{I_2} as input. (c) shows the proposed student network, which utilizes multi-scale features from the detector as inputs to learn discriminative features.

3. Landmark Retrieval

Table 1 (top) compares different landmark retrieval approaches with ImageNet pretrained models, including sum pooling of convolutions (**SPoC**) [1], maximum activation of convolutions (**MAC**) [10], regional maximum activation (**R-MAC**) [13], generalized mean pooling (**GeM**) [9], and our object embedding approach with the ResNet50 network, (**OE-SIR**) on $\mathcal{R}$Oxford5K and $\mathcal{R}$Paris6K dataset with medium protocol. For a fair comparison, we employ the same ResNet50 pre-trained on ImageNet for all the methods. Also, we do not apply any additional post-processing except PCA whitening. Our approach achieves the best performance among other approaches using the same pre-trained network. Table 1 (bottom) compares different state-of-the-art approaches on the same dataset. Note that state-of-the-art methods utilize different additional training data. For example, Radenovic *et al*. [8] utilize training data pairs collected from spatial verification with local features while Teichmann *et al*. [11] utilize Google landmark dataset as additional training data. Here we also utilize the Google landmark dataset to fine-tune our model and extract object embeddings from the fine-tuned model (OE-SIR-FT). Details on the training process are described in Section 4. Our model achieves competitive performance without any post-processing except PCA

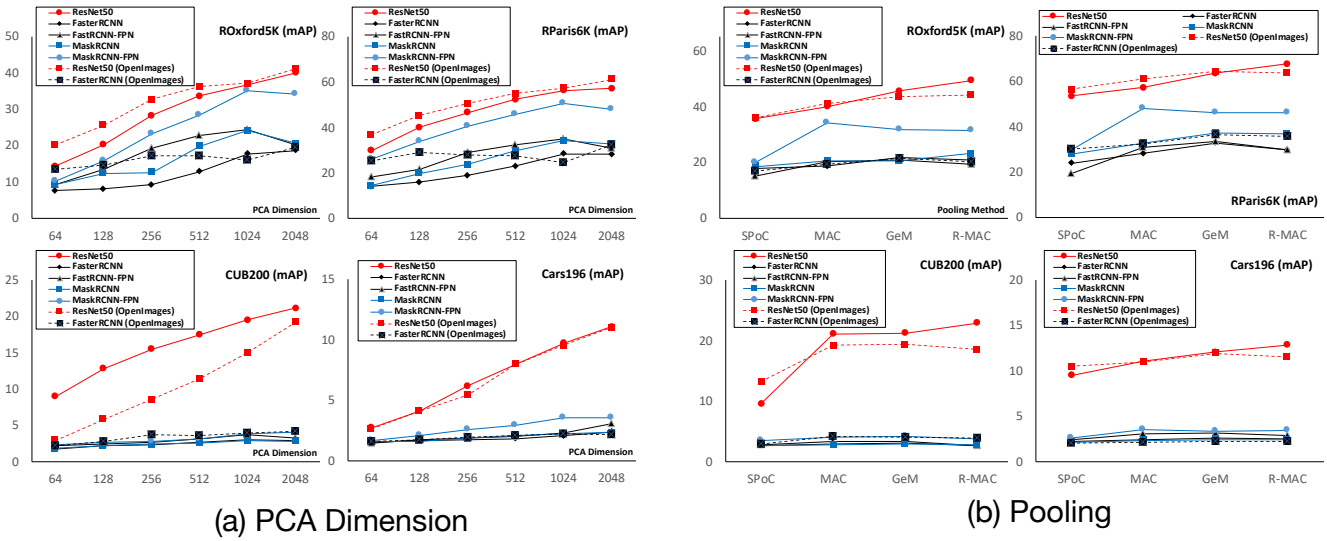

Figure 1. Analysis of embeddings with (a) different PCA dimension and (b) different pooling techniques. Embeddings learned from classification model consistently achieve the best performance.

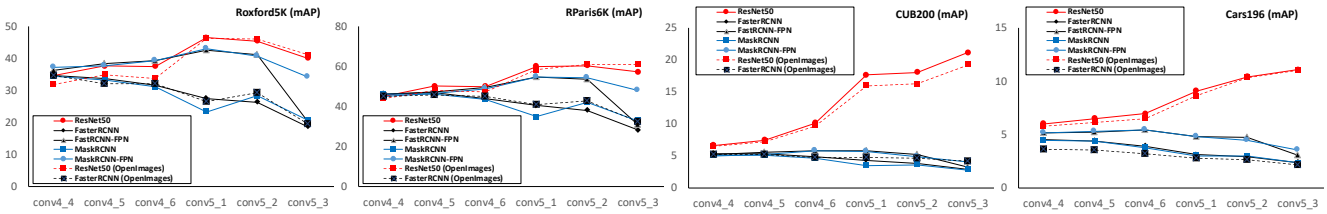

Figure 2. Performance of embeddings extracted from different layers of the pre-trained models. Embeddings from lower layers of classification and detection models have similar performance as they learn similar low-level texture features. However, their performance starts to diverge as we use higher layers, with the classification model achieving better performance.

whitening, compared to the state-of-the-art model that uses re-ranking techniques such as spatial verification.

4. Implementation Details

For all retrieval experiments, we follow [7] and resize the longer side of images down to 1024 if it is greater than 1024 and keep the original resolution if it is smaller than 1024. We use a batch size of 1 to extract features from images with different resolutions. We apply PCA whitening trained on 10K images randomly selected from OpenImagesV4 [5] dataset.

For training the student models in Section 2, we randomly resize and crop images of size 224×224 from the OpenImagesV4 dataset for training, we use a batch size of 64 and learning rate of 1e-3 with Adam optimizer for training.

For experiments with fine-tuning in Section 3, we use Google Landmark dataset [12] as the training set for experiments with ROxford5k and RParis6K. We use stochastic gradient descent with momentum for training, with a learning rate of 1e-3 and a momentum of 0.9. We use weight decay of 1e-5 and batch size of 64. Cosine annealing [6] is applied to expedite the training process. For training efficiency, we batch the images with similar size and resize them to some canonical size such as 512×384, 384×512, 448×448. We use the method described in [2] with softmax cross-entropy loss to fine-tune the networks. We use standard data argumentation such as random cropping and resizing, and training the network for five epochs. For OE-SIR experiments with landmark retrieval, we use landmark-related classes for detection, while for spliced image retrieval, we use all 600 classes in the OpenImagesV4 dataset.

5. Qualitative Results

Figure 4 shows some failure cases of the proposed method on the PIR dataset. The first column shows the query Spliced image, the second column shows the top-1 retrieved results with the proposed method and the third column shows the corresponding authentic image. In the first example (first row), OE-SIR matches the incorrect objects

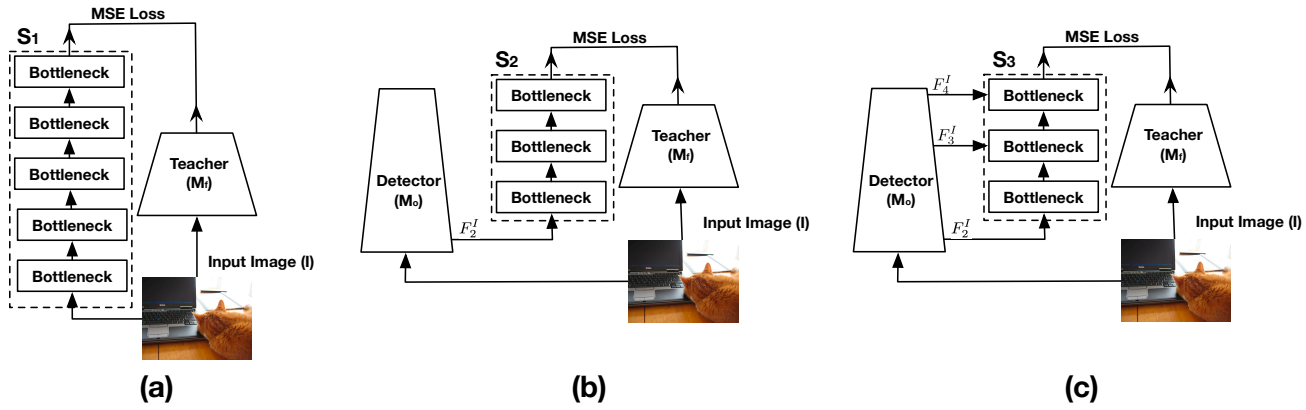

Figure 3. An illustration of the baseline student networks and the proposed student network. (a) S_1 directly takes image as input to learn discriminative features. (b) S_2 share the low-level feature map F_2^I from detector. (c) The proposed student network leverage multi-scale features from the detector to efficiently learn the discriminative features.

Method	$\mathcal{R}$Oxford5K		$\mathcal{R}$Paris6K	
w/ ImageNet pretrained model	mAP	P@10	mAP	P@10
SPoC [1]	35.7	55.4	53.5	90.3
MAC [10]	40.1	61.3	57.3	96.7
R-MAC[13]	49.4	70.4	67.6	98.1
GeM [9]	45.7	67.2	63.6	96.3
OE-SIR (Ours)	**53.4**	**76.0**	**69.7**	**98.6**
w/ additional training data	mAP	P@10	mAP	P@10
ResNet101-R-MAC [8]	60.9	78.1	78.9	96.9
ResNet101-GeM [3]	64.7	84.7	77.2	98.1
DELF–D2R-R-ASMK [11]	73.3	90.0	80.7	99.1
DELF–D2R-R-ASMK+SP [11]	76.0	**93.4**	80.2	**99.1**
OE-SIR-FT (Ours)	**78.7**	91.8	**83.4**	98.3

Table 1. Comparison of different approaches on $\mathcal{R}$Oxford5K and $\mathcal{R}$Paris6K datasets with or without additional training data. Our approach achieves the best performance among other baselines even when a compact student model is deployed. For models with additional training data, our model achieves competitive performance even when comparing with the model using a re-ranking method such as spatial verification.

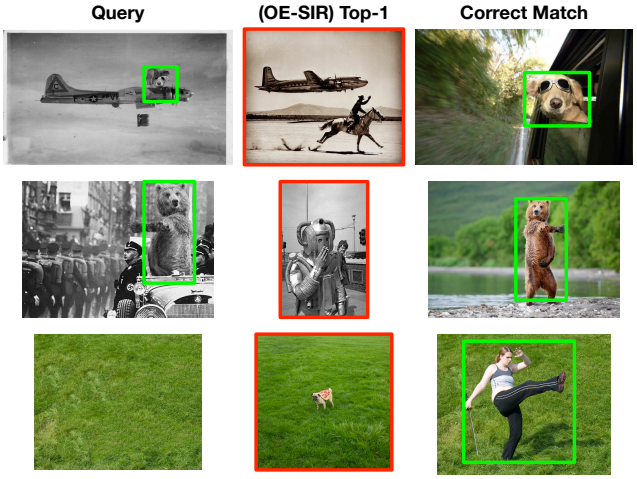

Figure 4. Some failure cases on PIR.

(airplane) and therefore fails to retrieve the authentic image. In the second example, OE-SIR fails to match the object because of the appearance changes. In the third example, OE-SIR fails to detect any objects since the object is removed from the authentic image. The failure cases show the limitation of the proposed method and provide some insight for future research in this direction.

References

[1] Artem Babenko and Victor Lempitsky. Aggregating local deep features for image retrieval. In *Proceedings of the IEEE international conference on computer vision*, pages 1269–1277, 2015. 1, 3

[2] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 2

[3] Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. Deep image retrieval: Learning global representations for image search. In *European conference on computer vision*, pages 241–257. Springer, 2016. 3

[4] Hervé Jégou and Ondřej Chum. Negative evidences and cooccurences in image retrieval: The benefit of pca and whitening. In *European conference on computer vision*, pages 774–787. Springer, 2012. 1

[5] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification,

object detection, and visual relationship detection at scale. *arXiv:1811.00982*, 2018. 2

[6] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 2

[7] Filip Radenović, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondřej Chum. Revisiting oxford and paris: Large-scale image retrieval benchmarking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5706–5715, 2018. 1, 2

[8] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples. In *European conference on computer vision*, pages 3–20. Springer, 2016. 1, 3

[9] Filip Radenović, Giorgos Tolias, and Ondrej Chum. Fine-tuning cnn image retrieval with no human annotation. *IEEE transactions on pattern analysis and machine intelligence*, 2018. 1, 3

[10] Ali S Razavian, Josephine Sullivan, Stefan Carlsson, and Atsuto Maki. Visual instance retrieval with deep convolutional networks. *ITE Transactions on Media Technology and Applications*, 4(3):251–258, 2016. 1, 3

[11] Marvin Teichmann, Andre Araujo, Menglong Zhu, and Jack Sim. Detect-to-retrieve: Efficient regional aggregation for image search. *arXiv preprint arXiv:1812.01584*, 2018. 1, 3

[12] Marvin Teichmann, Andre Araujo, Menglong Zhu, and Jack Sim. Detect-to-retrieve: Efficient regional aggregation for image search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5109–5118, 2019. 2

[13] Giorgos Tolias, Ronan Sicre, and Hervé Jégou. Particular object retrieval with integral max-pooling of cnn activations. *arXiv preprint arXiv:1511.05879*, 2015. 1, 3