Squeezed Bilinear Pooling for Fine-Grained Visual Categorization*

Qiyu Liao[1], Dadong Wang[1],Hamish Holewa[2], Min Xu[3]
CSIRO DATA61[1], NCMI[2], University of Technology, Sydney[3]
{qiyu.liao, dadong.wang, Hamish.Holewa}@csiro.au, min.xu@uts.edu.au

Abstract

In this paper, we propose a supervised selection based method to decrease both the computation and the feature dimension of the original bilinear pooling. Different from currently existing compressed second-order pooling methods, the proposed selection method is matrix normalization applicable. Moreover, by extracting the selected highly semantic feature channels, we proposed the Fisher-Recurrent-Attention structure and achieved state-of-the-art fine-grained classification results among the VGG-16 based models.

1. Introduction

Bilinear pooling was first proposed to address the challenge of Fine-Grained Visual Classification (FGVC) by Lin et al. [12]. Based on the bilinear pooling, Lin et al. [11] investigated matrix square-root normalization to significantly improve the representation of the bilinear feature. However, a neglected problem of the above feature encoding method is its extremely high output feature dimension. The tensor product makes c CNN output channels to c^2 dimension of pooled features. A relatively low $c = 512$ VGG-16[15] structure produces a $512 \times 512 \approx 262k$ dimension bilinear features. To deal with this problem, Tensor Sketching was investigated in [5] and similar accuracy was reported with 8K compact features. However, the linear combination significantly increases the computational complexity of bilinear features. To solve the computation dilemma, a low-rank approximation based method was proposed in [8] and obtained a similar performance of the original full bilinear pooling. Given these methods reduced the dimension and computational complexity by two orders of magnitude, one vital problem is that the matrix power function cannot propagate through the compact layer. Sub-normalization was employed in [7] to solve the problem, however, the performance is not as good as expected since the categorization accuracy drops around 1% comparing with that obtained with the baseline structure. It remains a problem to combine a compressed bilinear structure with matrix power normalization.

	Fast Computation	Matrix Normalization	Attention Interpretable
CBP	×	×	×
LRBP	√	×	×
MoNet	×	√	×
SBP	√	√	√

Table 1. Comparison on the proposed SBP and other compressed bilinear pooling based methods.

In this paper, we propose a novel compressed model, named Squeezed Bilinear Pooling (SBP), that can linearly reduce both the feature dimension and computation. With the same dimension, the proposed method outperforms the other compressed models (e.g. CBP[5], LRBP[8]), and is two orders of magnitude faster. The integration of the matrix square root layer is also investigated to enhance the categorization performance. Based on the selected discriminative feature channels, we propose the Fisher-Recurrent-Attention structure to achieve state-of-the-art classification accuracies on three common FGVC datasets.

2. The squeezed bilinear CNN

The overview of the proposed SBP architecture is shown in Fig. 1. For each input image I, the convolutional neural network outputs a feature matrix $M = \{m_1, m_2 \ldots, m_c\}$, where m_i is the expansion tensor of the i-th channel of CNN features. Following [12], the co-inner production is conducted on M to produce $c \times c$ second-order maps $\hat{M} = \{\hat{m}_1, \hat{m}_2, \ldots, \hat{m}_{c^2}\}$. Then the novel Fisher Selection Layer (FSL) and Global Average Pooling (GAP) are applied to generalize the d-dimension squeezed second-order feature vector, followed by the element-wise square root regularization, l_2-normalization layers and a full connection classification layer. In the parallel workflow, a matrix square root layer is applied as a bridge between the full bilinear layer and the FSL to further improve the performance

*The first workshop on Statistical Deep Learning for Computer Vision, in *Seoul, Korea*, 2019. Copyright by Author(s).

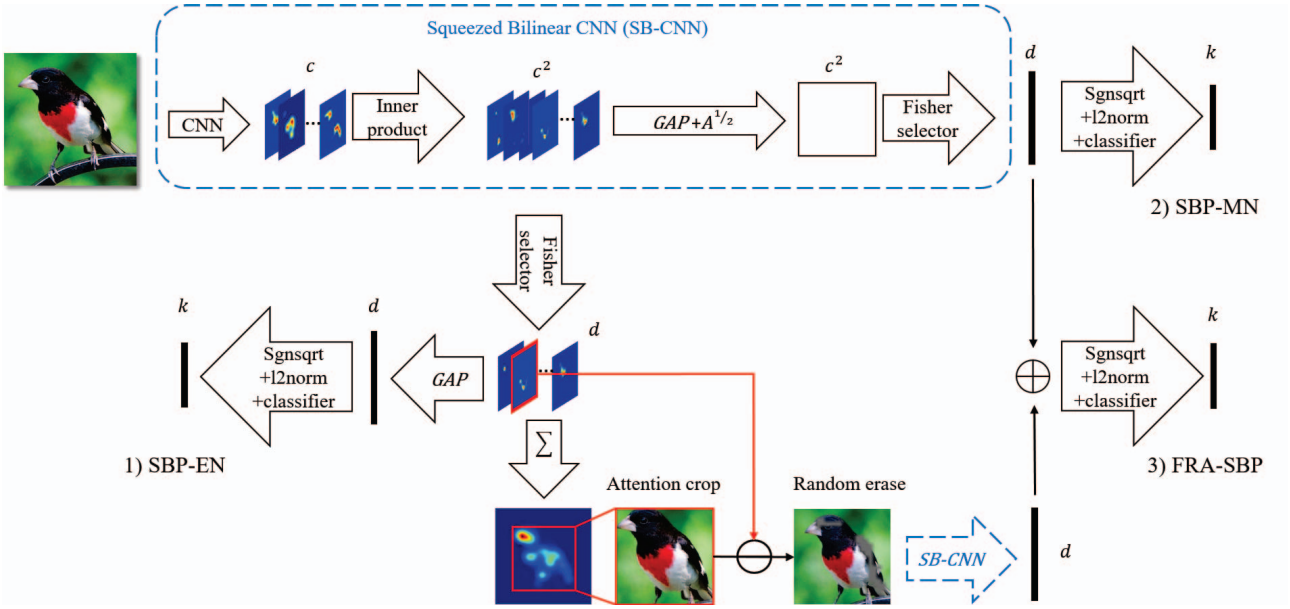

Figure 1. The proposed network architecture with three SBP based models: 1) the Squeezed Bilinear Pooling with Element-wise Normalization (SBP-EN) for fast computation, 2) the Squeezed Bilinear Pooling with Matrix Normalization (SBP-MN) by inserting the matrix square root function before the squeezing layer, and 3) the Fisher Recurrent Attention Squeezed Bilinear Pooling (FRA-SBP).

of the proposed SBP. The two flows are named as Squeezed Bilinear Pooling with element-wise (SBP-EN, illustrated in Fig. 1.(1)), and matrix normalization (SBP-MN, in Fig. 1.(2)), respectively. We also designed the Fisher Recurrent Attention SBP (FRA-SBP, shown in Fig. 1.(3)). We will discuss these structures in detail in the following sections.

Fisher Discriminant Analysis (FDA) discriminates patterns using the low-dimensional projection of high dimensional features with linear transformations. Its main idea is to maximize the interclass variations and minimize intraclass variations. We follow the similar idea to maximize the class separation measurement via the feature selection. Fukunaga [4] proved that the traces of the scatter matrice could be used to measure the class separation of the features. After mapping into and calculating the scatter matrices in the kernel space[17], the traces of the intra and interclass scatter matrices in kernel space, i.e., $\tilde{S}_w$ and $\tilde{S}_b$ can be obtained as:

$$\begin{aligned} Tr(\tilde{S}_w) &= \frac{1}{n}Tr(K) - \frac{1}{n}\sum_{i=1}^{g}\frac{1}{n_i}Sum(K^{(i)}), \\ Tr(\tilde{S}_b) &= \frac{1}{n}\sum_{i=1}^{g}\frac{1}{n_i}Sum(K^{(i)}) - \frac{1}{n^2}Sum(K). \end{aligned} \tag{1}$$

where the operators $Sum(\cdot)$ and $Tr(\cdot)$ calculate, respectively, the summation of all elements and the trace of a matrix, and K and $K^{(i)}$ are the $n\times n$ and $n_i\times n_i$ sized matrices defined by:

$$\{K\}_{kl} = k(x_k, x_l), \{K^{(i)}\}_{uv} = k(x_u^{(i)}, x_v^{(i)}). \tag{2}$$

where $x_i^{(j)}$ denotes the i-th feature observation in Class C_j.

The feature selector is denoted by $\alpha = [\alpha_1, ..., \alpha_p]^T \in \{0,1\}^p$ with $\alpha_k = 1$ indicating that the k-th feature is selected or 0 not-selected. Then the selected feature set from the original feature vector x is given by $x(\alpha) = x \odot \alpha$. With the feature selector, the traces of the scatter matrices are noted by $Tr(\tilde{S}_w)(\alpha)$, $Tr(\tilde{S}_b)(\alpha)$. To maximize the class separation, the optimization objective function can be formulated as:

$$\arg\max_{\alpha\in\{0,1\}^p}\{Tr(\tilde{S}_b)(\alpha) - \lambda Tr(\tilde{S}_w)(\alpha)\}. \tag{3}$$

The general Fisher selector formulation is a combinatorial optimization problem for many kernels, and it is far from feasible in our bilinear features whose p is up to over 262K. The polynomial kernel provides efficient and global optimization for (FS) for large p [16]:

$$k(x_1, x_2)(\alpha) = \langle x_1 \odot \alpha, x_2 \odot \alpha \rangle = \sum_{i=1}^{p} x_{1i}x_{2i}\alpha_i \tag{4}$$

Substitute (4) and the scatter matrix expression (1) into the object function (3) and normalize, we can get the Fisher discriminative score:

$$\begin{aligned} \theta_j =& \frac{1}{n}\sum_{i=1}^{g}\frac{1}{n_i}\sum_{u,v=1}^{n_i} x_{uj}^{(i)}x_{vj}^{(i)} \\ & - \frac{\lambda}{n}\sum_{i=1}^{n}x_{ij}^2 + \frac{(\lambda-1)}{n^2}\sum_{u,v=1}^{n} x_{uj}x_{vj} \end{aligned} \tag{5}$$

Considering the preset d object dimension of compressed image feature, the Fisher optimization objective function can be depicted as follows:

$$\arg\max_{\alpha\in\{0,1\}^p} \sum_{j=1}^{p} (\theta_j - \beta)\alpha_j, s.t. \|\alpha\|_0 = d. \tag{6}$$

This is the globally optimal solution, and the computational complexity for the calculation of α with n training samples and p feature dimensions is $O(n^2 p)$.

2.1. Squeezed bilinear

With the full bilinear feature $M = X^T X \in R^{c^2}$, the Fisher selector α, and the objective projection dimension d, the projection function for the Fisher selection layer can be represented as:

$$R^{c^2} \xrightarrow{\psi} R^d, \psi(M) = M \circ \alpha, \tag{7}$$

where the operator $(a \circ b)$ requires the same size of the tensors a and b, aiming at extracting the a values of position that is not 0 in b, to form a new d dimensional feature tensor ($\|a\|_0 = d$). The computational complexity is $O(hwd)$ for a d dimensional squeezed bilinear feature with CNN feature maps of size $h \times w \times c$.

The backpropagation is the converse process of the forwarding (7). The operator $\circ$ is not differentiable but is an element-wise first-order linear combination, hence it can be solved by the combination of element-wise derivatives. For each $\alpha_i \neq 0, \rho(i)$ is the projected index of i by the selection function ψ (if projected), the backpropagation function can thus be depicted as:

$$\frac{\partial L}{\partial M_i \alpha_i} = \frac{\partial L}{\partial \psi(M)_{\rho(i)}}. \tag{8}$$

For selected elements, $\alpha_i = 1$, we can directly pass the back gradient to the corresponding channels of CNN feature maps.

The implementation of the matrix power normalization $M' = M^{1/2}$ requires a positive definite forward input feature matrix, and a symmetric backward gradient matrix [11]. The SBP-MN structure satisfies the first requirement (Fig.1), while the second requirement can be met by a matrix diagonalization. Suppose $\delta(i)$ is the diagonal position of index i in the bilinear feature matrix M, we can describe the improved backpropagation function for matrix power normalization as follows:

$$\frac{\partial L}{\partial M'_i \alpha_i} = \frac{\partial L}{\partial M'_{\rho(i)} \alpha_i} = \frac{1}{2}\left(\frac{\partial L}{\partial \psi(M')_{\rho(i)}} + \frac{\partial L}{\partial \psi(M')_{\rho(\delta(i))}}\right). \tag{9}$$

Noticing that the Squeezed Bilinear Pooling selects the most discriminative second order features, it can be used to localize the object region and suppress the irrelevant background activation. As illustrated in Fig. 1. (3), the d selected second order feature maps $\{m_1, m_2, ..., m_d\} \in \psi(M)$ are element-to-element summed to produce the average activated map m_a. The m_a is then linearly resized to input image size, and its values are normalized into the range of $(0, 1)$, called the normalized attention map m_n. A threshold ε is applied to m_n to segment the attention activation map m_s. Following the spirit of [21], we randomly select an second order map $m_k \in \psi(M)$, and erase the activated region of m_k from m_s to obtain the attention erased map $m_e = (m_k == 0) \cdot m_s$.

After the random erasion, we crop the TRUE region of m_e from the input image I to create the attention image I_a. It is recurrently inputted into the SB-CNN and outputs the d dimensional attention feature f_a. The two stages of features f and f_a are cascaded and, after sgnsqrt and l_2 normalization layer, classified with a fully connected layer.

3. Experiment

In this section, we detail our experiments from two aspects. (1) In Section 3.1, we investigate the impact of the selected dimension and compare with the compact bilinear pooling. (2) In Section 3.2, we conduct an overall comparison of our proposed squeezed models against other methods on a variety of fine-grained datasets.

When evaluating our model on the VGG-16[15] structure (D-net in [12]), we used the convolutional layers of VGG-16 as the local feature extractor and retained the output of the Conv5_3+ReLU layer for the second-order encoding, as conducted in [12]. We resized the input images to 512×512 and randomly cropped them to 448×448 for training in all the second-order models. For Fisher Selective Layer, the λ was assigned a value of -0.5 and for attention cropping in FRA-SBP, the value of ε was assigned to 0.005.

3.1. Configuration and comparison with compact bilinear

To investigate the potential impact of the selected dimension on the squeezed bilinear pooling, we conducted experiments with the selected dimension in the range of 100 to 15,000 on the CUB-200-2011 dataset. As summarized in Fig. 2, with the increasing of projected dimension, the top-1 errors of both the CBP[5] and the SBP come down to a similar level of the full bilinear. With a lower dimension, the SBP's performance is more promising than the Tensor Sketch. SBP outperforms CBP by around 1.5% when the dimension ranges from 1K to 5K, and the gap widens as the dimension decreases. With 500 projected dimensions, significant disparity can be observed between the two methods (4.6% without and 2.6% with fine-tuning). In extremely low dimension cases, e.g. 100, SBP produces acceptable accuracy loss (22.7%) comparing with CBP (42.8%). This

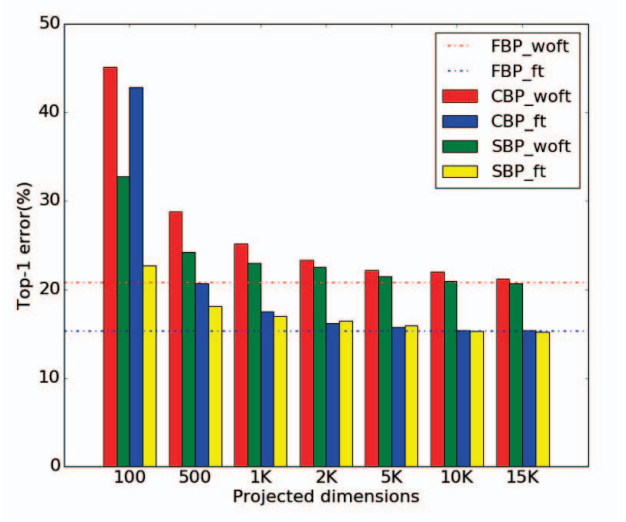

Figure 2. Classification error rate on the Cub dataset. Comparisons are made on the proposed SBP without matrix normalization and Compact Bilinear Polling (CBP) with Tensor Sketch. Horizontal lines are the baseline performance of Fully Bilinear Pooling (FBP). ft and woft stands for with and without global fine-tuning of CNN, respectively.

makes sense for a vast area of applications, for example, large scale fine-grained image retrieval [19] often needs quick, high representative but low dimensional image features.

3.2. Experiments with different datasets

We compared the proposed squeezed bilinear pooling against other approaches for the categorization of the following FGVC datasets: CUB-200-2011[18], FGVC-Aircrafts[13], and Stanford Cars[9]. The experimental results are summarized in Table 2. Note that for fairness, we only compare the results reported with the backbone of VGG-16. From Table 2, we can see that the fast version of the proposed SBP, SBP-EN obtained the best accuracy with planes and cars dataset, and achieved 84.6% in CUB-200-2011, only 0.4% lower than the recent proposed MoNet[7]. However, the computation of our SBP is only 24% of that required for the Tensor Sketching in MoNet. Comparing with CBP[5] and LRBP[8] with the same dimension, the accuracy of the proposed SBP-EN is around 0.5% higher on average for the 3 datasets.

In the other aspect, the matrix square root is not directly applicable to CBP[5] and LRBP[8], hence, the comparison of the compressed structures with matrix normalization is held between the SBP-MN and Tensor Sketching MoNet with and without the first-order information. SBP-MN without the first-order information outperforms MoNet_TS[7] by 0.4% to 1.0% in the three fine-grained datasets. When comparing with MoNet_2_TS[7], the classification accuracy of the proposed SBP-MN is 0.4% to 2.4% higher. To the best of our knowledge, the performance of our SBP-MN model is state-of-the-art among all compressed bilinear models with these datasets.

Method	Dim.	Mul.	Cub	Air	Car
BCNN[6]	262K	205M	84.0	86.9	90.6
iBCNN[11]	262K	205M	85.8	88.5	92.0
G2DeNet[20]	263K	206M	87.1	89.0	92.5
MoNet[7]	263K	206M	86.4	89.3	91.8
CBP_TS[5]	8.2K	105M	84.0	87.2	90.2
LRBP[8]	10K	48M	84.2	87.3	90.9
SMSO[22]	2K	—	**85.0**	—	—
MoNet_2U_TS[7]	10K	105M	**85.0**	86.1	89.5
SBP-EN	10K	**7.8M**	84.6	**87.8**	**90.9**
MoNet_2_TS[7]	10K	105M	85.7	86.7	90.3
MoNet_TS[7]	10K	105M	85.7	88.1	90.8
SBP-MN	10K	205M	**86.1**	**89.2**	**91.6**
FRA-SBP	20K	—	**86.8**	**90.4**	**93.2**
KP[1]	14.3K	420M	86.2	86.9	92.4
BoostCNN[14]	—	—	86.2	88.5	92.1
PC[3]	—	—	85.6	85.8	92.5
iSQRT-COV[10]	—	—	**87.2**	90.0	92.5
MA-CNN[23]	—	—	86.5	89.9	92.8

Table 2. Comparison of the classification performance on various FGVC datasets. From top to bottom, the four blocks respectively list fully bilinear based methods, compressed bilinear methods, compressed structures with matrix normalization, and other state-of-the-art methods. Dim. and Mul. stand for feature dimension and multiplies required for pooling, respectively.

The accuracy of the proposed FRA-SBP is 0.4% lower than the state-of-the-art fine-grained model, iSQRT-COV[10] on CUB-200-2011 dataset, but around 0.5% higher in the other two datasets. Note that iSQRT-COV needs to pre-train on ImageNet[2], while the FRA-SBP model achieved the overall better accuracy with a transferred model. Comparing with other duplicate or recurrent models, e.g. BoostCNN[14], KP[1] and MA-CNN[23], the accuracy of the FRA-SBP is 0.2% to 3.5% higher for the three datasets.

4. Conclusion

We presented a novel Squeezed Bilinear Pooling (SBP) network to solve the problem of extremely high feature dimension of bilinear pooling, and obtained the state-of-the-art results using VGG as backbone. Our model outperforms other compressed bilinear models in terms of classification accuracy and computation complexity. This is a promising step for the second-order pooling towards the replacement of global average pooling in other deep structures, e.g. ResNet, Inception, and DenseNet.

References

[1] Y. Cui, F. Zhou, J. Wang, X. Liu, Y. Lin, and S. Belongie. Kernel pooling for convolutional neural networks. In *CVPR*, volume 1, page 7, 2017.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[3] A. Dubey, O. Gupta, P. Guo, R. Raskar, R. Farrell, and N. Naik. Pairwise confusion for fine-grained visual classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 70–86, 2018.

[4] K. Fukunaga. *Introduction to statistical pattern recognition*. Elsevier, 2013.

[5] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell. Compact bilinear pooling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 317–326, 2016.

[6] E. Gavves, B. Fernando, C. G. Snoek, A. W. Smeulders, and T. Tuytelaars. Local alignments for fine-grained categorization. *International Journal of Computer Vision*, 111(2):191–212, 2015.

[7] M. Gou, F. Xiong, O. Camps, and M. Sznaier. Monet: Moments embedding network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3175–3183, 2018.

[8] S. Kong and C. Fowlkes. Low-rank bilinear pooling for fine-grained classification. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 7025–7034. IEEE, 2017.

[9] J. Krause, M. Stark, J. Deng, and L. Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 554–561, 2013.

[10] P. Li, J. Xie, Q. Wang, and Z. Gao. Towards faster training of global covariance pooling networks by iterative matrix square root normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 947–955, 2018.

[11] T.-Y. Lin and S. Maji. Improved bilinear pooling with cnns. 2017.

[12] T.-Y. Lin, A. RoyChowdhury, and S. Maji. Bilinear cnn models for fine-grained visual recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1449–1457, 2015.

[13] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

[14] M. Moghimi, S. J. Belongie, M. J. Saberian, J. Yang, N. Vasconcelos, and L.-J. Li. Boosted convolutional neural networks. In *BMVC*, 2016.

[15] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[16] V. Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

[17] V. N. Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

[18] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[19] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu. Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1386–1393, 2014.

[20] Q. Wang, P. Li, and L. Zhang. G2denet: Global gaussian distribution embedding network and its application to visual recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 3, 2017.

[21] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, and S. Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1568–1576, 2017.

[22] K. Yu and M. Salzmann. Statistically-motivated second-order pooling. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[23] H. Zheng, J. Fu, T. Mei, and J. Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In *Int. Conf. on Computer Vision*, volume 6, 2017.