EXTENDING HASHING TOWARDS FAST RE-IDENTIFICATION

Meihan Liu[§†] *Yongxing Dai*[⋆] *Shengsen Wu*[§] *Yan Bai*[⋆] *Ling-Yu Duan* [⋆†]

§ The SECE of Shenzhen Graduate School, Peking University, Shenzhen, China
⋆ Institute of Digital Media, Peking University, China
† Peng Cheng Laboratory, Shenzhen, China

ABSTRACT

Searching accuracy and efficiency are two challenges in person and vehicle Re-identification (Re-ID), which one focuses on robust representations learning which usually generating high-dimensional features while the other has not been fully explored. Hashing is a suitable solution to make RE-ID efficient. However, directly extending the existing hashing methods to fast Re-ID faces two challenges: one is the non-overlap between training and testing set which need more discriminative hash codes, the other is the large identities in Re-ID tasks which will lead to slow convergence and hard optimization. In this work, we propose an attention pooling operator to exploit both local and global visual attributes which can break limited discriminative power in hash methods. To further make training procedure converge faster and optimize the network more easily, we substitute non-differentiable l_1-regularization with smooth l_1-regularization. In experiments, our work outperforms state-of-the-art hashing and quantization methods on both person and vehicle Re-ID datasets. Besides, the results can serve as a strong baseline in the field of deep hashing for fast Re-ID.[1]

Index Terms— Hashing, Re-Identification, Pooling

1. INTRODUCTION

Person and vehicle re-identification (Re-ID) have attracted more and more attention in recent years. Given a query image of person or vehicle, the goal of Re-ID is to search the images of the same person or vehicle under different cameras. There have been many research works focused on solving the challenges of Re-ID [1], including pose changes, occlusions and background clutter. However, most of them are dedicated to learn robust representation to improve the accuracy of Re-ID but ignore the searching efficiency. Hashing is a powerful tool for fast Re-ID [2]. Existing hashing works about fast Re-ID focus on small datasets and show the poor performance, which do not exploit the discriminative information of the raw images to learn binary codes.

Hashing is one of the solutions of Approximate Nearest Neighbor (ANN) search, which aims to convert high-dimensional float features into compact hamming codes and meanwhile preserve the similarity. Compared with quantization, which is another solution of ANN, hashing achieves more fast speed with simple bit-wise operation. Hashing methods mainly contains two parts: unsupervised hashing and supervised hashing. For more details, please refer to the survey [3]. Deep supervised hashing has achieved great success in instance level image retrieval [4]. Deep neural network can learn compact binary code in an end-to-end way by preserving similarity and minimizing the quantization error.

However, the existing deep hashing protocols concentrate on image classification datasets with less categories while deep hashing methods for fast Re-ID have been less explored. As mentioned in [5], they fail to capture desirable properties of supervised hashing schemes, since the testing identities and training identities of Re-ID do not overlap. This means the hash codes for fast Re-ID must be more discriminative. On the other hand, the existing non-smooth quantization regularization between the activated binary codes and their real-valued counterparts leads to slow convergence and hard optimization since there are large identities in Re-ID. To solve above challenges, we propose a plug and play hashing module for fast Re-ID, whose input is feature maps produced by backbone.

The discriminative information of binary codes heavily relies on the continuous features which are refined by pooling layer from feature maps. Hence, in this work, we propose a novel pooling method, which is used to promote discrimination of hash codes for fast Re-ID. By fusing local and global visual attributes with attention mechanism, the more discriminative hash codes can be obtained. In practice, the attention pooling operator combines global max pooling (GMP) and global average pooling (GAP) to exploit both local and global information for person or vehicle. Besides, we substitute non-differentiable l_1-regularization with smooth l_1-regularization to speed up the convergence of training procedure and make optimization easier.

Overall, the main contributions of our paper can be summarized as follows: (i) We introduce a novel attention pooling operator combined with GAP and GMP to generate discriminative hashing codes for person or vehicle images. (ii) We substitute non-differentiable l_1-regularization with smooth l_1-

[1] Code is available at https://github.com/cynthia951031/Hashing_ReID

Authorized licensed use limited to: Peking University. Downloaded on December 31,2020 at 08:58:13 UTC from IEEE Xplore. Restrictions apply.

regularization to make training process more stable and converge faster under the supervision of many identities. (iii) We provide a strong baseline of deep hashing on Re-ID for the field of fast Re-ID.

2. METHODOLOGY

In this section, we introduce our proposed methods in details. The detailed pipeline is depicted in Fig.1.

2.1. Attention Pooling

Given an input image I, pre-trained backbone outputs a 3D feature maps X of $W \times H \times C$, where C is the channel number of feature maps. In general, a compact image representations $P \in \mathbb{R}^C$ are constructed from these 3D feature maps by a global pooling operation with dimensionality equivalent. In the task of classification, most works propose to use global max pooling(GMP), which retains max activation per channel. GMP can be written as Eq.1, where C_j means j^{th} channel.

$$P_{max}(C_j) = \max_{m=0,\dots,W} \max_{n=0,\dots,H} X(C_j, m, n) \quad (1)$$

As shown in the left of Fig.1, each dimension of compact representation produced by GMP corresponds to an image patch with the limitation of receptive field, since the neurons in network cannot abstract the whole image. Relatively, global average pooling(GAP) (Eq.2) retains global universal attributes of original images as shown in the middle of Fig.1. It is obvious that, max pooling is more concerned about local visual feature, while average pooling gather global visual feature.

$$P_{avg}(C_j) = \frac{1}{H \times W} \sum_{0<m\leq W, 0<n\leq H} X(C_j, m, n) \quad (2)$$

However, most existing works select fixed hand-tuned global pooling operator which restrict network to find discriminative feature either locally or globally. Hence, there are some works preform a hybrid scheme which explore better way of global pooling. R-MAC [6] performs GMP over regions and finally sum pooling of the regional descriptors. GeM(generalizaed-mean pooling) [7] select global pooling method by control pooling parameter. But they cannot fuse local and global signal in self-inspired fashion.

Here, to encode the most distinguished signal from feature maps, whether it is local or global under the supervision of regualrization term, we propose an attention pooling mechanism, which apply attention mechanism to weight and fuse two different global pooling operators. It can be formed as Eq.3.

$$P_{attn}(C_j) = W(C_j)P_{max}(C_j) + W(C_j)P_{avg}(C_j) \quad (3)$$

The weights set $W_{avg} \in \mathbb{R}^C$ and $W_{max} \in \mathbb{R}^C$ are learnable weights corresponding to GMP and GAP respectively. We normalize them to $W \in \mathbb{R}^{2\times C}$ along each channel.

2.2. Loss Functions

2.2.1. supervised loss functions

Most methods apply cross entropy function as supervised loss, which requires an additional linear layer appended to the end of backbone and vector $s \in \mathbb{R}^C$ which indicate the confidence level of prediction for C classes. For given score vector s, cross entropy function can be written as Eq.4.

$$L_{CE} = -log(\frac{exp(s[class])}{\sum_j exp(s[j])}) \quad (4)$$

However, in Re-ID datasets, the number of identity is too large. For example, there are 30671 classes in VERI-Wild [8] train set which leads to 30671 neurons in last linear layer and this will lead to bad results for it's difficult to optimize.Hence, we replenish cross entropy loss with triplet loss [9].

As a complement to cross entropy loss, triplet loss exploits better embeddings instead of learning limited predictions for each sample. As shown in formulation Eq.5, $[a, p, n]$ constitute triplet for triplet loss, where a and p are images of the same identity while n is an image of a different identity relatively. The core idea of triplet loss is to close the distance between $[a, p]$ and pull the distance between $[a, p]$ in embedding space. Therefore, triplet loss will not be limited by scanty categories in train set when classes in test set are "unseen".

$$L_T(a, p, n) = \max\{d(a, p) - d(a, n) + margin, 0\} \quad (5)$$

In experiment, we apply hard sample mining in training stage. For each training batch, we follow [10], picking the most dissimilar sample with the same identity and most similar sample with a different identity to form a triplet.

2.2.2. smooth l_1-regularization

To make the features close enough to the final binary codes, we apply l_1-regularization to sparse features [11]. Besides, there is no clear bound of target binary codes, training with l_1-regularization is robust since it is less sensitive to outliers and can prevent exploding gradients.

However, as mentioned above, sparse l_1-norm regularization based hashing methods are difficult to convergence for l_1-norm regularization is convex but non-smooth, namely non-differentiable. Hence, our strategy is to substitutes the l_1-norm regular term with smooth l_1-norm regular term.

Authorized licensed use limited to: Peking University. Downloaded on December 31,2020 at 08:58:13 UTC from IEEE Xplore. Restrictions apply.

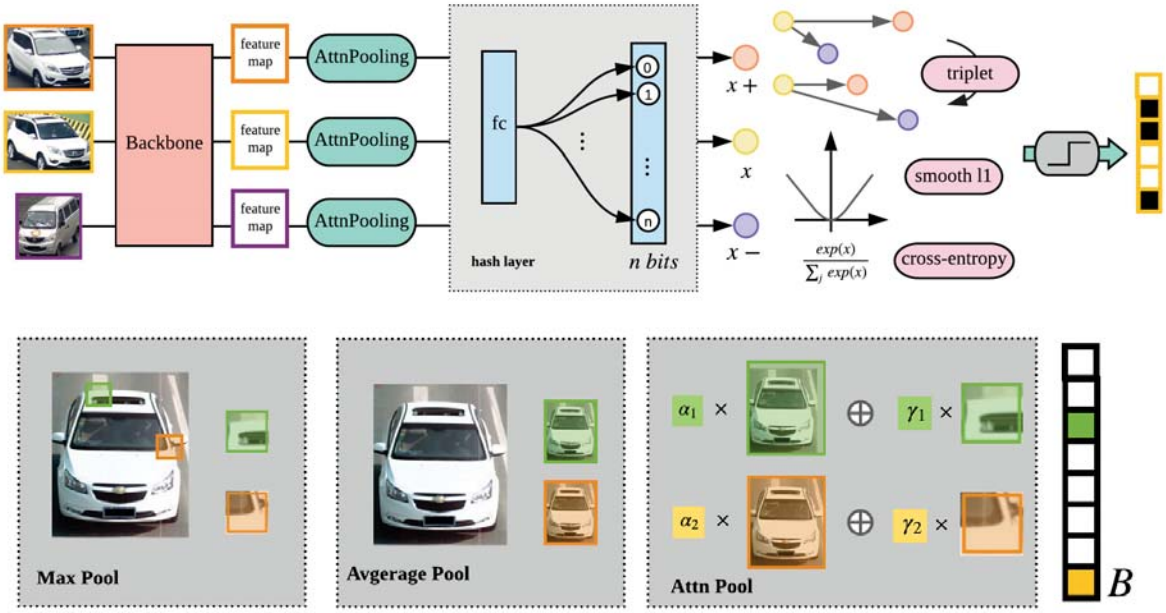

Fig. 1. Top: The pipeline of our method. The complete pipeline has two components: backbone which produces feature maps and our proposed module containing attention pooling layer, hash layer and supervised module in order. Among them, hash layer is composed by a linear layer which mapping original M dimension features to length of the target compressed codes. Bottom: Comparision between three different global pooling method. Green and yellow rectangles correspond to different channels in binary code B.

Smooth l_1-norm regular term is firstly used in object detectionfor bounding box regression. It is formulated as Eq.6, where replace l_1-regularization with smooth $0.5x^2$ when $|x|$ is less than 1.

$$R_{smoothl_1} = \begin{cases} 0.5x^2 & |x| < 1 \\ |x| - 0.5 & otherwise \end{cases} \tag{6}$$

To our knowledge, our work is the first method that uses this function for hashing and smooth l_1-regularization is so efficient compared to previous methods as shown in Sec.3 yet.

Finally, the loss function of our method is:

$$L = \beta * L_{CE} + \mu * L_T + \sigma * R_{smoothl_1}, \tag{7}$$

where our approach is affected by coefficients of loss functions and regularization. In experiment, we set σ as 1.61 while β and μ fixed to 1.

2.3. Binary Quantization and Distance Metric

The binary hash code h of the input feature maps X can be obtained by hard threshold quantization: $h = sgn(F(X, W))$, where F and W represents the network mapping and weights of proposed module. After that, towards fast Re-ID, we apply hamming distance to compute dissimilarity between two images.

3. EXPERIMENT

3.1. Datasets and Evaluation Protocol

We evaluate the proposed method on four large-scale vehicle and person Re-ID benchmarks: VehicleID [12], VERI-Wild [8], Market1501 [13] and MSMT17 [14] with 13136 identities, 30671 identities, 1501 identities, and 4101 identities respectively. The performance of our method is evaluated with mean average precision (mAP) and rank1 accuracy of cumulative matching characteristics (CMC) in Tab.1 and Tab.2 under different hashing code length: 64, 128, 256, 512 and 1024 bits.

3.2. Experiment Settings

3.2.1. Comparative Methods

We compare our method with supervised deep hashing method: DSH [15], DBH [16], DCH [17], DHN [18] and unsupervised hashing method: KMH [19], ITQ [20]. We evaluate these methods with source code provided by their authors on four Re-ID datasets. Besides, for the dataset Market1501, we copy the results of PDH [2] from [2]. ITQ and KMH will not produce codes that longer than input features.

Authorized licensed use limited to: Peking University. Downloaded on December 31,2020 at 08:58:13 UTC from IEEE Xplore. Restrictions apply.

	VehicleID										VERI-Wild									
	64		128		256		512		1024		64		128		256		512		1024	
	mAP	r=1	mAP	r=1	mAP	r=1	mAP	r=1	mAP	r=1	mAP	r=1	mAP	r=1	mAP	r=1	mAP	r=1	mAP	r=1
Ours	**55.66**	**50.07**	**64.24**	**57.67**	**69.03**	**62.50**	**71.07**	**64.85**	**72.28**	**65.89**	**45.37**	**68.29**	**55.34**	**76.81**	**60.25**	**80.72**	**62.33**	**82.15**	**63.13**	**82.88**
DBH	47.14	40.68	56.42	49.60	64.03	56.81	68.43	61.93	70.62	63.79	32.92	38.17	36.89	42.77	41.65	50.28	47.80	61.06	54.92	72.82
DSH	53.86	47.84	56.90	50.08	56.97	50.29	54.93	48.41	55.53	47.96	40.82	56.25	42.58	54.82	36.79	43.04	33.62	38.88	19.53	22.73
DCH	42.32	36.76	48.77	42.25	54.17	47.89	56.29	49.22	57.93	50.74	40.91	60.19	50.53	70.30	55.64	75.44	58.48	78.77	60.07	80.70
DHN	54.17	49.99	62.85	57.16	67.18	61.73	69.76	62.59	71.09	64.49	41.65	60.23	50.85	70.73	55.75	76.46	58.54	79.20	59.95	80.80
KMH	37.23	33.76	57.98	52.69	-	-	-	-	-	-	29.67	49.93	43.48	63.75	51.79	72.09	-	-	-	-
ITQ	52.87	47.10	60.73	53.35	-	-	-	-	-	-	39.04	58.62	51.21	69.80	55.36	74.71	57.43	77.39	58.78	79.69

Table 1. This table shows experimental CMC top1 and mAP@all for retrieval of protocols VehicleID test2400 and VERI-Wild test10k. The results of the other four protocols of vehicle datasets can be found in supplementary material. "r1" in table represents CMC top1 and "mAP" represents mAP@all.

	Market1501										MSMT17									
	64		128		256		512		1024		64		128		256		512		1024	
	mAP	r=1	mAP	r=1	mAP	r=1	mAP	r=1	mAP	r=1	mAP	r=1	mAP	r=1	mAP	r=1	mAP	r=1	mAP	r=1
Ours	**58.43**	**77.40**	**68.77**	**86.02**	**73.46**	**88.12**	75.78	88.75	77.02	88.93	24.02	**48.92**	32.55	**58.63**	**38.13**	**64.89**	**41.31**	67.41	42.85	**68.74**
DBH	39.76	62.50	51.77	75.15	61.15	81.18	67.05	84.98	70.09	86.25	10.41	25.50	16.25	35.82	21.79	44.99	26.18	51.19	30.44	56.32
DSH	42.70	63.45	33.76	54.19	20.06	40.47	18.68	37.29	17.36	32.18	4.62	7.76	5.50	10.96	5.43	11.90	4.82	11.21	5.76	13.77
DCH	51.20	72.92	65.67	83.52	71.45	86.79	75.52	88.78	77.15	89.58	18.67	41.18	28.85	54.30	36.71	62.72	41.27	66.67	**43.50**	68.68
DHN	58.20	78.38	68.13	85.54	73.31	87.83	**75.93**	**89.16**	**77.34**	**89.73**	23.32	48.91	31.71	58.05	37.27	64.41	40.62	**67.54**	42.33	68.40
KMH	40.04	66.39	66.65	83.70	-	-	-	-	-	-	9.46	29.22	27.09	52.43	-	-	-	-	-	-
ITQ	58.96	78.09	68.18	83.97	-	-	-	-	-	-	**24.98**	48.09	**33.18**	58.50	-	-	-	-	-	-
PDH	-	-	19.59	36.31	22.43	42.07	24.30	44.60	26.09	49.58	-	-	-	-	-	-	-	-	-	-

Table 2. Experimental CMC top1 and mAP@all for retrieval on Re-ID person datasets, including Market1501 and MSMT17.

	mAP	r=1
VehicleID 128-dimensional features	77.52	71.75
VERI-Wild 2048-dimensional features	69.49	86.67
Market1501 128-dimensioanl features	79.13	90.65
MSMT17 128-dimensional features	46.83	71.16

Table 3. Experimental CMC top1 and mAP@all for retrieval of four datasets with float features. This table shows the average results of 3 protocols on VehicleID and VERI-Wild.

3.2.2. Implementation Details

When we reproduce these comparative methods, we obtain 2048-dimensional features of VERI-Wild and 128-dimensional features of the other three datasets from 2048-channel feature maps produced by pre-tarined backbone ResNet50 [21] and other essential layers.

3.3. Results and Analysis

Experimental mAP results on VehicleID and VERI-Wild show that our module outperforms existing best retrieval performance by at most 3.51% and 4.5% respectively on vary test bits. It should be noted that, VERI-Wild is a more challenging dataset compared to VehicleID, and our method outperforms state-of-the-art methods by a large gap. For dataset Market1501, DHN show the best performance on 512 and 1024 bits while our method outperforms the comparative methods on 64, 128, 256 test bits. For the test bit 512 and 1024, our method only differ the best CMC top1 by at most 0.80% and differ the best mAP@all by at most 0.32%. For the dataset MSMT17, our method achieve best mAP on 256 and 512 bits and differ state-of-the-art mAP on 64 bits, 128 bits, 1024 bits by at most 0.96%. Also, we achieve best CMC top1 on most test bits expect 512 bits on where our method differ by 0.13%. However, the results show that directly extending existing deep hashing methods to fast Re-ID can achieve better performance compare to specially designed PDH. Moreover, we evaluate retrieval performance of float features which are produced by pre-trained ResNet-50 backbone, and the results are shown in Tab.3. Significantly, our method show the better performance compared to float features on VERI-Wild dataset.

4. CONCLUSION

In this paper, we propose a novel deep hashing network for fast Re-ID. First, we design an attention module which combine global max pooling and global average pooling to fully utilize the local and global information of person or vehicle images.Second, a smooth l_1 quantization loss is adopted to minimize the quantization error when converting float features into binary codes, which can speed up convergence in training stage as well. Experimental results on four large-scale Re-ID benchmarks have shown the superiority of the proposed method.

5. ACKNOWLEDGEMENT

This work was supported by National Natural Science Foundation of China under Grant U1611461.

Authorized licensed use limited to: Peking University. Downloaded on December 31,2020 at 08:58:13 UTC from IEEE Xplore. Restrictions apply.

6. REFERENCES

[1] Liang Zheng, Yi Yang, and Alexander G Hauptmann, "Person re-identification: Past, present and future," *arXiv preprint arXiv:1610.02984*, 2016.

[2] Fuqing Zhu, Xiangwei Kong, Liang Zheng, Haiyan Fu, and Qi Tian, "Part-based deep hashing for large-scale person re-identification," *IEEE Transactions on Image Processing*, vol. 26, no. 10, pp. 4806–4817, 2017.

[3] Jingdong Wang, Ting Zhang, Nicu Sebe, Heng Tao Shen, et al., "A survey on learning to hash," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 769–790, 2017.

[4] Venice Erin Liong, Jiwen Lu, Gang Wang, Pierre Moulin, and Jie Zhou, "Deep hashing for compact binary codes learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2475–2483.

[5] Alexandre Sablayrolles, Matthijs Douze, Nicolas Usunier, and Hervé Jégou, "How should we evaluate supervised hashing?," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 1732–1736.

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

[7] Filip Radenović, Giorgos Tolias, and Ondřej Chum, "Fine-tuning cnn image retrieval with no human annotation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 7, pp. 1655–1668, 2018.

[8] Yihang Lou, Yan Bai, Jun Liu, Shiqi Wang, and Lingyu Duan, "Veri-wild: A large dataset and a new method for vehicle re-identification in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3235–3243.

[9] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[10] Alexander Hermans, Lucas Beyer, and Bastian Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[11] Zhixiang Chen, Jiwen Lu, Jianjiang Feng, and Jie Zhou, "Nonlinear sparse hashing," *IEEE Transactions on Multimedia*, vol. 19, no. 9, pp. 1996–2009, 2017.

[12] Hongye Liu, Yonghong Tian, Yaowei Wang, Lu Pang, and Tiejun Huang, "Deep relative distance learning: Tell the difference between similar vehicles," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2167–2175.

[13] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian, "Scalable person re-identification: A benchmark," in *Computer Vision, IEEE International Conference on*, 2015.

[14] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian, "Person transfer gan to bridge domain gap for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 79–88.

[15] Haomiao Liu, Ruiping Wang, Shiguang Shan, and Xilin Chen, "Deep supervised hashing for fast image retrieval," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2064–2072.

[16] Huei-Fang Yang, Kevin Lin, and Chu-Song Chen, "Supervised learning of semantics-preserving hash via deep convolutional neural networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 2, pp. 437–451, 2017.

[17] Yue Cao, Mingsheng Long, Bin Liu, and Jianmin Wang, "Deep cauchy hashing for hamming space retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1229–1237.

[18] Han Zhu, Mingsheng Long, Jianmin Wang, and Yue Cao, "Deep hashing network for efficient similarity retrieval," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[19] Kaiming He, Fang Wen, and Jian Sun, "K-means hashing: An affinity-preserving quantization method for learning binary compact codes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 2938–2945.

[20] Yunchao Gong, Svetlana Lazebnik, Albert Gordo, and Florent Perronnin, "Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2916–2929, 2012.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

Authorized licensed use limited to: Peking University. Downloaded on December 31,2020 at 08:58:13 UTC from IEEE Xplore. Restrictions apply.