Part-based Fusion Feature Learning for Person Re-Identification

Titipakorn Prakayaphun
Information, Computer, and Communication Technology
Sirindhorn International Institute of Technology
Prathumthani, Thailand
titipakorn.p@gmail.com

Sasiporn Usanavasin
Information, Computer, and Communication Technology
Sirindhorn International Institute of Technology
Prathumthani, Thailand
sasiporn.us@siit.tu.ac.th

***Abstract*—Person re-identification is the task to recognize the same person in gallery images from different cameras. Many previous researches aim to improve feature representation to separate different persons but features are only focus on local parts and a body part. Therefore, in this paper, we propose the Part-based Fusion Network (PFN) that extracted two global features from two layers of the ResNet50, split one global feature to form part-based features, and utilized both local and global features and concatenated to be a final feature for discriminating the same person. In addition, we combine the visual feature with the spatial temporal information to gain the better result on the testing phase. The experiment result shows that our method gained essential improvement and outperformed other state-of-the-art algorithms on two public datasets which are Market-1501 and DukeMTMC-reID.**

***Index Terms*—Person Re-Identification, Image Classification, Metric Learning, and Deep Learning.**

I. Introduction

Person re-identification (re-ID) is a retrieval task for searching an interesting person from multiple images in galleries. The camera caused many variations to a person image such as brightness, background, and angles. By using many cameras, an occlusion and illumination have been occurred in different scenes. In the early approaches, the handcraft-features are built from the observation of images, the results failed to discriminate the same person with different view-poses. The more powerful methods adopt Deep Convolutional Neural Network(CNN) to extract features as a feature representation. With the transfer learning technique, the model is able to learn the low-level features from the whole body. Partial information around local regions is essential to distinguish persons. For example, a person who carries a bag on the back is different from the one without it.

In order to overcome this problem, some algorithms in [1]–[4] are proposed to extract semantic partitions and learn part-level features. Some methods leverage the external information such as human pose estimation to locate the semantic parts. However, occlusion and various poses are affected to local information. Opposed to the part-based learning methods, we argued that combining the global feature with the local features will ensemble discriminative information.

A network named Part-based Fusion Network (PFN) is introduced to learn granularity information which defines the whole image as the global information and uniformly partitioned stripes as the local part features. PFN takes the input image through the backbone network with some modifications. Local parts are horizontally split in the equal portion. By increasing the number of stripes, the features become more fine-grained. From the 4th stage of the ResNet-50 backbone, the global feature is concatenated to the last layer output to balance the impact between local and global features.

To achieve the better performance, we utilize the spatial-temporal operation to boost up the accuracy. The performance of the Market-1501 dataset is increased from 95.01% rank-1 accuracy and 87.41% mAP to 98.09% (+3.08%) rank-1 accuracy and 89.09% (+1.68%), which is superior compared to other state-of-the-art methods by a significant number.

II. Related Work

A. Visual features

In early years, handcrafted features had been developed to capture visual information. In [5], a whole image is divided into stripes to exploit color and texture patterns. [6] apply HSV histograms on parts to extract spatial information. Deep CNN methods bring the new standard for generic object classification such as ImageNet 1000 objects [7]. Several recent works employ deep learning methods to learn fine-grained information that represented a person. [2], [8], [9] learn the local parts such as head, torso, and legs using attention model. [1], [10], [11] use local cues to extract body parts from the image by extracting features and concatenating them. Multiple-level features are extracted at different layers and combined in [4], [12]–[14]. [15] use multiple-scale features from every layer of the networks. [16] use multiple feature from branches to learn fine-grained features for both global and local parts. [17] use GAN to obtain more data for training original data and generated data to improve the model robustness.

B. Attributes

The hybrid deep network are introduced by using multiple networks to extract individually features and integrate to metric learning methods in [18]. For instance, the network is trained to classify whether a person or not, female or male, and person classes. The network is trained separately not end-to-end learning and may not correlate the attributes and ID

classes. The attribute is used as supervision for unsupervised learning such as [19] leverage the model trained from labeled source data and transfer the knowledge to unlabeled data by using a joint attribute across domains. In [20], the attribute is also used as the query to retrieve the same person. [21] use the visual feature concatenated with attribute feature to represent a person .

C. Spatial-temporal information

The spatial-temporal information is typically used to eliminate irrelevant candidates in [22]–[24]. For example, a person at the timestamp t would be appeared between $t - \Delta t$ and $t + \Delta t$. [22], [24] integrate spatial-temporal information into the visual feature representation.

D. Re-ranking

The Euclidean or cosine distances are computed to measure the similarity of people for object retrieval. The k-reciprocal nearest neighbors are the most relevant identities, and used to build a group for re-ranking others in the dataset. [25] calculate a new distance between two images by comparing their k-reciprocal nearest neighbors.

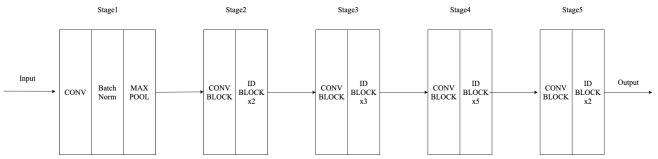

Fig. 1. The ResNet50 modification model.

III. Methodology

To obtain a robust feature, we chose the ResNet50 as the backbone network due to its performance. The ResNet50 is a network that trained on a million images from the ImageNet database. ResNet has 50 layers and 5 stages that have a different number of convolutional layers as shown Fig.1. The default image size is 224x244 for a single object, however, the human body is high in height. Therefore, we set the image size to be high in the height for smoothly semantic partitions.

A. The Architecture of PFN

The ResNet50 has been modified by removing the ReLU layer on the first stage to allow nonlinear features fed to the second stage. The structures after the fifth stage are also removed. The global average pooling layer is applied to the fourth stage and the fifth stage for extracting global features that represented overall parts. The output of the 5th stage is horizontally stripped into six parts by using a part-based average pooling layer as illustrated in Fig.2. Then, PFN employs a 1x1 convolution that followed with batch normalization and ReLu activation to reduce the feature dimension to 512-dim. Finally, each feature is fed to the batch normalization and a fully-connected layer.

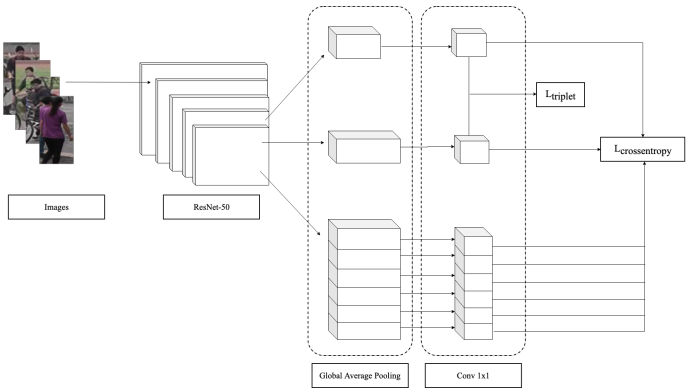

Fig. 2. Part-based Fusion Network architecture. The input image is input to the backbone network. The outputs of two layers are used to form global features. One of the outputs is split into six semantic partitions. A following 1 x 1 convolution reduces the feature dimension to lower computation complexity. Each classifier is deployed for each feature respectively. During training, global features are supervised by both Triplet loss and Cross-Entropy loss while local features are merely supervised by Cross-Entropy loss. During testing, all features are concatenated to form the final feature.

B. Loss Functions

To gain more discriminative features, we employ the cross-entropy loss proposed in [26] for classification and triplet loss for metric learning. The cross-entropy loss with label smoothing is formulated as:

$$-\frac{1}{P}\sum_{j=1}^{P}\sum_{i=1}^{N} q_{ij} log(p_{ij}), \quad \left\{ \begin{matrix} q_{ij} = 1 - \frac{N-1}{N}\epsilon, & y = i \\ q_{ij} = \epsilon / N, & y \neq i \end{matrix} \right\} \quad (1)$$

where p_{ij} is prediction logit of class i in part j, y is the truth label, P is the number of feature parts, N is the number of classes, and q_{ij} is the mask to prevent the overfitting model. In this paper, ϵ is set to be 0.1.

While the model is trained to classify the same person based on the feature, features are used to find the closest distance or the most similarity. To ensure that the same class is close to each others, the triplet loss is employed to boost up the ranking performance. In [27], batch-hard triplet loss is an improved version of the triplet loss that covered on the hardest negative and positive pairs. For instance, the different person that has similar appearance is the hard negative cases and vice versa. This loss function is formulated as:

$$\sum_{i=1}^{P}\sum_{j=1}^{K}\left[m + \max_{p=1\ldots K} D(f_j^i, f_p^i) - \min_{n=1\ldots K} D(f_j^i, f_n^{i_n})\right]_+ \quad (2)$$

which is defined for a mini-batch with P selected identities and K images for each identity, f_j^i,f_p^i,$f_n^{i_n}$ are the features from anchor, positive, and negative samples respectively, and m is the margin hyperparameter to control the differences between in-class and between-class distances. ReLU function is applied at the end and defined as $max(0, x)$.

When training the model, global features are optimized by the triplet loss to avoid the misalignment due to disconnected

information among local features. All features are supervised by the cross-entropy loss to enhance the feature ability.

IV. Experiment

A. Problem Definition

A pedestrian image cropped from a surveillance image is used to retrieve the same person from the images captured from different cameras. The traditional approach of person re-identification is to use a feature extracted from the image to calculate the similarity score among other features. The sorting order of the score will be used as the ranking result. To increase more accuracy, the additional data are needed to overcome the complexity.

B. Implementation

We set the image size to 384x192 for gaining more information on local parts. We adopt the pre-trained weight from the ResNet50 to initialize the network. On the training, input images are horizontally flipped, padded for cropping, and randomly erased for data augmentation. Each mini-batch is sampled with selected P identities and K images for each identity to be computed in the triplet loss. The recommend values is to set $P = 16$ and $K = 4$ to train the model. The margin parameter for the triplet loss is set to 0.3 in all experiments. We use Adam as the optimizer with the initial learning rate 0.00035, which is decreased by 0.1 at the 40th epoch and 70th epoch. The total training are 160 epochs. During testing, we extract the feature from a given image and the horizontally flipped image, then concatenate these as the final feature. Our model is implemented on PyTorch framework. All our experiments on different datasets follow the same settings.

C. Dataset and Protocol

a) Market-1501: The Market-1501 dataset in [28] are collected from six cameras in front of a supermarket which contains 32,668 labeled images of 1,501 identities. This dataset is divided into two parts: 12,936 images of 751 identities for training set and 3,368 query images and 19,732 gallery images of 750 identities for testing set.

b) DukeMTMC-reID: The DukeMTMC-reID is a subset of the DuckMTMC and used for image-based re-identification. This dataset contains 36,411 labeled images of 1,812 identities that collected from eight surveillance cameras. The DukeMTMC-reID also separated into training and testing set. The training set consists of 16,522 images of 702 identities, and 2,228 query images and 17,661 gallery images from the remaining 702 identities are the testing set.

c) Evaluation Protocol: The performance of our model is evaluated in term of Cumulative Matching Characteristic (CMC) curves which are the most popular evaluation metrics for person re-identification. We measure the ranking accuracy and mean Average Precision (mAP) on both datasets. The ranking accuracy is defined by ordering the smallest distance between the searching image and the gallery images. In addition, the rank-1 accuracy is the important factor to compare the performance of methods.

V. Result

We evaluate our method on two datasets and clearly observe the advantage of combining global and local features increased the discriminative feature. The performance of our model with the spatial temporal method gains significant improvement compared with the visual feature.

A. Evaluations on Market-1501

Market-1501 is a popular dataset for person re-identification approaches. We compare our model against several methods. When only the ResNet50 is used with the triplet loss, Triplet Loss [27] obtains 84.9% rank-1 accuracy and 68.1% mAP. Local features of 6 parts with a refine method in PCB+RPP [1] obtain 93.8% rank-1 accuracy and 77.4% mAP. BagofTricks [26] use only global feature of a model trained with tricks, and achieve 94.5% rank-1 accuracy and 85.9%. MGN [16] achieves 95.7% rank-1 accuracy and 86.9% mAP without a post-processing technique, however different scales are used to extract global and local features to concatenate as the final feature. With the spatial temporal scheme, our PFN obtains the rank-1 accuracy of 98.0% and mAP of 89.0%.

B. Evaluations on DukeMTMC-reID

DukeMTMC-reID is a challenge dataset and consists of 408 distractor identities in the testing set. We compare our method with [1], [9], [12], [14], [16], [17], [21], [24], [26], [28]–[33] state-of-the-art methods on the DukeMTMC-reID dataset. With the spatial temporal method, our PFN achieves the rank-1 accuracy of 94.2% and mAP of 83.9%. This dataset contains many distractors in the testing set and the result of combining global and local features seem to be inferior than local features.

Fig. 3. Visualization examples of our FPN for retrieving a query image across gallery images on the Martket-1501 dataset. The following images are ranked according to the similarity score. Red borders denote the wrong class.

VI. Discussion

Deep learning model is capable of extracting discriminative features but they are not represented meaningful features such as a gender, wearing a bag, and etc. To overcome appearance ambiguity, extra information is needed to cut off the candidates. By exploiting the given data of camera ID and timestamp, the probability distribution is formed to reorder the ranked results to the best possibility of the person from one camera to another.

TABLE I
COMPARISON OF THE PROPOSED METHOD ON MARKET-1501 WITH THE STATE-OF-ART METHODS. "RK" REFERS TO IMPLEMENTING RE-RANKING OPERATION. "ST" REFERS TO IMPLEMENTING SPATIAL-TEMPORAL METHOD. * DENOTES THE METHODS ARE REPRODUCED BY OURSELVES.

Methods	Rank-1	Rank-5	Rank-10	mAP
Bow+kissme [28]	44.4	63.9	72.2	20.8
KLFDA [34]	46.5	71.1	79.9	-
SVDNet [33]	82.3	92.3	95.2	62.1
PAN [29]	82.8	-	-	63.4
Triplet Loss [27]	84.9	94.2	-	69.1
HydraPlus [4]	76.9	91.3	94.5	-
PAR [11]	81.0	92.0	94.7	63.4
MultiLoss [35]	85.1	-	-	65.5
DuATM [9]	91.4	-	-	76.6
MultiScale [14]	88.9	-	-	73.1
GLAD [36]	89.9	-	-	73.9
HPM [13]	94.2	-	-	82.7
MFML [30]	92.5	-	-	89.3
APR [21]	87.0	95.1	96.4	66.8
PCB+RPP [1]	93.8	97.2	98.2	77.4
GAN [17]	83.9	-	-	66.0
Auto-ReID [3]	94.5	-	-	85.1
BagofTricks [26]	94.5	-	-	85.9
MLFN [12]	90.0	-	-	74.3
DeepCRF [31]	93.5	-	-	81.6
Mancs [32]	93.1	-	-	82.3
MGN [16]	95.7	-	-	86.9
OSNet [15]	94.8	-	-	84.9
Ours	95.0	98.1	98.8	87.4
PCB+ST* [24]	97.5	99.3	99.5	87.8
BagofTricks+RK [26]	95.4	-	-	**94.2**
MGN+RK [16]	96.6	-	-	**94.2**
Ours+ST	**98.0**	**99.3**	**99.7**	89.0

TABLE II
COMPARISON OF RESULTS ON DUKEMTMC-REID. * DENOTES AS THE RESULT REPRODUCED BY OURSELVES.

Methods	Rank-1	mAP
BoW+kissme [28]	25.1	12.2
GAN [17]	67.6	47.1
PAN [29]	71.6	51.5
APR [21]	73.9	55.5
MFML [30]	84.0	80.0
DeepCRF [31]	84.9	69.5
MLFN [12]	81.0	62.8
Mancs [32]	84.9	71.8
DuATM [9]	81.8	64.6
SVDNet [33]	76.7	56.8
MultiScale [14]	79.2	60.6
PCB+RPP [1]	83.3	69.2
BagofTricks [26]	86.4	76.4
MGN [16]	88.7	78.4
OSNet [15]	88.6	73.5
Ours	86.8	76.0
BagofTricks+RK [26]	90.3	**89.1**
PCB+ST* [24]	**94.4**	84.6
Ours+ST	94.2	83.9

Two datasets have the distractor class which is the unknown classes to lure the model. The ranking results are reviewed and apparently the distractor class is appeared to be in the testing class as shown in 3. This leads to the degrading performance of our model.

VII. CONCLUSION

In this paper, we have proposed PFN as a high-performance model combining local and global information to extract embedding representation of a person. Experiments show that our model outperforms other state-of-art algorithms. With external information, our method achieves rank-1 accuracy of 98.09% on Market-1501 and 94.25% on DukeMTMC-reID, improving from the baseline 95.01% and 86.80% respectively.

In the future work, we will use the augmentation methods to generate more robust data for classifying each augmented data. For example, the classifier can predict data for what rotation of the image taken [37] and what location of the original image [38]. Inspired from the self-supervised representation learning, the feature is fine-tune to perform better. Besides producing dataset with the label is expensive, this technique is practical in the real application.

ACKNOWLEDGMENT

This research is partially supported by DTAC under a collaborative research grant between SIIT and DTAC in Thailand, the Center of Excellence in Intelligent Informatics, Speech and Language Technology and Service Innovation (CILS) and by NRU grant at Sirindhorn International Institute of Technology (SIIT), Thammasat University Thailand.

REFERENCES

[1] Y. Sun, L. Zheng, Y. Yang, and S. Wang, "Beyond part models: Person retrieval with refined part pooling," 11 2017.

[2] H. Lawen, A. Ben-Cohen, M. Protter, I. Friedman, and L. Zelnik-Manor, "Attention Network Robustification for Person ReID," *arXiv e-prints*, p. arXiv:1910.07038, Oct 2019.

[3] R. Quan, X. Dong, Y. Wu, L. Zhu, and Y. Yang, "Auto-reid: Searching for a part-aware convnet for person re-identification," 03 2019.

[4] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, and X. Wang, "Hydraplus-net: Attentive deep features for pedestrian analysis," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 350–359, 2017.

[5] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," vol. 5302, 10 2008, pp. 262–275.

[6] A. Das, A. Chakraborty, and A. Roy-Chowdhury, "Consistent re-identification in a camera network," 09 2014.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 06 2016, pp. 770–778.

[8] J. Xu, R. Zhao, F. Zhu, H. Wang, and W. Ouyang, "Attention-aware compositional network for person re-identification," 06 2018, pp. 2119–2128.

[9] J. Si, H. Zhang, C.-G. Li, J. Kuen, X. Kong, A. Kot, and G. Wang, "Dual attention matching network for context-aware feature sequence based person re-identification," 03 2018.

[10] J. Liu, B. Ni, Y. Yan, P. Zhou, S. Cheng, and J. Hu, "Pose transferrable person re-identification," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2018, pp. 4099–4108.

[11] L. Zhao, X. Li, Y. Zhuang, and J. Wang, "Deeply-learned part-aligned representations for person re-identification," 10 2017, pp. 3239–3248.

[12] X. Chang, T. Hospedales, and T. Xiang, "Multi-level factorisation net for person re-identification," 03 2018.

[13] Y. Fu, Y. Wei, Y. Zhou, H. Shi, G. Huang, X. Wang, Z. Yao, and T. Huang, "Horizontal pyramid matching for person re-identification," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 04 2018.

[14] Y. Chen, X. Zhu, and S. Gong, "Person re-identification by deep learning multi-scale representations," in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, Oct 2017, pp. 2590–2600.

[15] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person re-identification," 05 2019.

[16] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *Proceedings of the 26th ACM International Conference on Multimedia*, ser. MM '18. New York, NY, USA: ACM, 2018, pp. 274–282. [Online]. Available: http://doi.acm.org/10.1145/3240508.3240552

[17] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 3774–3782, 2017.

[18] A. Franco and L. Oliveira, "Convolutional covariance features: Conception, integration and performance in person re-identification," *Pattern Recognition*, vol. 61, 07 2016.

[19] J. Wang, X. Zhu, S. Gong, and W. Li, "Transferable joint attribute-identity deep learning for unsupervised person re-identification," 06 2018, pp. 2275–2284.

[20] Z. Yin, W.-S. Zheng, A. Wu, H.-X. Yu, H. Wan, X. Guo, F. Huang, and J.-H. Lai, "Adversarial attribute-image person re-identification," in *IJCAI*, 2017.

[21] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu, C. Yan, and Y. Yang, "Improving person re-identification by attribute and identity learning," *Pattern Recognition*, vol. 95, pp. 151 – 161, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320319302377

[22] W. Huang, R. Hu, C. Liang, Y. Yu, Z. Wang, X. Zhong, and C. Zhang, "Camera network based person re-identification by leveraging spatial-temporal constraint and multiple cameras relations," 01 2016, pp. 174–186.

[23] Y.-J. Cho, S.-A. Kim, J.-H. Park, K. Lee, and K.-J. Yoon, "Joint person re-identification and camera network topology inference in multiple cameras," *Computer Vision and Image Understanding*, vol. 180, 10 2017.

[24] G. Wang, J. Lai, P. Huang, and X. Xie, "Spatial-Temporal Person Re-identification," *arXiv e-prints*, p. arXiv:1812.03282, Dec 2018.

[25] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," 07 2017, pp. 3652–3661.

[26] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bags of tricks and a strong baseline for deep person re-identification." 2019.

[27] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," 03 2017.

[28] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec 2015, pp. 1116–1124.

[29] Z. Zheng, L. Zheng, and Y. Yang, "Pedestrian alignment network for large-scale person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. PP, 07 2017.

[30] H. Wu, M. Xin, W. Fang, H. Hu, and Z. Hu, "Multi-level feature network with multi-loss for person re-identification," *IEEE Access*, vol. 7, pp. 91 052–91 062, 2019.

[31] D. Chen, D. Xu, H. Li, N. Sebe, and X. Wang, "Group consistent similarity learning via deep crf for person re-identification," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2018, pp. 8649–8658.

[32] C. Wang, Q. Zhang, C. Huang, W. Liu, and X. Wang, "Mancs: A multi-task attentional network with curriculum sampling for person re-identification," in *ECCV*, 2018.

[33] Y. Sun, L. Zheng, W. Deng, and S. Wang, "Svdnet for pedestrian retrieval," 03 2017.

[34] S. Karanam, M. Gou, Z. Wu, A. Rates-Borras, O. I. Camps, and R. J. Radke, "A comprehensive evaluation and benchmark for person re-identification: Features, metrics, and datasets," *ArXiv*, vol. abs/1605.09653, 2016.

[35] W. Li, X. Zhu, and S. Gong, "Person re-identification by deep joint learning of multi-loss classification," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, ser. IJCAI'17. AAAI Press, 2017, pp. 2194–2200. [Online]. Available: http://dl.acm.org/citation.cfm?id=3172077.3172193

[36] L. Wei, S. Zhang, H. Yao, and W. Gao, "Glad: Global-local-alignment descriptor for pedestrian retrieval," 09 2017.

[37] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised Representation Learning by Predicting Image Rotations," *arXiv e-prints*, p. arXiv:1803.07728, Mar 2018.

[38] C. Doersch, H. Mulam, and A. Efros, "Unsupervised visual representation learning by context prediction," 05 2015.