Learning Deep Features via Congenerous Cosine Loss for Person Recognition

Yu Liu¹*, Hongyang Li²* and Xiaogang Wang²
¹ SenseTime Group Ltd., Beijing, China
² The Chinese University of Hong Kong, New Territories, Hong Kong
liuyu@sensetime.com, {yangli, xgwang}@ee.cuhk.edu.hk

Abstract

Person recognition aims at recognizing the same identity across time and space with complicated scenes and similar appearance. In this paper, we propose a novel method to address this task by training a network to obtain robust and representative features. A key observation is that traditional cross entropy loss only enforces the inter-class variation among samples and ignores to narrow down the similarity within each category. We propose a congenerous cosine loss to enlarge the inter-class distinction as well as alleviate the inner-class variance. Such a design is achieved by minimizing the cosine distance between sample and its cluster centroid in a cooperative way. Our method differs from previous work in person recognition that we do not conduct a second training on the test subset and thus maintain a good generalization ability. The identity of a person is determined by measuring the similarity from several body regions in the reference set. Experimental results show that the proposed approach achieves better classification accuracy against previous state-of-the-arts.

1 Introduction

With an increasing demand of intelligent cellphones and digital cameras, people today take more photos to jot down daily life and stories. Such an overwhelming trend is generating a desperate demand for smart tools to recognize the same person (known as query), across different time and space, among thousands of images from personal data, social media or Internet. Previous work [Anguelov et al., 2007; Oh et al., 2016; Oh et al., 2015; Li et al., 2016c; Zhang et al., 2015; Li et al., 2016a] has demonstrated that person recognition in such unconstrained settings remains a challenging problem due to many factors, such as non-frontal faces, varying light and illumination, the variability in appearance, texture of identities, etc.

The recently proposed PIPA dataset [Zhang et al., 2015] contains thousands of images with complicated scenarios and similar appearance among persons. The illumination, scale and context of the data varies a lot and many instances have partial or even no faces. Figure 1 shows some samples from both the training and test sets. Previous work [Zhang et al., 2015; Oh et al., 2015; Li et al., 2016a] resort to identifying the same person via a multi-cue, multi-level manner where the training set is used only for extracting features and the follow-up classifier (SVM or neural network) is trained on the test set⁴. The recognition system is evaluated on the test set. We argue that such a practice is infeasible and ad hoc in realistic application since the second training on test set is auxiliary and needs re-training if new samples are added. Instead, we aim at providing a set of robust and well

* The first two authors contribute equally.

⁴ We denote this set as the reference set.
generalized feature representations, which is trained directly on the training set, and at identifying the person by measuring the feature similarity between two splits on test set. There is no need to train on test_0 and the system still performs well even if new data comes in.

As shown in the bottom of Figure 1, the woman in green box wears similar scarf as does the person in light green box. Her face is shown partially or does not appear in some cases. To obtain robust feature representations, we train several deep models for different regions and combine the similarity score of features from different regions to have the prediction of one’s identity. Our key observation is that during training, the cross-entropy loss does not guarantee the similarity among samples within a category. It magnifies the difference across classes and ignore the feature similarity of the same class. To this end, we propose a *congeneric cosine loss*, namely COCO, to enlarge the inter-class distinction as well as narrow down the inner-class variation. It is achieved by measuring the cosine distance between sample and its cluster centroid in a cooperative manner. Moreover, we also align each region patch to a pre-defined base location to further make samples within a category be more closer in the feature space. Such an alignment strategy could also make the network less prone to overfitting.

Figure 2 illustrates the training pipeline of our proposed algorithm at a glance. Each instance in the image is annotated with a ground truth head and we train a face and human body detector respectively, using the RPN framework [Ren et al., 2015] to detect these two regions. Then a human pose estimator [Wei et al., 2016] is applied to detect key parts of the person in order to localize the upper body region. After cropping four region patches, we conduct an affine transformation to align different patches from training samples to a ‘base’ location. Four deep models are trained separately on the PIPA training set using the COCO loss to obtain a set of robust features. To sum up, the contributions in this work are as follows:

- Propose a congeneric cosine loss to increase inner-class similarity as well as enlarge inter-class distinction in a cooperative way.
- Design a training pipeline that take advantage of several deep models to obtain robust features, without the necessity of conducting a second training on the test set.
- Align region patches to the base location via affine transformation, making the network less prone to overfitting and archiving better accuracy.

Our recognition model performs favourably against previous state-of-the-arts in terms of classification accuracy on the PIPA test set. The source code and results are available^2^.

2 Related Work

**Person recognition** in photo albums [Anguelov et al., 2007; Oh et al., 2016; Oh et al., 2015; Li et al., 2016c; Zhang et al., 2015; Li et al., 2016a] aims at recognizing the identity of people in daily life photos, where such scenarios can be complex with cluttered background. [Anguelov et al., 2007] first address the problem by proposing a Markov random field framework to combine all contextual cues to recognize the identity of persons. Recently, [Zhang et al., 2015] introduce a large-scale dataset for this task. They accumulate the cues of poselet-level person recognizer trained by a deep model to compensate for pose variations. In [Oh et al., 2015], a detailed analysis of different cues is explicitly investigated and three additional test splits are proposed for evaluation. [Li et al., 2016c] embed scene and relation contexts in LSTM and formulate person recognition as a sequence prediction task.

**Person re-identification** is to match pedestrian images from various perspectives in cameras for a typical time period and has led to many important applications in video [Li et al., 2014; Yi et al., 2014; Zhao et al., 2014; Tapaswi et al., 2012; Prosser et al., 2010]. Existing work employ metric learning and mid-level feature learning to address this problem. [Li et al., 2014] propose a deep network using pairs of people to encode photometric transformation. [Yi et al., 2014] incorporate a Siamese deep network to learn the similarity metric between pairs of images. The main difference between person recognition and re-identification resides in the data logistics. The former is to identify the same person across places and time. In most cases, the identity varies a lot in appearance under different occasions. The latter is to detect person in a consecutive video, meaning that the appearance and background do not vary much in terms of time.

**Deep neural networks** [Krizhevsky et al., 2012; He et al., 2016] have dramatically advanced the computer vision community in recent years, with high performance boost in tremendous tasks, such as image classification [He et al., 2016; Li et al., 2016b], object detection [Girshick, 2015; Li et al., 2017], object tracking [Chi et al., 2017], etc. The essence behind the success of deep learning resides in both its superior expression power of non-linear complexity in high-dimension space [Hinton and Salakhudinov, 2006] and large-scale datasets [Deng et al., 2009; Guo et al., 2016] where the deep networks could, in full extent, learn complicated patterns and representative features. In this work, we extend the ability of deep models to a higher level in order to make features of the human recognition task more robust and generalized, based on the state-of-the-art detector [Ren et al., 2015] to recognize face and body, and pose estimator [Wei et al., 2016] to predict key points of an identity.

3 Algorithm

3.1 Robust feature representation

The features of four regions \( r \in \{1, \cdots, 4\} \), namely, *face, head, whole body* and *upper body*, are utilized to identify a person. Each region follows a similar procedure of training and feature extraction.

We pre-train a face detector in a region proposal network (RPN) spirit following Faster RCNN [Ren et al., 2015]. The source of data comes from Internet and the number of images is roughly 300,000. The network structure is a shallow version of the ResNet model [He et al., 2016] where we remove layers after res\_3\_b and add two loss heads (classification and regression) after res\_3\_b. Then we finetune the

^2^ [https://github.com/sciencefans/coco_loss](https://github.com/sciencefans/coco_loss)
Figure 2: Training workflow of our algorithm. (a) Given an image, persons are labelled with a ground-truth box of the head. (b) For a typical training sample, we first detect face and body regions. (c) Pose estimator [Wei et al., 2016] is employed to identify keypoints of a human body in order to find the upper body region. We align each region (patch) to a base position to alleviate the inner-class variance. (d) Each aligned patch is further fed into a deep model to obtain representative and robust features f, where COCO loss is applied afterwards.

face model on PIPA training set. The face detector identifies m keypoints of the face (eye, brow, mouth, etc.) and we align the detected face patch to a ‘base’ shape via translation, rotation and scaling. Such a scheme is to ensure samples within each category do not have large variance and to prevent the model from overfitting. Let \( p, q \in \mathbb{R}^{m \times 2} \) denote m keypoints detected by the face model and the aligned results, respectively. We define \( P, Q \) as two affine spaces, then an affine transformation \( \mathcal{A} : P \rightarrow Q \) is defined as:

\[
p \mapsto q = \mathcal{A}p + b,
\]

where \( \mathcal{A} \in \mathbb{R}^{m \times m} \) is a linear transformation matrix in \( P \) and \( b \in \mathbb{R}^{m \times 2} \) being the bias in \( Q \). Note that in practice, if the confidence of a keypoint is below some threshold, we do not depend on such a point to align the patch; when the number of keypoints is less than three, we heuristically obtain the aligned patch based on the ground truth of the head, since at least three points can determine an affine transformation.

The head region is given as the ground truth for each person and the detection of face is stated previously. To detect a head, \( \sum \), \( \sum \delta \) and \( \sum \) are the labels of sample \( i, j \), where \( K \) is the total number of categories and \( W \) be the weights of a network, we have the loss as follows to maximize:

\[
L(B, W) = \sum_{i,j \in B} \frac{\delta(l_i, l_j)C(f^{(i)}, f^{(j)})}{(1 - \delta(l_i, l_j))C(f^{(i)}, f^{(j)}) + \epsilon},
\]

where \( \delta(\cdot, \cdot) \) is an indicator function and \( \epsilon \) is a trivial number for computation stability. Such a design is reasonable in theory and yet suffers from computational inefficiency. Since the complexity of the loss above is \( O(M^2) \), the loss increases quadratically as batch size \( M \) goes bigger. Also the network suffers from unstable parameter update and is hard to converge if we directly compute loss from two arbitrary samples from a mini-batch.

Inspired by the center loss [Wen et al., 2016], we define the centroid of class \( k \) as the average of features over a mini-batch \( B \):

\[
c_k = \frac{\sum_{i \in B} \delta(l_i, k)f^{(i)}}{\sum_{i \in B} \delta(l_i, k) + \epsilon} \in \mathbb{R}^D.
\]

Incorporating the spirit of Eqn. 3 with class centroid, we have the following output of sample \( i \) to maximize:

\[
p_{l_i}^{(i)} = \frac{\exp C(f^{(i)}, c_{l_i})}{\sum_{k \neq l_i} \exp C(f^{(i)}, c_k)} \in \mathbb{R}.
\]

The direct intuition behind Eqn. 5 is to measure the distance of one sample against other samples by way of a class centroid, instead of a direct pairwise comparison as in Eqn. 3. The numerator ensures sample \( i \) is close enough to its own class \( l_i \) and the denominator enforces a minimal distance against samples in other classes. The exponential operation
is to transfer the cosine similarity to a normalized probability output, ranging from 0 to 1.

To this end, we propose the congenericous cosine (COCO) loss, which is to increase similarity within classes and enlarge variation across categories in a cooperative way:

$$\arg \min_w \mathcal{L} = \arg \min_w - \sum_{i \in B} \log p^{(i)}_m. \quad (6)$$

In practice, COCO loss can be implemented in a neat way via the softmax operation. For Eqn. 5, if we constrain the feature and centroid to be normalized (i.e., $f = f/\|f\|$, $c = c/\|c\|$) and loose the summation in the denominator to include $k = l_i$, the probability output of sample $i$ becomes:

$$p^{(i)}_m = \frac{\exp(c^T_m \cdot f^{(i)})}{\sum_k \exp(c^T_k \cdot f^{(i)})} = \text{softmax}(c^T_m \cdot f^{(i)}), \quad (7)$$

where $m$ indexes along the class dimension in $\mathbb{R}^K$. Therefore, the normalized centroids $c_m$ can be seen as weights in the original classification layer before softmax with bias term being zero. The gradients with respect to input feature $f^{(i)}$ and the centroid $c_m$, in an element-wise form, are as follows:

$$z_k = c^T_k \cdot \hat{f},$$

$$\frac{\partial \mathcal{L}^{(i)}}{\partial f^d} = \frac{\partial \mathcal{L}^{(i)}}{\partial \hat{f}^d} \cdot \frac{\partial \hat{f}^d}{\partial f^d} = \sum_k \frac{\partial \mathcal{L}^{(i)}}{\partial z_k} \cdot \frac{\partial z_k}{\partial \hat{f}^d} \cdot \frac{\partial \hat{f}^d}{\partial f^d}, \quad (9)$$

$$\frac{\partial \mathcal{L}^{(i)}}{\partial c^d} = \frac{\partial \mathcal{L}^{(i)}}{\partial \hat{c}^d} \cdot \frac{\partial \hat{c}^d}{\partial c^d} = \sum_k \frac{\partial \mathcal{L}^{(i)}}{\partial z_k} \cdot \frac{\partial z_k}{\partial \hat{c}^d} \cdot \frac{\partial \hat{c}^d}{\partial c^d} + \frac{(p_k - t_k) \cdot \hat{f}^d}{\|\hat{f}\|}, \quad (10)$$

where $t_k$ is the vectorized mapping from label $l_i$. Note that the cluster centroids $c_m$ are learnable parameters and we only use Eqn. 4 to initialize the network.

### 3.3 COCO loss compared with counterparts

COCO loss is formulated as a metric learning approach in the feature space, using cluster centroid in the cosine distance as metric to both enlarge inter-class variation as well as narrow down inner-class distinction. It can be achieved via a softmax operation under several constraints. Figure 3 shows the visualization of feature clusters under different loss schemes. For softmax loss, it only enforces samples across categories to be far away while ignores the similarity within one class

---

1. We drop the sample index $i$ in $f^{(i)}$ for brevity and use $d$ (i.e., $f_d$) to index along the feature dimension.

Figure 3: Feature visualization using different losses, trained on MNIST [LeCun et al., 1998] with 10 classes. The softmax loss tries to choke up the feature space while COCO enlarges inter-class distance as well as alleviates inner-class variation. 3D cases in insets of (c) and (d) are more evident.

3(c)). In COCO loss, we replace the weights in the classification layer before softmax, with a clearly defined and learnable cluster centroid (3(d)). The center loss [Wen et al., 2016] is similar in some way to ours. However, it needs an external memory to store the center of classes and thus the computation is twice as ours (3(a-b)). The coupled cluster loss [Liu et al., 2016] is a further adaptation from the triplet loss [Wang and Gupta, 2015] where in one mini-batch, the positives of one class will get as far as possible from the negatives of other classes. It optimizes the inter-class distance in some sense but fails to differentiate among the negatives.

### 3.4 Inference

At testing stage, we measure the similarity of features between two test splits to recognize the identity of each instance in test 1 based on the labels in test 0. The similarity between two patches $i$ and $j$ from test 1 and test 0 is denoted by $s^{(r)}_{ij} = C(f^{(r)}_i, f^{(r)}_j)$, where $r$ indicates a specific region model. A key problem is how to merge the similarity scores from different regions. We first normalize the preliminary result $s^{(r)}_{ij}$ in order to have scores across different regions comparable, which is achieved by the logistic regression; then the final score $S_{ij}$ is a weighted mean of the normalized scores $\hat{s}^{(r)}_{ij}$ of each region:

$$\hat{s}^{(r)}_{ij} = \left(1 + \exp \left(-\beta_0 - \beta_1 s^{(r)}_{ij}\right) \right)^{-1}, \quad (11)$$

$$S_{ij} = \sum_{r=1}^R \gamma^r \cdot \hat{s}^{(r)}_{ij}, \quad (12)$$

where $\beta_0, \beta_1$ are parameters of the logistic regression; $R$ is the total number of regions and $\gamma^r$ being the weight of each
Figure 4: Visualization results of our method on the PIPA test set. Given the image as input to be predicted in (a), its nearest neighbour in the feature space from text 0 is shown in (b). The identity of the input is determined by the label of its neighbour. Our model can handle complex scenes with non-frontal faces and body occlusion. The last two columns show failure cases.

Table 1: Ablation study on body regions and feature alignment using softmax as loss. We report the classification accuracy (%) of the non-alignment and alignment cases in left and right column within each region category, respectively. Note that for the face model, we only evaluate instances with faces.

| Test split | Face  | Head  | Upper body | Whole body |
|------------|-------|-------|------------|------------|
| original   | 95.47 | 97.45 | 74.23      | 82.69      | 76.67      | 80.75      | 75.04      | 79.06      |
| album      | 94.66 | 96.57 | 65.47      | 73.77      | 66.23      | 69.58      | 64.21      | 67.27      |
| time       | 91.03 | 93.36 | 55.88      | 64.31      | 55.24      | 57.40      | 55.53      | 54.62      |
| day        | 90.36 | 91.32 | 35.27      | 44.24      | 26.49      | 32.09      | 32.85      | 29.59      |

Implementation details. For the face model, the initial learning rate is set to 0.001 and decreased by 10% after 20 epochs. For other three models, the initial learning rate is set to 0.005 and decreased by 20% after 10 epochs. The weight decay and momentum are 0.005 and 0.9 across models. We use stochastic gradient descent with the Adam optimizer. Note that during training, we resize the image for each person based on the longer dimension of the head to ensure that the scale of body regions for each instance is the same. Moreover, for the whole body model we do not apply the affine operation but simply use the similarity transformation (scale, transform, rotation) since the region is relatively large in the image.

4 Experiments

4.1 Setup

Dataset and evaluation metric. The People In Photo Albums (PIPA) dataset [Zhang et al., 2015] is adopted for evaluation. The PIPA dataset is divided into train, validation, test and leftover sets, where the head of each instance is annotated in all sets. The test set is split into two subsets, namely test 0 and test 1 with roughly the same number of instances. Such a division is called the original split. As did in [Oh et al., 2015; Li et al., 2016; Zhang et al., 2015; Li et al., 2016a], the training set is only used for learning feature representations; the recognition system is trained on test 0 and evaluated on test 1. In this work, as mentioned previously, we take full advantage of the training set and remove the second training on test 0. Moreover, [Oh et al., 2015] introduced three more challenging splits, including album, time and day. Each case emphasizes different temporal distance (different albums, events, days, etc.) between the two subsets. The evaluation metric is the averaged classification accuracy over all instances on test 1.

Feature alignment in different regions. Table 1 reports the performance of using feature alignment and different body regions, where several remarks could be observed. First, the alignment case in each region performs better by a large margin than the non-alignment case, which verifies the motivation of patch alignment to alleviate inner-class variance stated in Section 3.1. Second, for the alignment case, the most representative features to identify a person reside in the region of face, followed by head, upper body and whole body at last. Such a clue is not that obvious for the non-alignment case. Third, we notice that for the whole body region, accuracy in the non-alignment case is higher than that of the alignment region’s score. The identity of patch $i$ in test 1 is decided by the label corresponding to the maximum score in the reference set test 0: $l_i = \arg \max_j S_{ij}$. Such a scheme guarantees that when new training data are added into test 0, there is no need to train a second model or SVM on the reference set, which is quite distinct from previous work.

4.2 Component Analysis
Table 2: Investigation on the combination of merging similarity score from different body regions during inference. The top two results in each split are marked in **bold** and *italic*. COCO loss is applied in all cases except the last one (softmax loss).

| Face | Head | Upper body | Whole body | original | album | time | day |
|------|------|------------|------------|----------|-------|------|-----|
| ✓    | ✓    | ✓          | ✓          | 84.17    | 80.78 | 74.00| 53.75|
| ✓    | ✓    | ✓          | ✓          | 89.24    | 81.46 | 76.84| 61.48|
| ✓    | ✓    | ✓          | ✓          | 88.40    | 82.15 | 70.90| 57.87|
| ✓    | ✓    | ✓          | ✓          | 88.76    | 79.15 | 68.64| 42.91|
| ✓    | ✓    | ✓          | ✓          | 87.43    | 77.54 | 67.40| 42.30|
| ✓    | ✓    | ✓          | ✓          | 81.93    | 73.84 | 62.46| 34.77|
| ✓    | ✓    | ✓          | ✓          | 87.86    | 80.85 | 71.65| 59.03|
| ✓    | ✓    | ✓          | ✓          | 88.13    | 82.87 | 73.01| 55.52|
| ✓    | ✓    | ✓          | ✓          | 89.71    | 78.29 | 66.60| 52.21|
| ✓    | ✓    | ✓          | ✓          | 91.43    | 80.67 | 70.46| 55.56|
| ✓    | ✓    | ✓          | ✓          | **92.78**| **83.53**| **77.68**| **61.73**|
| ✓    | ✓    | ✓          | ✓          | 88.73    | 80.26 | 71.56| 50.36|

Table 3: Recognition accuracy (%) comparison with state-of-the-arts on PIPA, where test sets are divided into four cases to indicate different settings of the identity.

| Methods                                   | original | album | time | day |
|-------------------------------------------|----------|-------|------|-----|
| PIPER [Zhang et al., 2015]                | 83.05    | -     | -    | -   |
| RNN [Li et al., 2016c]                    | 84.93    | 78.25 | 66.43| 43.73|
| Naeil [Oh et al., 2015]                   | 83.05    | 78.25 | 66.43| 43.73|
| Ours                                      | **92.78**| **83.53**| **77.68**| **61.73**|

Figure 5: Histogram of the cosine distance for positive and negative pairs during inference using different losses.

**Congenereous cosine loss.** Figure 5 shows the histogram of the cosine distance among positive pairs (*i.e.*, same identity in test_0 and test_1) and negative pairs. We can see that in the COCO trained model, the discrepancy between inter-class (blue) and inner-class (green) samples in the test set is well magnified; whereas in the softmax trained case, such a distinction is not obvious. This verifies the effectiveness of our COCO loss that maximizes the distinction across classes as well as the similarity within one class.

**Similarity integration from regions.** Table 2 depicts the ablation study on the combination of merging the similarity score from different body regions during inference. Generally speaking, taking all regions into consideration could result in the best accuracy of 92.78 on the original set. It is observed that the performance is still fairly good on day and time if the two scores of face and upper body alone are merged.

### 4.3 Comparison to state-of-the-arts

We can conclude from Table 3 that our recognition system outperforms against previous state-of-the-arts in all four test splits. Figure 4 visualizes a few examples of the predicted instances by our model, where complex scenes with non-frontal faces and body occlusion can be handled properly in most scenarios. Failure cases are probably due to the almost-the-same appearance configuration in these scenarios (same-view of frontal face, similar clothes and background).

### 5 Conclusion

In this work, we propose a person recognition method to identify the same person, where four models for different body regions are trained. Region patches are further aligned via affine transformation to make the model less prone to overfitting. Moreover, the training procedure employs a COCO loss to reduce the inner-class variance as well as enlarge inter-class variation. Our pipeline requires only one-time training of the model; we utilize the similarity between test_0 and test_1 to determine the person’s identity during inference. Experiments show that the proposed method outperforms against other state-of-the-arts on the PIPA dataset.
References

[Anguelov et al., 2007] Dragomir Anguelov, Kuang chih Lee, Salih Burak Gokturk, and Baris Sumengen. Contextual identity recognition in personal photo albums. In CVPR, 2007.

[Chi et al., 2017] Zhizhen Chi, Hongyang Li, Huchuan, and Ming-Hsuan Yang. Dual deep network for visual tracking. IEEE Trans. on Image Processing, 2017.

[Deng et al., 2009] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In CVPR, 2009.

[Girshick, 2015] Ross Girshick. Fast R-CNN. In ICCV, 2015.

[Guo et al., 2016] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. MS-Celeb-1M: a dataset and benchmark for large-scale face recognition. arXiv preprint:1607.08221, 2016.

[He et al., 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, 2016.

[Hinton and Salakhutdinov, 2006] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. Science, 313, 2006.

[Ioffe and Szegedy, 2015] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep neural networks. In ICML, 2015.

[Krizhevsky et al., 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In NIPS, pages 1106–1114, 2012.

[LeCun et al., 1998] Yann LeCun, Loui Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In Proceedings of the IEEE, volume 86, pages 2278–2324, 1998.

[Li et al., 2014] Wei Li, Rui Zhao, Tong Xiao, and Xiaoang Wang. DeepReID: Deep filter pairing neural network for person re-identification. In CVPR, 2014.

[Li et al., 2016a] Haoxiang Li, Jonathan Brandt, Zhe Lin, Xiaohui Shen, and Gang Hua. A multi-level contextual model for person recognition in photo albums. In CVPR, 2016.

[Li et al., 2016b] Hongyang Li, Wanli Ouyang, and Xiaoang Wang. Multi-bias non-linear activation in deep neural networks. In ICML, 2016.

[Li et al., 2016c] Yao Li, Guosheng Lin, Bohan Zhuang, Lingqiao Liu, Chunhua Shen, and Anton van den Hengel. Sequential person recognition in photo albums with a recurrent network. arXiv preprint:1611.09967, 2016.

[Li et al., 2017] Hongyang Li, Yu Liu, Wanli Ouyang, and Xiaoang Wang. Zoom out-and-in network with recursive training for object proposal. arXiv preprint:1702.05711, 2017.

[Liu et al., 2016] Hongye Liu, Yonghong Tian, Yaowei Wang, Lu Pang, and Tiejun Huang. Deep relative distance learning: Tell the difference between similar vehicles. In CVPR, 2016.

[Oh et al., 2015] Seong Joon Oh, Rodrigo Benenson, Mario Fritz, and Bernt Schiele. Person recognition in personal photo collections. In ICCV, 2015.

[Oh et al., 2016] Seong Joon Oh, Rodrigo Benenson, Mario Fritz, and Bernt Schiele. Faceless person recognition: privacy implications in social media. In ECCV, 2016.

[Prosser et al., 2010] Bryan Prosser, Wei-Shi Zheng, Shao-gang Gong, and Tao Xiang. Person re-identification by support vector ranking. In BMVC, 2010.

[Ren et al., 2015] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In NIPS, 2015.

[Tapaswi et al., 2012] Makarand Tapaswi, Martin Bäuml, and Rainer Stiefelhagen. “Knock! Knock! Who is it?” Probabilistic Person Identification in TV Series. In CVPR, 2012.

[Wang and Gupta, 2015] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In ICCV, 2015.

[Wei et al., 2016] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In CVPR, 2016.

[Wen et al., 2016] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In ECCV, 2016.

[Yi et al., 2014] Dong Yi, Zhen Lei, and Stan Z. Li. Deep metric learning for practical person re-identification. arXiv preprint:1407.4979, 2014.

[Zhang et al., 2015] Ning Zhang, Manohar Paluri, Yaniv Taigman, Rob Fergus, and Lubomir Bourdev. Beyond frontal faces: Improving person recognition using multiple cues. In CVPR, 2015.

[Zhao et al., 2014] Rui Zhao, Wanli Ouyang, and Xiaoang Wang. Unsupervised salience learning for person re-identification. In CVPR, 2014.