Quality Aware Network for Set to Set Recognition

Yu Liu
SenseTime Group Limited
liuyuisanai@gmail.com

Junjie Yan
SenseTime Group Limited
yanjunjie@sensetime.com

Wanli Ouyang
University of Sydney
wanli.ouyang@gmail.com

Abstract

This paper targets on the problem of set to set recognition, which learns the metric between two image sets. Images in each set belong to the same identity. Since images in a set can be complementary, they hopefully lead to higher accuracy in practical applications. However, the quality of each sample cannot be guaranteed, and samples with poor quality will hurt the metric. In this paper, the quality aware network (QAN) is proposed to confront this problem, where the quality of each sample can be automatically learned although such information is not explicitly provided in the training stage. The network has two branches, where the first branch extracts appearance feature embedding for each sample and the other branch predicts quality score for each sample. Features and quality scores of all samples in a set are then aggregated to generate the final feature embedding. We show that the two branches can be trained in an end-to-end manner given only the set-level identity annotation. Analysis on gradient spread of this mechanism indicates that the quality learned by the network is beneficial to set-to-set recognition and simplifies the distribution that the network needs to fit. Experiments on both face verification and person re-identification show advantages of the proposed QAN. The source code and network structure can be downloaded at GitHub[1]

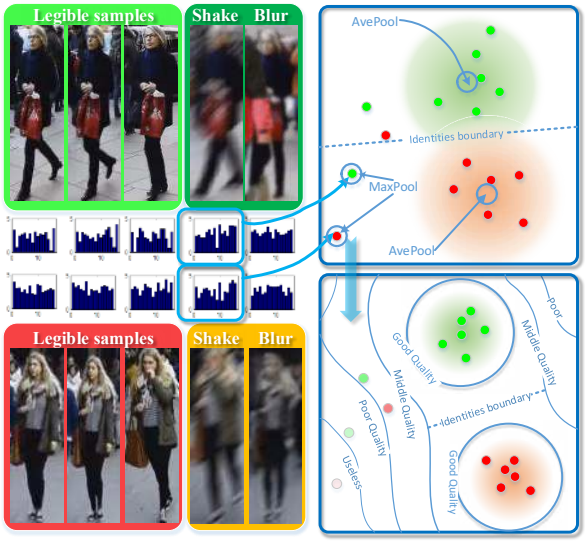

Figure 1. Illustration of our motivation, best viewed in color. ***Left column:*** A classical puzzle in set-to-set recognition. Both set A (upper) and B (lower) contain noisy image samples caused by shake and blur. Their features (shown by histograms in middle row) are more similar to samples in other class than the inner class. ***Right column:*** Distributions and samples of two identities in hyperspace. Top: Due to the noisy, variances of two identities are large and they both have hard negative samples. Bottom: Quality aware network (QAN) weaken the noisy samples and narrow down identities' variances, which makes them more discriminative.

1. Introduction

Face verification [12, 26, 27, 28, 30] and person re-identification [5,6,20,42] have been well studied and widely used in computer vision applications such as financial identity authentication and video surveillance. Both the two tasks need to measure the distance between two face or person images. Such tasks can be naturally formalized as a metric learning problem, where the distance of images from the same identity should be smaller than that from different identities. Built on large scale training data, convolutional neural networks and carefully designed optimization criterion, current methods can achieve promising performance on standard benchmarks, but may still fail due to appearance variations caused by large pose or illumination.

In practical applications, instead of one single image, a set of images for each identity can always be collected. For example, the image set of one identity can be sampled from the trajectory of the face or person in videos. Images in a set can be complementary to each other, so that they provide more information than a single image, such as images from different poses. The direct way to aggregate identity infor-

[1] github.com/sciencefans/Quality-Aware-Network
Note that we are developing P-QAN (a fine-grained version of QAN, see Sec.5) in this repository so the performance of the code may be higher than that we report in this paper.

mation from all images in a set can be simply max/average pooling appearance features of all images. However, one problem in this pooling is that some images in the set may be not suitable for recognition. As shown in Figure 1, both sets from left-top and left-bottom hold noisy images caused by shake or blur. If the noisy images are treated equally and max/average pooling is used to aggregate all images' features, the noisy images will mislead the final representation.

In this paper, in order to be robust to images with poor quality as described above and simultaneously use the rich information provided by the other images, our basic idea is that each image can have a quality score in aggregation. For that, we propose a quality aware network (QAN), which has two branches and then aggregated together. The first branch named *feature generation part* extracts the feature embedding for each image, and the other branch named *quality generation part* predicts quality score for each image. Features of images in the whole set are then aggregated by the final *set pooling unit* according to their quality.

A good property of our approach is that we do not supervise the model by any explicit annotations of the quality. The network can automatically assign low quality scores to images with poor quality in order to keep the final feature embedding useful in set-to-set recognition. To implement that, an elaborate model is designed in which embedding branch and score generation branch can be jointly trained through optimization of the final embedding. Specially in this paper, we use the joint triplet and softmax loss on top of image sets. The designed gradient of image set pooling unit ensures the correctness of this automatic process.

Experiments indicate that the predicted quality score is correlated with the quality annotated by human, and the predicted quality score performs better than human in recognition. In this paper, we show the applications of the proposed method on both person re-identification and face verification. For person re-identification task, the proposed quality aware network improves top-1 matching rates over the baseline by 14.6% on iLIDS-VID and 9.0% on PRID2011. For face verification, the proposed method reduces 15.6% and 29.32% miss ratio when the false positive rate is 0.001 on YouTube Face and IJB-A benchmarks.

The main contributions of the paper are summarized as follows.

- The proposed quality aware network automatically generates quality scores for each image in a set and leads to better representation for set-to-set recognition.
- We design an end-to-end training strategy and demonstrate that the quality generation part and feature generation part benefit from each other during back propagation.
- Quality learnt by QAN is better than quality estimated by human and we achieves new state-of-the-art performance on four benchmarks for person re-identification and face verification.

2. Related work

Our work is build upon recent advances in deep learning based person re-identification and unconstrained face recognition. In person re-identification, [20, 37, 41] use features generated by deep convolutional network and obtain state-of-the-art performance. To learn face representations in unconstrained face recognition, Huang et al. [11] uses convolutional Restricted Boltzmann Machine while deep convolutional neural network is used in [28, 30]. Furthermore, [26, 29] use deeper convolutional network and achieved accuracy that even surpasses human performance. The accuracy achieved by deep learning on image-based face verification benchmark LFW [12] has been promoted to 99.78%. Although deep neural network has achieved such great performance on these two problems, in present world, unconstrained set-to-set recognition is more challenging and useful.

Looking backward, there are two different approaches handling set-to-set recognition. The first approach takes image set as a convex hull [2], affine hull [10] or subspace [1, 13]. Under these settings, samples in a set distribute in a Hilbert space or Grassmann mainfold so that this issue can be formulated as a metric learning problem [23, 39].

Some other works degrade set-to-set recognition to point-to-point recognition through aggregating images in a set to a single representation in hyperspace. The most famous approach in this kind is the Bag of features [17], which uses histogram to represent the whole set for feature aggregation. Another classical work is vector of locally aggregated descriptors (VLAD) [14], which aggregates all local descriptors from all samples. Temporal max/average pooling is used in [36] to integrate all frames' features generated by recurrent convolutional network. This method uses the 1st order statistics to aggregate the set. The 2nd order statistics is used in [32, 43] in assuming that samples follow Gaussian distribution. In [8], original faces in a set are classified into 20 bins based on their pose and quality. Then faces in each bin are pooled to generate features and finally feature vectors in all bins are merged to be the final representation. [38] uses attention mechanism to summarize several sample points to a single aggregated point.

The proposed QAN belongs to the second approach. It discards the dross and selects the essential information in all images. Different from recent works which learn aggregation based on fixed feature [38] or image [8], the QAN learns feature representation and aggregation simultaneously. [7] proposed a similar quality aware module named "memorability based frame selection" which takes "visual entropy" to be the score of a frame. But the score of a frame

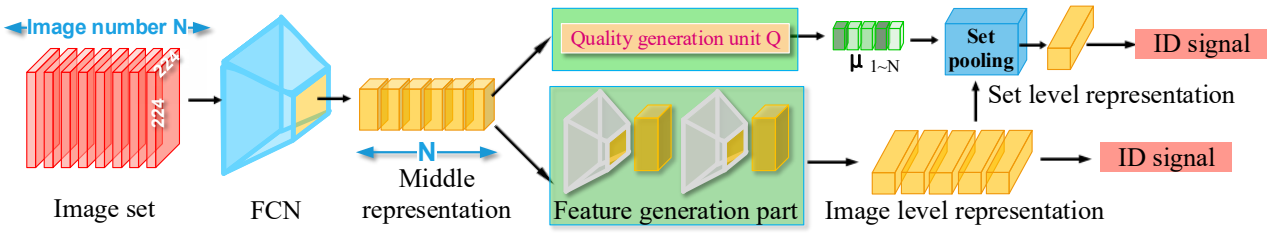

Figure 2. The end-to-end learning structure of quality aware net. The input of this structure is three image sets S_{anchor}, S_{pos} and S_{neg} belong to class A, A and B. Each of them pass through the fully convolutional network (FCN) to generate the middle representations, which will be fed to quality generation part and feature generation part. The former generates quality score for each image and the latter generates final representation for each image. Then the scores and representations of all image will be aggregated by set pooling unit and the final representation of the image set will be produced. We use softmax-loss and triplet-loss to be the supervised ID signal.

is defined by human and independent with feature generation unit. In QAN, score is automatically learned and quality generation unit is joint trained with feature generation unit. Due to mutual benefit between the two parts during training, performance is improved significantly by jointly optimizing images aggregation parameter and images' feature generator.

3. Quality aware network (QAN)

In our work we focus on improving image set embedding model, which maps an image set $S = \{I_1, I_2, \cdots, I_N\}$ to an representation with fixed dimension so that image sets with different number of images are comparable with each other. Let $R_a(S)$ and R_{I_i} denote representation of S and I_i. $R_a(S)$ is determined by all elements in S, therefore it can be denoted as

$$R_a(S) = \mathcal{F}(R_{I_1}, R_{I_2}, \cdots, R_{I_N}). \tag{1}$$

The R_{I_i} is produced by a feature extraction process, containing traditional hand-craft feature extractors or convolutional neural network. $\mathcal{F}(\cdot)$ is an aggregative function, which maps a variable-length input set to a representation of fixed dimension. The challenge is to find an optimized $\mathcal{F}(\cdot)$, which aggregate features from the whole image set to obtain the most discriminative representation. Based on notion that images with higher quality are easier for recognition while images with lower quality containing occlusion and large pose have less effect on set representation, we denote $\mathcal{F}(\cdot)$ as

$$\mathcal{F}(R_{I_1}, R_{I_2}, \cdots, R_{I_N}) = \frac{\sum_{i=1}^{N} \mu_i R_{I_i}}{\sum_{i=1}^{N} \mu_i} \tag{2}$$

$$\mu_i = Q(I_i), \tag{3}$$

where $Q(I_i)$ predicts a quality score μ_i for image I_i. So the representation of a set is a fusion of each images' features, weighted by their quality scores.

3.1. QAN for image set embedding

In this paper, feature generation and aggregation module is implemented through an end-to-end convolutional neural network named QAN as shown in Fig. 2. Two branches are splited from the middle of it. In the first branch, quality generation part followed by a set pooling unit composes the aggregation module. And in the second branch, feature generation part generates images' representation. Now we introduce how an image set flows through QAN. At the beginning of the process, all images are sent into a fully convolutional network to generate middle representations. After that, QAN is divided into two branches. The first one (upper) named quality generation part is a tiny convolution neural network (see Sec. 3.4 for details) which is employed to predict quality score μ. The second one (lower), called feature generation part, generates image representations R_I for all images. μ and R_I are aggregated at set pooling unit $\mathcal{F}$, and then pass through a fully connected layer to get the final representation $R_a(S)$. To sum up, this structure generates quality scores for images, uses these quality scores to weight images' representations and sums them up to produce the final set's representation.

3.2. Training QAN without quality supervision

We train the QAN in an end-to-end manner. The data flow is shown in Fig. 2. QAN is supposed to generate discriminative representations for images and sets belonging to different identities. For image level training, a fully connection layer is established after feature generation part, which is supervised by Softmax loss L_{class}. For set level training, a set's representation $R_a(S)$ is supervised by L_{veri} which is formulated as:

$$L_{veri} = \|R_a(S_a) \quad R_a(S_p)\|^2 \quad \|R_a(S_a) \quad R_a(S_n)\|^2 + \delta \tag{4}$$

The loss function above is referred as *Triplet Loss* in previous works [26]. We define S_a as *anchor set*, S_p as *positive set*, and S_n as *negative set*. This function minimizes variances of intra-class samples while Softmax loss cannot

guarantee that because softmax-loss directly optimizes the probability of each class, but not the discrimination of representation.

Keeping this in mind, we consider the set pooling operation $\mathcal{F}$. The gradients back propagated through set pooling unit can be formulated as follows,

$$\frac{\partial \mathcal{F}}{\partial R_{I_i}} = \frac{\partial R_a(S)}{\partial R_{I_i}} = \mu_i \tag{5}$$

$$\frac{\partial \mathcal{F}}{\partial \mu_i} = \frac{\partial R_a(S)}{\partial \mu_i} = R_{I_i} \quad R_a(S) \tag{6}$$

So we can formulate propagation process of the final loss as

$$\frac{\partial L_{veri}}{\partial R_{I_i}} = \frac{\partial R_a(S)}{\partial R_{I_i}} \cdot \frac{\partial L_{veri}}{\partial R_a(S)} = \frac{\partial L_{veri}}{\partial R_a(S)} \cdot \mu_i \tag{7}$$

$$\begin{aligned} \frac{\partial L_{veri}}{\partial \mu_i} &= \frac{\partial R_a(S)}{\partial \mu_i} \cdot (\frac{\partial L_{veri}}{\partial R_a(S)})^T \\ &= \sum_{j=1}^{D} (\frac{\partial L_{veri}}{\partial R_a(S)_j} \cdot (x_{ij} \quad R_a(S)_j)) \end{aligned} \tag{8}$$

Where D is the dimension of images' representation. We discuss how a quality score μ is automatically learned by this back propagation process.

3.3. Mechanism for learning quality score

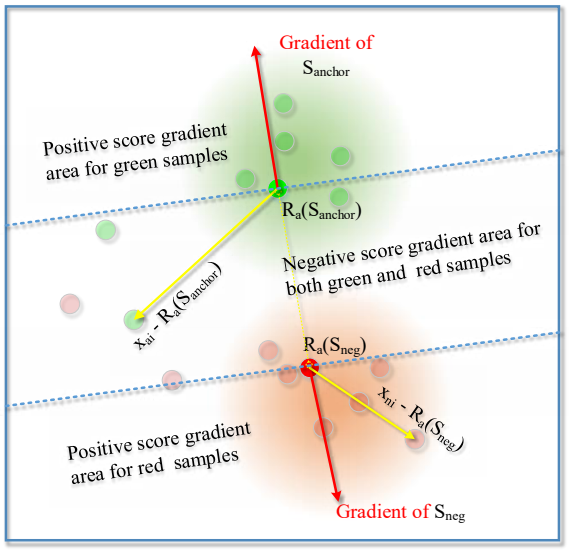

Figure 3. Two different identities in training, best viewed in color. Red translucent dots and green translucent dots indicate images in sets of two different identities. And the two solid dots denote the weighted centers of the two sets, which are also the representations of two sets S_{anchor} and S_{neg}. The gradients of S_{anchor} and S_{neg} are shown with red arrows. The x_{ni} and x_{ai} are two image representations in two sets.

Automatic gradient of μ. After back-propagation through set pooling unit, gradient of μ_i with regard to L_{veri} can be calculated according to the Eq. 8, which is the dot product of gradient from $R_a(S)$ and R_{I_i}. So if angle of $\nabla R_a(S)$ and R_{I_i} belongs to ($90°, 90°$), μ_i's gradient will be positive. For example, as shown in Fig. 3, the angle of $\nabla R_a(S_{neg})$ and x_{ni} $R_a(S_{neg})$ is less than $90°$, so the $x'_{ni}s$ quality score μ_{ni} will become larger after this back propagation process. In contrast, the relative direction of $x_a i$ is in the opposite side of the gradient of $R_a(S_{anchor})$, making it obviously a hard sample, so its quality score μ_{ai} will tend to be smaller. Obviously, samples in the "correct" directions along with set gradient always score higher in quality, while those in the "wrong" directions gain lower weight. For example in Fig. 3, green samples in the upper area and red samples in the lower area keep improving their quality consistently while in the middle area, sample's quality reduces. To this end, μ_i represents whether i th image is a good sample or a hard sample. This conclusion will be further demonstrated by experiments.

μ regulates the attention of R_{I_i}. The gradient of R_{I_i} is shown in Eq. 7 with a factor μ_i, together with the gradient propagated from Softmax loss. Since most of hard samples with lower μ_i are always poor images or even full of background noises, the factor μ_i in gradient of R_{I_i} weaken their harmful effect on the whole model. That is, their impact on parameters in feature generation part is negligible during back propagation. This mechanism helps feature generation part to focus on good samples and neglect ones, which benefits set-to-set recognition.

3.4. Details of quality generation part

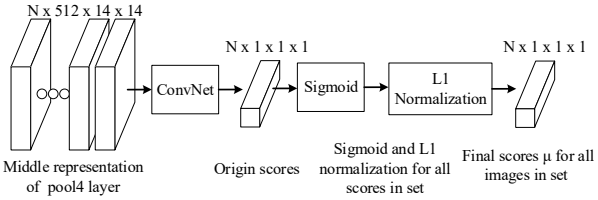

Figure 4. Structure of quality generation unit. The input of this unit is middle representations of a set which contains N images and it produces the normalized weights of all N images.

In quality aware network (QAN), quality generation part is a convolution neural network. We design different score generation parts start at different feature maps. We use QAN split at Pool4 as an instance. As shown in Fig. 4, the output spatial of Pool4 layer is $512 \times 14 \times 14$. In order to generate a 1×1 quality score, the convolution part contains a 2-stride pooling layer and a final pooling layer with kernel size 7×7. A fully connected layer is followed by the final pooling layer to generate the original quality score. After that, the origin scores of all images in a set are sent to

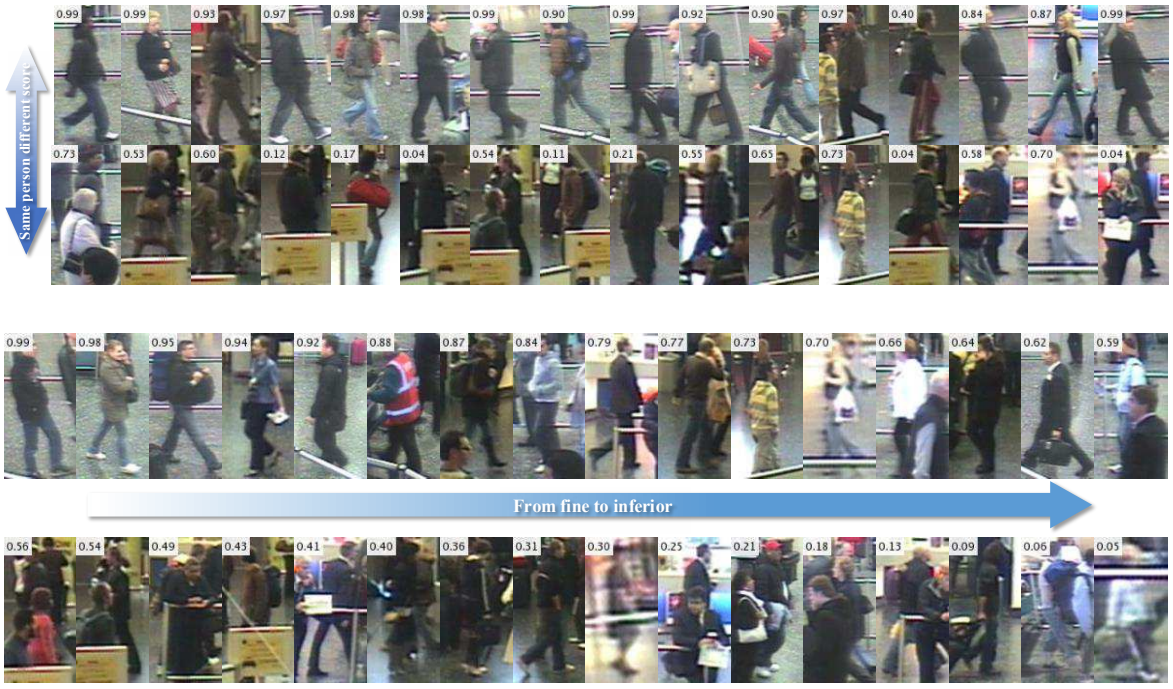

Figure 5. Samples with their qualities predicted by QAN, best viewed in color. **Top:** Comparison between two images from same person. From **up to down**, each column shows the two frames of a same person. The quality of the top one is better than the bottom one. **Bottom:** Random selected images in test set sorted by quality scores from **left to right**, best viewed in color.

sigmoid layer and group L1-normalization layer to generate the final scores μ. For QAN split at `Pool3`, we will add a block containing three 1-stride convolution layer and a 2-stride pooling layer at the beginning of quality generation unit.

4. Experiments

In this section, we first explore the meaning of the quality score learned by QAN. Then QAN's sensitivity to level of feature is analysed. Based on above knowledge, we evaluate QAN on two human re-identification benchmarks and two unconstrained face verification benchmarks. Finally, we analyse the concept learned by QAN and compare it with score labelled by human.

4.1. What is learned in QAN?

Qualitative analysis We visualize images with their μ generated by QAN to explore the meaning of μ. Instances of same person with different qualities are shown in the first two rows in Fig. 5. All images are selected from test set. The two images in the same column belong to a same person. The upper images are random selected from images with quality scores higher than 0.8 and the lower images are selected from images with quality scores lower than the corresponding higher one. It is easy to find that images with deformity, superposition, blur or extreme light condition tend to obtain lower quality scores than normal images.

The last two rows in Fig. 5 give some examples of other images random selected from test set. They are sorted by their quality scores from left to right. We can observe that instances with quality scores larger than 0.70 are easy to recognize by human while the others are hard. Especially many of hard images include two or more bodies in the center and we can hardly discriminate which one is the right target.

Quantitative analysis In order to measure the relationship between the quality labelled by human and μ predicted by QAN, 1000 images in YouTube Face are selected randomly and the quality of them are rated subjectively by 6 volunteers, where each volunteer estimates a quality score for each image, ranging from 0 to 1. All the ratings of each volunteer are aligned by logistic regression. Then the 6 aligned scores of each image are averaged and finally normalized to $[0, 1]$ to get the final quality score from human.

We divide the images into ten partitions based on human's score as shown in Fig. 6. In which we show the corresponding quality statistics generated by QAN. It is obvious that the scores given by the QAN are strongly correlated with human-defined quality. We further analyse the 499,500 image pairs from these 1000 images and ask hu-

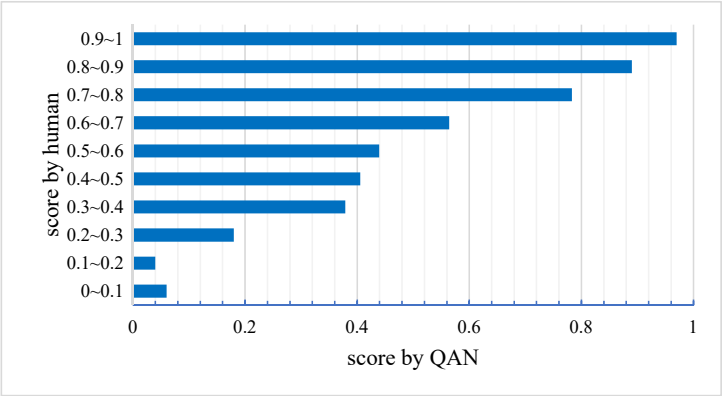

Figure 6. Comparison of qualities estimated by human and predicted by QAN.

man and QAN to select the better one in each pair. Result shows that the decision made by QAN has 78.1% in common with human decision.

4.2. Person re-identification

Datasets. For person re-identification, we collect 134,942 frames with 16,133 people and 212,726 bounding boxes as the training data. Experiments are conducted on PRID2011 [9] and iLiDS-VID [33] datasets. PRID2011 contains frames in two views captured at different positions of a street. `CameraA` has 385 identities while `CameraB` has 749 identities, and the two videos have a overlap of 200 people. Each person has 5 to 675 images, and the average number is 100. iLIDS-VID dataset has 300 people, and each person has two sets also captured from different positions. Each person has 23 to 192 images.

Evaluation procedure. The results are reported in terms of *Cumulative Matching Characteristics* (CMC) table, each column in which represents matching rate in a certain top-N matching. Two settings are used for comprehensive evaluation. In the first setting, we follow the state-of-the-art method described in [40] and [34]. The sets whose frame number is larger than 21 are used in PRID2011, and all the sets in iLIDS-VID are used. Each dataset is divided into two parts for fine-tuning and testing, respectively. For the testing set, sets form `CameraA` are taken as probe set while sets from `CameraB` are taken as the gallery. The final number is reported as the average of "10-fold cross validation". In the second setting, we conduct cross-dataset testing. Different from the first setting, we ignore the finetuning process and use all data to test our model. That is, in PRID2011, the first 200 people from `CameraA` serve as probes, and all sets from `CameraB` are used as the gallery set. In iLIDS-VID, `CameraA` are used as the probe set, and Camera B serve as gallery set.

Baseline. We implement two baseline approaches. In the first baseline, we use average pooling to aggregate all images' representations. In the second baseline, a minimal cosine distance between two closures is used to be their similarity.

4.2.1 Evaluation on common setting

Results of evaluation obeying "10-fold cross validation" on PRID2011 and iLIDS-VID are shown in Table 1 and Table 2. Benefiting from the large scale training dataset, our CNN+AvePool and CNN+Min(cos) baselines are close to or even better than the state-of-the-art. Notice that most of the leading methods listed in table consider both appearance and spatio-temporal information while our method only considers appearance information. On PRID2011 dataset, QAN increase top-1 matching rate by 11.1% and 29.4% compared with CNN+AvePool and CNN+Min(cos). On iLIDS-VID dataset, inherent noise is much more than that in PRID2011, which significantly influence the accuracy of CNN+Min(cos) since operator "Min(cos)" is more sensitive than "AvePool" to noisy samples . However, QAN achieves more gain on this noisy dataset. It increase top-1 matching rate by 12.21% and 37.9%.

PRID2011				
Methods	CMC1	CMC5	CMC10	CMC20
QAN	**90.3**	**98.2**	**99.32**	**100.0**
CNN+AvePool	81.3	96.6	98.5	99.6
CNN+Min(cos)	69.8	91.3	97.1	99.8
CNN+RNN [36]	70	90	95	97
STFV3D [22]	42.1	71.9	84.4	91.6
TDL [40]	56.7	80.0	87.6	93.6
eSDC [34]	48.3	74.9	87.3	94.4
DVR [34]	40.0	71.7	84.5	92.2
LFDA [25]	43.7	72.8	81.7	90.9
KISSME [16]	34.4	61.7	72.1	81.0
LADF [21]	47.3	75.5	82.7	91.1
TopRank [19]	31.7	62.2	75.3	89.4

Table 1. Comparison of QAN, `AvePool`, `Min(cos)` and other state-of-the-art methods on PRID2011, where the number represents the cumulative matching rate in CMC curve.

Based on these two experiments, QAN significantly outperforms two baselines on both datasets. It also performs better than many state-of-the-art approaches and pushes top-1 matching rate 20.3% higher than previous best CNN+RNN [36] on PRID2011 and 10% on iLIDS-VID. The performance gain is more significant on noisy iLIDS-VID dataset, which meets the expectation and proves QAN's ability to deal with images of poor quality.

4.2.2 Dataset cross evaluation

To prevent our model from over-fitting the quality distribution of test set, we conduct dataset cross evaluation. We

iLIDS-VID				
Methods	CMC1	CMC5	CMC10	CMC20
QAN	**68.0**	**86.8**	95.4	97.4
CNN+AvePool	60.6	84.9	89.8	93.6
CNN+Min(cos)	49.3	79.4	88.2	91.9
CNN+RNN [36]	58	84	91	96
STFV3D [22]	37.0	64.3	77.0	86.9
TDL [40]	56.3	87.6	**95.6**	**98.3**
eSDC [34]	41.3	63.5	72.7	83.1
DVR [34]	39.5	61.1	71.7	81.0
LFDA [25]	32.9	68.5	82.2	92.6
KISSME [16]	36.5	67.8	78.8	87.1
LADF [21]	39.0	76.8	89.0	96.8
TopRank [19]	22.5	56.1	72.7	85.9

Table 2. Comparison of QAN, `AvePool`, `Min(cos)` and other human re-identification methods on iLIDS-VID, where the number represents the cumulative matching rate on CMC curve.

PRID2011				
Methods	CMC1	CMC5	CMC10	CMC20
QAN	**34.0**	**61.3**	**74.0**	**83.1**
CNN+AvePool	29.4	57.5	68.8	80.2
CNN+Min(L2)	28.5	57.1	67.1	78.6
CNN+RNN [36]	28	57	69	81

Table 3. Cross-dataset performance of QAN on PRID2011, where the number represents the cumulative accuracy on CMC curve.

iLIDS-VID				
Methods	CMC1	CMC5	CMC10	CMC20
QAN	**47.7**	**70.4**	**83.9**	**91.3**
CNN+AvePool	44.1	65.8	78.5	88.9
CNN+Min(L2)	41.9	61.7	75.5	79.5

Table 4. Cross-dataset performance of QAN on iLIDS-VID, where the number represents the cumulative accuracy on CMC curve.

extract set representation of iLIDS-VID and PRID2011 directly using trained QAN without fine-tuning. The QAN representation is then evaluated for CMC scores. Table 3 and 4 shows the results of QAN and the two baselines. It can be found that the QAN is robust even in cross-dataset setting. It improves top-1 matching by 15.6% and 8.2% compared to the baselines. This result shows that the quality distribution learned from different datasets by QAN is able to generalize to other datasets.

4.3. Unconstrained face verification

Datasets. For face verification, we train our base model on extended version of VGG Face dataset [24], in which we extend the identity number from 2.6K to 90K and image number from 2.6M to 5M. The model is evaluated on YouTube Face Database [35] and IARPA Janus Benchmark A (IJB-A) dataset. YouTube Face contains 3425 videos of 1595 identities. It is challenging in that most faces are blurred or has low resolution. IJB-A dataset contains 2042 videos of 500 people. Faces in IJB-A have large pose variance.

Evaluation procedure. We follow the 1:1 protocol in both two benchmarks and evaluate results using receiver operating characteristic (ROC) curves. Area under curve (AUC) and accuracy are two important indicators of the ROC. The datasets are evaluated using 10-fold cross-validation.

Training details. All faces in training and testing sets are detected and aligned by a multi-task region proposal network as described in [3]. Then we crop the face regions and resize them to 256×224. After that, a convolutional neural networks with 256×224 inputs are used for face verification. It begins with a 2-stride convolution layer, followed by 4 basic blocks, while each block has three 1-stride convolution layers and one 2-stride pooling layers. After that, a fully connected layer is used to get the final feature. Quality generation branch is built on top of the third pooling layer, where the spatial size of middle representation response is $256 \times 16 \times 14$. We pre-train the network supervised by classification signal and then train the whole QAN.

4.3.1 Results on YouTube Face and IJB-A benchmark

Method	Accuracy(%)	AUC
QAN	**96.17± 0.09%**	**99.14± 0.12%**
CNN+AvePool	95.46± 0.07%	98.66± 0.04%
CNN+Min(cos)	94.87± 0.10%	98.37± 0.06%
NAN [38]	95.52±0.06%	98.7%
FaceNet [26]	95.12±0.39%	-
DeepID2+ [29]	93.2±0.2%	-
DeepFace-single [30]	91.4±1.1%	96.3%
EigenPEP [18]	84.8±1.4%	92.6%

Table 5. Average accuracy and AUC of QAN on YouTube Face dataset, compared with baselines and other state-of-the-arts.

TPR@FPR	1e-3	1e-2	1e-1
QAN	**89.31±3.92%**	**94.20±1.53%**	**98.02±0.55%**
CNN+AvePool	85.30±3.48%	93.81±1.44	97.85±0.61%
CNN+Min(cos)	82.74±3.61%	92.06±1.98	97.29±0.67%
NAN [38]	78.5±2.8%	89.7±1.0%	95.9±0.5%
DCNN+metric [4]	-	78.7±4.3%	94.7±1.1%
LSFS [31]	51.4±6.0%	73.3±3.4%	89.5±1.3%
OpenBR [15]	10.4±1.4%	23.6±0.9%	43.3±0.6%

Table 6. TPRs of QAN at specific FPRs on IJB-A dataset, compared with baselines and other state-of-the-arts.

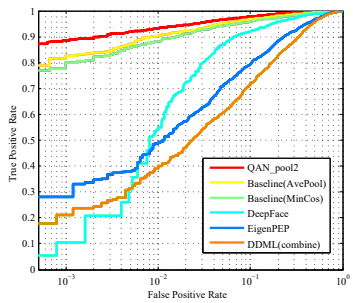

Figure 7. Average ROC curves of different methods on YouTube Face Dataset

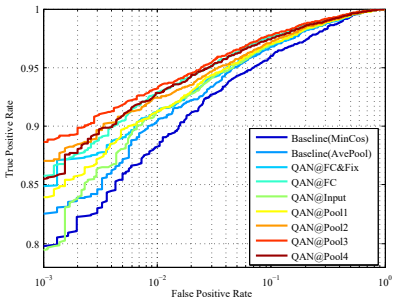

Figure 8. ROC results for score generation part learned by different level of feature.

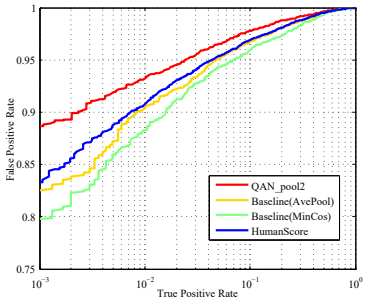

Figure 9. QAN with human score performs better than the two baseline but worse than that scored by network.

On YouTube Face dataset, it can be observed in Fig. 7 and Table 5 that the accuracy and AUC of our baselines are similar with the state-of-the-art methods such as FaceNet and NAN. Based on this baseline, QAN further reduces 15.6% error ratio. Under ROC evaluation metric, QAN surpasses NAN by 8% and DeepFace by 80% at 0.001 FPR (false positive rate), which ensembles 25 models.

On IJB-A dataset, QAN significantly outperforms the state-of-the-art algorithm NAN by 10.81% at 0.001 FPR, 4.5% at 0.01 FPR and 2.12% at FPR=0.1, as shown in Table 6. Compared with average pooling baseline, QAN reduces false negative rate at above three FPRs by 29.32%, 6.45% and 7.91%.

Our experiments on the two tasks show that QAN is robust for set-to-set recognition. Especially on the point of low FPR, QAN can recall more matched samples with less errors.

4.4. Quality by QAN VS. quality by human

There is no explicit supervision signals for the cascade score generation unit in training. So another problem arises: is it better to use human-defined scores instead of letting the network learn itself? In YouTube Face experiment, we replace the quality score $Q(I)$ with volunteer-rated scores and got the following result in Fig.9, which is better than the two baselines but inferior to the result of original QAN. It shows that Q is similar with human thoughts, but more suitable for recognition. Quality score by human can also enhance the accuracy but is still worse than QAN's.

4.5. Diagnosis experiments

Level of middle representation may affect the performance of QAN. We use YouTube face to analyse this factor by comparing different configurations.

In the first configuration, the weight generation part is connected to the image. In the second to fifth configurations, weight generation part is set after four pooling layers in each block, respectively. In the sixth configuration, we connect weight generation part to a fully connected layer. For the final configuration, we fix all parameters before the final fully connection layer in the sixth configuration and only update parameters in weight generation part, which is taken as the seventh structure. To minimize the influence by parameters' number, the total size of different models is restricted to the same by changing the channel number.

Results are shown in Fig. 8. It can be found that the performance of QAN improves at the beginning and reaches the top accuracy at Pool3. The end-to-end training version of feature generation part with quality generation part performs better than that of fixed. So we can make the conclusion that 1) the middle level feature is better for QAN to learn and 2) significant improvement can be achieved by jointly training feature generation part and quality generation part.

5. Conclusion and future work

In this paper we proposes a Quality Aware Network (QAN) for set-to-set recognition. It automatically learns the concept of quality for each sample in a set without supervised signal and aggregates the most discriminative samples to generate set representation. We theoretically and experimentally demonstrate that the quality predicted by network is beneficial to set representation and better than human labelled.

QAN can be seen as an attention model that pay attention to high quality elements in a image set. However, an image with poor quality may still has some discriminative regions. Considering this, our future work will explore a fine-grained quality aware network that pay attention to high quality regions instead of high quality images in a image set.

References

[1] Ronen Basri, Tal Hassner, and Lihi Zelnik-Manor. Approximate nearest subspace search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(2):266–278, 2011. 2

[2] Hakan Cevikalp and Bill Triggs. Face recognition based on image sets. In *CVPR'10*, pages 2567–2573. IEEE, 2010. 2

[3] Dong Chen, Gang Hua, Fang Wen, and Jian Sun. Supervised transformer network for efficient face detection. In *European Conference on Computer Vision*, pages 122–138. Springer, 2016. 7

[4] Jun-Cheng Chen, Rajeev Ranjan, Amit Kumar, Ching-Hui Chen, Vishal Patel, and Rama Chellappa. An end-to-end system for unconstrained face verification with deep convolutional neural networks. In *ICCV Workshops*, pages 118–126, 2015. 7

[5] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR, 2010 IEEE Conference on*, pages 2360–2367. IEEE, 2010. 1

[6] Shaogang Gong, Marco Cristani, Shuicheng Yan, and Chen Change Loy. *Person re-identification*, volume 1. Springer, 2014. 1

[7] Gaurav Goswami, Romil Bhardwaj, Richa Singh, and Mayank Vatsa. Mdlface: Memorability augmented deep learning for video face recognition. In *Biometrics (IJCB), 2014 IEEE International Joint Conference on*, pages 1–7. IEEE, 2014. 2

[8] Tal Hassner, Iacopo Masi, Jungyeon Kim, Jongmoo Choi, Shai Harel, Prem Natarajan, and Gerard Medioni. Pooling faces: template based face recognition with pooled face images. In *CVPR'16 Workshops*, pages 59–67, 2016. 2

[9] Martin Hirzer, Csaba Beleznai, Peter M. Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *Proc. Scandinavian Conference on Image Analysis (SCIA)*, 2011. 6

[10] Yiqun Hu, Ajmal S Mian, and Robyn Owens. Sparse approximated nearest points for image set classification. In *CVPR'11*, pages 121–128. IEEE, 2011. 2

[11] Gary B. Huang. Learning hierarchical representations for face verification with convolutional deep belief networks. In *CVPR*, CVPR '12, pages 2518–2525, Washington, DC, USA, 2012. IEEE Computer Society. 2

[12] Gary B Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst, 2007. 1, 2

[13] Zhiwu Huang, Ruiping Wang, Shiguang Shan, and Xilin Chen. Projection metric learning on grassmann manifold with application to video based face recognition. In *CVPR'15*, pages 140–149, 2015. 2

[14] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. Aggregating local descriptors into a compact image representation. In *CVPR'10*, pages 3304–3311. IEEE, 2010. 2

[15] Joshua C Klontz, Brendan F Klare, Scott Klum, Anubhav K Jain, and Mark J Burge. Open source biometric recognition. In *Biometrics: Theory, Applications and Systems (BTAS), 2013*, pages 1–8. IEEE, 2013. 7

[16] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *CVPR'12*, pages 2288–2295. IEEE, 2012. 6, 7

[17] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR'06*, volume 2, pages 2169–2178. IEEE, 2006. 2

[18] Haoxiang Li, Gang Hua, Xiaohui Shen, Zhe Lin, and Jonathan Brandt. Eigen-pep for video face recognition. In *Computer Vision–ACCV 2014*, pages 17–33. Springer, 2014. 7

[19] Nan Li, Rong Jin, and Zhi-Hua Zhou. Top rank optimization in linear time. In *Advances in Neural Information Processing Systems*, pages 1502–1510, 2014. 6, 7

[20] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *ICCV*, pages 152–159, 2014. 1, 2

[21] Zhen Li, Shiyu Chang, Feng Liang, Thomas Huang, Liangliang Cao, and John Smith. Learning locally-adaptive decision functions for person verification. In *CVPR'13*, pages 3610–3617, 2013. 6, 7

[22] Kan Liu, Bingpeng Ma, Wei Zhang, and Rui Huang. A spatio-temporal appearance representation for viceo-based pedestrian re-identification. 6, 7

[23] Jiwen Lu, Gang Wang, Weihong Deng, Pierre Moulin, and Jie Zhou. Multi-manifold deep metric learning for image set classification. In *CVPR'15*, pages 1137–1145, 2015. 2

[24] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015. 7

[25] Sateesh Pedagadi, James Orwell, Sergio Velastin, and Boghos Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In *ICCV'13*, pages 3318–3325, 2013. 6, 7

[26] 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*, pages 815–823, 2015. 1, 2, 3, 7

[27] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. In *NIPS*, pages 1988–1996, 2014. 1

[28] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation from predicting 10,000 classes. In *CVPR*, pages 1891–1898, 2014. 1, 2

[29] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deeply learned face representations are sparse, selective, and robust. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2892–2900, 2015. 2, 7

[30] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *ICCV*, pages 1701–1708, 2014. 1, 2, 7

[31] Dayong Wang, Charles Otto, and Anil K Jain. Face search at scale: 80 million gallery. *arXiv preprint arXiv:1507.07242*, 2015. 7

[32] Ruiping Wang, Huimin Guo, Larry S Davis, and Qionghai Dai. Covariance discriminative learning: A natural and efficient approach to image set classification. In *CVPR'12*, pages 2496–2503. IEEE, 2012. 2

[33] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by video ranking. In *ECCV 2014*, pages 688–703. Springer, 2014. 6

[34] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by discriminative selection in video ranking. 2016. 6, 7

[35] Lior Wolf, Tal Hassner, and Itay Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR'11*, pages 529–534. IEEE, 2011. 7

[36] Lin Wu, Chunhua Shen, and Anton van den Hengel. Deep recurrent convolutional networks for video-based person re-identification: An end-to-end approach. *arXiv preprint arXiv:1606.01609*, 2016. 2, 6, 7

[37] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learning deep feature representations with domain guided dropout for person re-identification. *arXiv preprint arXiv:1604.07528*, 2016. 2

[38] Jiaolong Yang, Peiran Ren, Dong Chen, Fang Wen, Hongdong Li, and Gang Hua. Neural aggregation network for video face recognition. *arXiv preprint arXiv:1603.05474*, 2016. 2, 7

[39] Meng Yang, Pengfei Zhu, Luc Van Gool, and Lei Zhang. Face recognition based on regularized nearest points between image sets. In *Automatic Face and Gesture Recognition (FG), 2013 Workshops on*, pages 1–7. IEEE, 2013. 2

[40] Jinjie You, Ancong Wu, Xiang Li, and Wei-Shi Zheng. Top-push video-based person re-identification. *arXiv preprint arXiv:1604.08683*, 2016. 6, 7

[41] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *European Conference on Computer Vision*, pages 868–884. Springer, 2016. 2

[42] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Person re-identification by probabilistic relative distance comparison. In *CVPR, 2011 IEEE conference on*, pages 649–656. IEEE, 2011. 1

[43] Pengfei Zhu, Lei Zhang, Wangmeng Zuo, and David Zhang. From point to set: Extend the learning of distance metrics. In *ICCV'13*, pages 2664–2671, 2013. 2