Confidence-Driven Network for Point-to-Set Matching

Mengjun Leng and Ioannis A. Kakadiaris
Computational Biomedicine Lab
Computer Science, University of Houston
{mleng2, ikakadia}@uh.edu

***Abstract*—The goal of point-to-set matching is to match a single image with a set of images from a subject. Within an image set, different images contain various levels of discriminative information and thus should contribute differently to the results. However, the discriminative level is not accessible directly. To this end, we propose a confidence driven network to perform point-to-set matching. The proposed system comprises a feature extraction network (FEN) and a performance prediction network (PPN). Given an input image, the FEN generates a template, while the PPN generates a confidence score which measures the discriminative level of the template. At matching time, the template is used to compute a point-to-point similarity. The similarity scores from different samples in the set are integrated at a score level, weighted by the predicted confidence scores. Extensive multi-probe face recognition experiments on the IJB-A and UHDB-31 datasets demonstrate performance improvements over state of the art algorithms.**

I. Introduction

Many real-life applications can be mapped into the point-to-set matching problem. In this paper, we focus on multi-probe face recognition where each probe sample contains multiple images from the same subject (i.e., set data). These images can be video frames or independent photos captured from different cameras. Only a single image is stored for each subject in the gallery (i.e., point data). The task is to match each probe image set to its corresponding gallery subject. Although recent single-probe (i.e., point-to-point) face identification systems have achieved remarkable performance [1], [2], there is still a need to investigate multi-probe face recognition. In real-life scenarios, many problems can be naturally formulated as multi-probe face identification. When a child goes missing, or the police are looking for a crime suspect, a single frontal face image (e.g., a photo ID) may be matched to a set of face images captured through a camera network. Such face images naturally form a set. Additionally, being able to leverage the additional information provided by multiple images is a potential solution for challenges such as extreme poses of the subject depicted in the image and poor illumination conditions. However, multi-probe settings introduce many challenges. First, not all the images provided are informative, and some (e.g., with extreme poses and occlusion) could impact negatively the whole set performance. Second, the point and set data are usually represented using different models, and hence they share very distinct properties, which make it hard to match them directly [3], [4].

To address the challenges mentioned above, we propose a confidence driven network (CDN). The CDN has the following properties: (i) each probe template in the set is matched independently with the gallery template, and thus, there is no model difference between point and set data; and (ii) each probe template contributes differently to the final decision according to its discriminative level, as measured by the confidence score. As illustrated in Figure 1, CDN comprises two parts: (i) a feature extraction network (FEN), and (ii) a performance prediction network (PPN). The FEN is a distance-based ConvNet architecture, and a pre-trained point-to-point network can be employed. The PPN is a binary classification network. It takes as input an intermediate feature representation from FEN and generates a confidence score. This score indicates the probability that the extracted feature vector will contribute towards a correct decision. To estimate the ground-truth confidence score, we propose a ground-truth generation paradigm. Specifically, we perform single-sample tests on batches of triplets in a leave-one-out manner and compute the average rank-1 rate on each sample. This estimated value is then assigned as ground truth. In the training phase, the confidence score will guide the feature extraction network to focus less "attention" on the samples with lower confidence levels. By doing so, it avoids overfitting on samples that are possibly difficult, or others that the model is uncertain about its predictions. In the matching phase, the confidence scores are used to integrate the results from different samples of the same set. In summary, the contributions of the proposed CDN are the following: (i) a weighted-by-confidence point-to-set triplet loss that enables us to adapt a point-to-point network to a point-to-set network; and (ii) a single-sample test mechanism to quantify the discriminative level of a sample.

II. Related Work

Point-to-set matching: A straightforward approach to address the point-to-set matching problem is to model the point-to-set distance. Depending on how the point-to-set distance is formulated, existing methods can be grouped into two categories: (i) set-based approaches, and (ii) sample-based approaches. In set-based approaches, an image set is represented using different models, such as hull [5], [6], [7], [8], linear subspace [9], [3], [4], Grassmann manifold [10], [11], and statistical models [12], [13], [14]. Then, the point-to-set distance is mapped to a cross-domain matching problem.

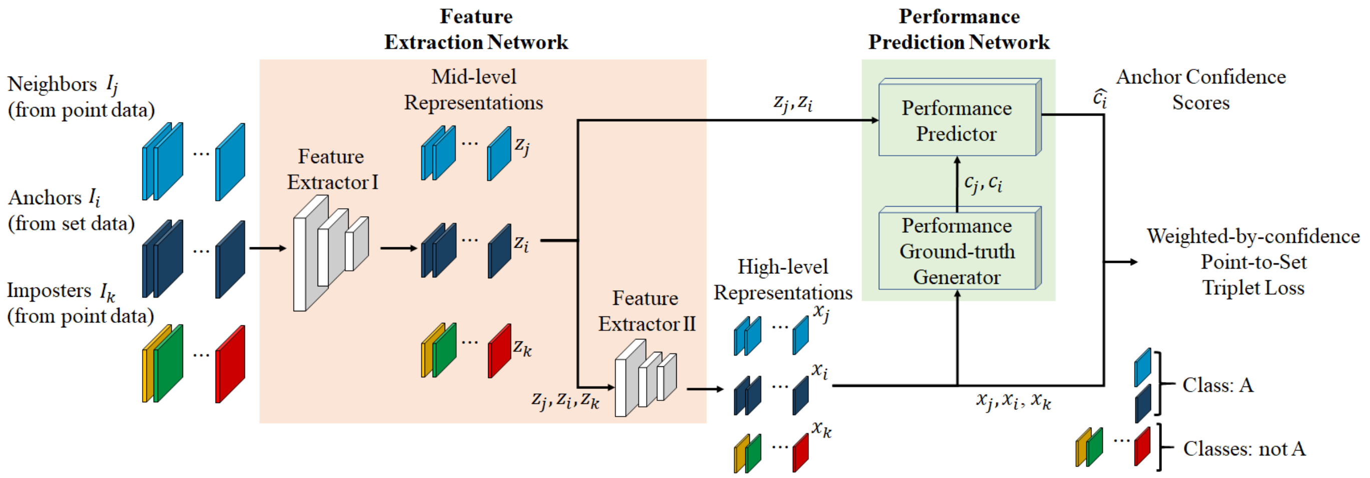

Feature Extraction Network
Performance Prediction Network
Neighbors I_j (from point data)
Anchors I_i (from set data)
Imposters I_k (from point data)
Mid-level Representations
Feature Extractor I
z_j
z_i
z_k
z_j, z_i
Performance Predictor
c_j, c_i
Performance Ground-truth Generator
Anchor Confidence Scores
$\hat{c}_i$
Weighted-by-confidence Point-to-Set Triplet Loss
Feature Extractor II
z_j, z_i, z_k
High-level Representations
x_j
x_i
x_k
x_j, x_i, x_k
Class: A
Classes: not A

Fig. 1: The proposed CDN architecture contains two parts. The first part is a feature extraction network, which consists of Feature Extractor I that contains the first block of a pre-trained classification network, and Feature Extractor II that contains the rest of the blocks. The second part is a performance prediction network, which comprises a ground-truth generator and a performance predictor. The high-level features that contain information about the identity are leveraged to generate "ground-truth" using a single-sample test discussed in Section III-B2. The middle-level features along with the generated "ground-truth" are used to train the performance predictor which generates confidence scores $\hat{c}_i$ for each of the anchors in the set.

Algorithms in this category leverage intra-set relationships, but rely on strong model assumptions which do not always hold. Another limitation of such approaches is that the model estimation is computationally expensive, especially in a high dimensional feature or sample space. In contrast, sample-based approaches treat each image in the set independently, and the final point-to-set distances are fused at different levels [15], [16], [17], [18]. Following the deep learning revolution, most recent techniques employ deep architectures and simple fusion rules to improve the point-to-set matching performance [1], [2], [19]. Compared with the set-based approaches, the sample based ones are more flexible and efficient in the matching phase. However, the within-set correlations are not fully explored.

Image set weighting schemes: Our solution follows a sample-based approach, and thus the set weighting scheme is of paramount importance. Recent set-to-set matching methods [16], [17], [20] employ deep-learning architectures to learn the set weighting scheme. Specifically, an image set is embedded into a single template using a weighted average. Yang *et al.* [17] introduced the Neural Aggregation Network, in which the learned features are fed to an attention mechanism which organizes the input through accessing external memory. These features are then aggregated into a fixed length feature vector adaptively via two attention blocks. Liu *et al.* [16] followed a different approach and proposed a quality-aware network (QAN) in which the template and image quality scores are learned jointly by minimizing a weighted loss function. The proposed Confidence Driven Network shares the same intuition with the aforementioned approaches, but has four key distinct differences: (i) we estimate the confidence score "ground truth" using a single-sample test, which makes our framework flexible enough to be plugged into other pre-trained networks, (ii) instead of embedding features into a single template, the aggregation happens at a distance level without model differences, (iii) we focus on point-to-set matching instead of set-to-set matching; and (iv) our architecture is less complicated and easier to train.

III. Confidence-Driven Network

A. *Feature Extraction Network*

Our objective is to leverage the discriminative power of templates in a point-to-point matching setup and adjust them to a point-to-set matching protocol, by introducing only a few changes to the original architecture. Thus, we selected the center loss face ResNet [21] developed by Wen *et al.* [19] as a base architecture, due to its outstanding performance. To adapt FEN to point-to-set matching, a weighted-by-confidence point-to-set triplet loss is proposed. The triplet loss [2] is chosen because of its ability to learn discriminative templates that can be generalized to distinguish unseen classes. In particular, the training data are split into triplet batches

$$\mathcal{B} = \{(I_i, I_j, I_k) \,|, l_j = l_i = l_+, l_k \neq l_+\}\,, \tag{1}$$

with the restrictions: (i) all the anchor images I_i are sampled from an image set; (ii) the neighbors I_j and imposters I_k are sampled from the point data; (iii) the anchors and the neighbors are sampled from the same subject (i.e., $l_j = l_i = l_+$); and (iv) the anchors and the imposter are from different subjects (i.e., $l_k \neq l_+$). The weighted-by-confidence point-to-set triplet loss in a batch is formulated as follows:

$$\mathcal{L}_{\mathcal{B}} = \frac{\sum_{\mathcal{B}} \hat{c}_i \left[d^2\left(\mathbf{x}_i, \mathbf{x}_j\right) - d^2\left(\mathbf{x}_i, \mathbf{x}_k\right) + \alpha \right]_+}{\sum_{\mathcal{B}} \hat{c}_i}. \tag{2}$$

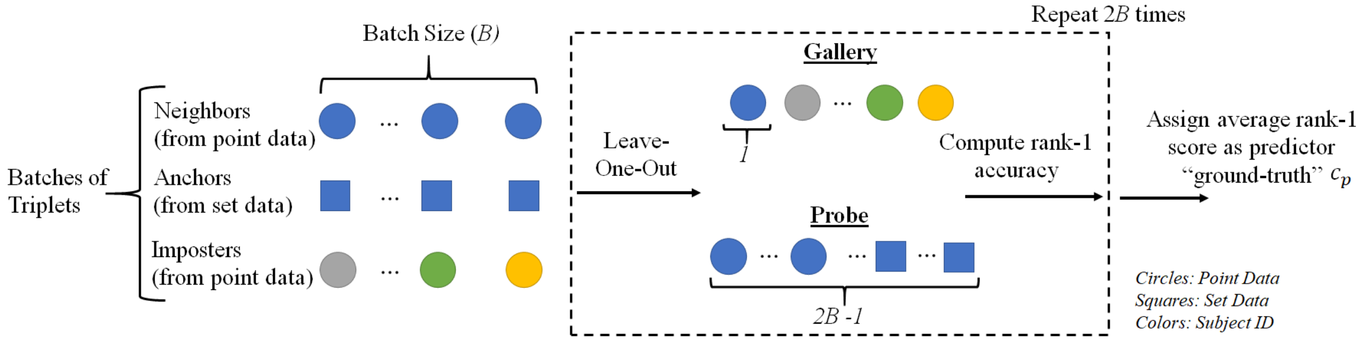


Fig. 2: Single-sample test mechanism to generate ground truth targets for the performance predictor. Given high-level batches of triplets, a single sample from the neighbors or the anchors is placed in the gallery along with all the imposters. The rest of the neighbors and the anchors form the probe. The rank accuracy is computed across 2B iterations.

where $\mathbf{x}_i$, $\mathbf{x}_j$, $\mathbf{x}_k$ are the templates of the anchors, neighbors, and imposters respectively. The objective of the first term is to the pull the neighbors $\mathbf{x}_j$ "closer", while the objective of the second term is to push the imposters $\mathbf{x}_k$ "far away", by a margin α. The proposed loss $\mathcal{L}_\mathcal{B}$ is the sum of the losses over all the triplets in the batch, weighted by the corresponding confidence scores $\hat{c}_i$ of the anchors $\mathbf{x}_i$ from an image set. The confidence score $\hat{c}_i$ generated by the performance predictor guides the FEN to focus less attention on the less informative samples.

B. *Performance Prediction Network*

1) Performance Predictor: The performance predictor is a binary classification network that distinguishes between informative and non-informative samples. In our implementation, we used the softmax output probabilities as confidence scores. The inputs of the performance predictor are the middle-level features from Feature Extractor I. The middle-level representation is used because it preserves more information from the original image, as the high-level templates only preserve the identity-related information.

2) Ground Truth Generator: In order to train the performance predictor in a supervised manner, "ground truth" confidence scores are necessary. Such measurements are not available and thus, we opted for a conceptually straightforward approach to estimate the "ground truth" targets for the confidence scores. The confidence score proposed in this paper is the likelihood of a correct decision,

$$c_i = P(\hat{l}_i = l_i | \mathbf{x}_i), \tag{3}$$

where $\mathbf{x}_i$ is the anchor template from Feature Extractor II, $\hat{l}_i$ is the rank-1 identity label returned from a distance based ranking, and l_i is the ground truth identity label. The confidence score c_i, indicates the probability of returning a correct decision (i.e., $\hat{l}_i = l_i$), based on the given template $\mathbf{x}_i$. To estimate this likelihood ground truth c_i, we propose a single-sample test mechanism, which is performed during training within a triplet batch. An overview of this mechanism is provided in Figure 2.

The input batch comprises B sets of triplets,

(i) Anchors $\mathcal{A} = \{(\mathbf{x}_i, l_+) \mid i \in [1, B]\}$, from the set data;
(ii) Neighbors $\mathcal{A}_+ = \{(\mathbf{x}_j, l_+) \mid j \in [B+1, 2B]\}$, from the point data;
(iii) Imposters $\mathcal{A}_- = \{(\mathbf{x}_k, l_k) \mid k \in [2B+1, 3B]\}$, from the point data.

To further distinguish among the anchors, the neighbors, and the imposters, they are indexed using different ranges within the batch. The neighbors and the anchors are selected from the same subject l_+, whereas the imposters are selected such that they are from B different subjects l_k. The single sample test is conducted via simulating an identification scenario, with the following steps.

Step 1: Gallery and Probe Enrollment. The gallery comprises all imposter samples, and one sample $\mathbf{x}_b$ from the union of the anchors and neighbors:

$$\mathcal{G}^b = \mathcal{A}_- \cup \{(\mathbf{x}_b, l_+)\}\,, \tag{4}$$

which contains one sample for every identity in the batch. Then, the rest of the samples from the anchor and the neighbor sets serve as the probe:

$$\mathcal{P}^b = \mathcal{A} \cup \mathcal{A}_+ - \{(\mathbf{x}_b, l_+)\}\,, \tag{5}$$

where all probes are from the same identity l_+.

Step 2: Single Sample Test. Each probe sample $\mathbf{x}_p \in \mathcal{P}^b$ is compared with each gallery sample $\mathbf{x}_g$ by computing the distances $d(\mathbf{x}_p, \mathbf{x}_g)$. Then, the label for $\mathbf{x}_p$ is assigned to be the same as the identity in the gallery that has the smallest distance:

$$\hat{l}_p^b = l_{\hat{g}}, \text{ where } \hat{g} = \arg\min_g d(\mathbf{x}_p, \mathbf{x}_g). \tag{6}$$

The decision is compared with the subject ID ground truth l_+, and the rank-1 hit for sample $\mathbf{x}_p$ is computed as:

$$r_p^b = \begin{cases} 1 & \text{when} \quad \hat{l}_p^b = l_+ \\ 0 & \text{when} \quad \hat{l}_p^b \neq l_+ \end{cases}. \tag{7}$$

Step 3: Steps 1 and 2 are repeated $2B$ times until each element in $\mathcal{A} \cup \mathcal{A}_+$ has been enrolled in the gallery only once. The likelihood ground truth for each sample which has served as

Algorithm 1: Confidence Driven Network

```
input : Batches of image triplets $(I_j, I_i, I_k)$.
output: Parameters $\theta_c, \theta_p$ for FEN $f_c$ and PPN $f_p$.
1  Initialization: step $s = 0$, $\theta_c^s$, $\theta_p^s$ ;
2  while (validation loss decreases) do
3      $s \leftarrow s + 1$;
4      $(\mathbf{z}_j, \mathbf{z}_i, \mathbf{z}_k) \leftarrow f_I(\mathbf{I}_j, \mathbf{I}_i, \mathbf{I}_k)$ ; // Feed-forward of Feature Extractor I;
5      $\mathbf{x}^{s-1} = (\mathbf{x}_j^{s-1}, \mathbf{x}_i^{s-1}, \mathbf{x}_k^{s-1}) \leftarrow f_c(\mathbf{z}_j, \mathbf{z}_i, \mathbf{z}_k)$ ; // Feed-forward of Feature Extractor II;
6      $c^{s-1} = (c_j^{s-1}, c_i^{s-1}) \leftarrow SST(\mathbf{x}^{s-1})$ ; // Single-sample test to estimate confidence ``ground truth'';
       /* Update $\theta_p$:                                  */
7      $\theta_p^s = SGD(\mathcal{L}_p\left(f_p((z_j, z_i), \theta_p^{s-1}), c^{s-1}\right))$;
8      $\hat{c}_i^s = f_p(z_i, \theta_p^s)$ ; // Feed-forward to compute confidence scores;
       /* Update $\theta_c$:                                  */
9      $\theta_c^s = SGD(\mathcal{L}_\mathcal{B}\left(\mathbf{x}^{s-1}, \hat{c}_i^s, \theta_c^{s-1}\right))$ ; // Eq. (2);
10 end
```

probe $\mathbf{x}_p$, is then computed as the average rank-1 hit rate across all testing iterations:

$$c_p = \frac{\sum_{b=1}^{2B} r_p^b}{2B - 1}, \quad (8)$$

where each sample from the anchors and the neighbors has served as a probe for $2B - 1$ times.

IV. Implementation

A. Training

During training, the Feature Extractor I remains frozen, whereas the Feature Extractor II and the PPN are trained jointly so as to enable information sharing through the weight updates. An overview is provided in Algorithm 1.

Lines 4-5: During training, triplets of images are fed to the Feature Extractor I (denoted by f_I) to obtain intermediate features $(\mathbf{z}_j, \mathbf{z}_i, \mathbf{z}_k)$. These are then provided to the Feature Extractor II which outputs high-level representations $\mathbf{x}^{s-1} = (\mathbf{x}_j^{s-1}, \mathbf{x}_i^{s-1}, \mathbf{x}_k^{s-1})$.

Line 6: The single-sample test (denoted by *SST*) is performed as described in Section III-B2 using $\mathbf{x}^{s-1}$, and the ground truth confidence scores $c^{s-1} = (c_j^{s-1}, c_i^{s-1})$ are estimated for both the neighbors and the anchors.

Line 7: The pair $(\mathbf{z}_j, \mathbf{z}_i)$ are fed to the performance predictor to obtain the predicted confidence score $f_p(z_j, z_i)$. Then, the binary cross-entropy loss denoted by L_p is computed and the respective weights θ_p are updated using SGD.

Line 8: The anchor middle-level features z_i is fed forward through the performance predictor to obtain the confidence predictions $\hat{c}_i^s$.

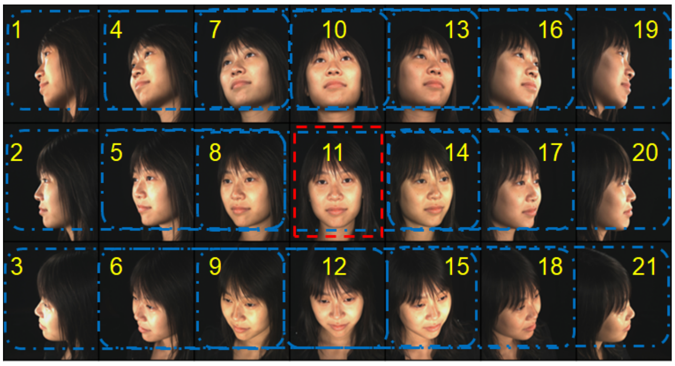


Fig. 3: Using pose 11 as the only image in the gallery, we construct 15 different probe sets. For example, the first set contains the poses $[1, 4, 7]$, the second $[4, 7, 10]$, and so on. In order to ensure that each pose appears the same amount of times in the probe image set, and since pose 11 dose not appear in any probe image sets, we construct some sets with just two images (e.g., $[5, 8, \text{NaN}]$).

Line 9: The confidence score predictions $\hat{c}_i^s$ are utilized along with the high-level features $\mathbf{x}^{s-1}$ to compute the weighted-by-confidence point-to-set triplet loss in Eq. (2). Using this loss, the weights of the Feature Extractor II are updated and then the process is repeated by fetching the next batch of samples.

B. Matching

At matching time, the gallery $\mathcal{G} = \{(\mathbf{x}_m, l_m) \mid m \in [1, N_\mathcal{G}]\}$ contains only one image per subject. For a probe image set, the high-level template and confidence score are computed for each image $\mathcal{P} = \{(\mathbf{x}_n, \hat{c}_n) \mid n \in [1, N_\mathcal{P}]\}$, where n is the index of the image in the set. $\mathcal{P}$ is a set of images belonging to the same subject and $N_\mathcal{G}, N_\mathcal{P}$ are the number of samples in the gallery and the probe, respectively. Then, for each template in the gallery $\mathbf{x}_m$, we compute its distance from the probe template set $\mathcal{P}$ as follows:

$$\mathcal{D}_m\left(\mathbf{x}_m, \mathcal{P}\right) = \frac{\sum_{n=1}^{N_\mathcal{P}} \hat{c}_n d\left(\mathbf{x}_m, \mathbf{x}_n\right)}{\sum_{n=1}^{N_\mathcal{P}} \hat{c}_n}, \quad (9)$$

where $d\left(\mathbf{x}_m, \mathbf{x}_n\right)$ corresponds to the distance between the single template in the gallery and the n^{th} template $\mathbf{x}_n$ of the probe set. The final distance is a fusion of point-to-point distances weighted by the corresponding confidence score $\hat{c}_n$. The distance D_m is computed for every image in the gallery and the one with the minimum distance is selected.

V. Experiments

A. Datasets

The CDN is trained on the CASIA WebFace Database [22], and evaluated on IARPA Janus Benchmark A [23] (with fine-tuning) and UHDB-31 [24] datasets (without fine-tuning).

IJB-A: The IARPA Janus Benchmark A (IJB-A) [23] dataset comprises $5,397$ still images and $2,042$ videos from 500 subjects. The original protocol is designed for set-to-set

TABLE I: Summary of rank-1 accuracy (%) results for Experiment 1. For IJB-A, the values denote average and standard deviation over the 10 splits. In UHDB-31, CDN is tested under three different illumination conditions.

Method	IJB-A	UHDB-31		
		I01	I03	I05
QAN [16]	71.53 ± 2.65	63.25	94.22	89.19
CLFR [19]	83.74 ± 2.72	96.42	98.41	96.38
CDN	**84.56** ± 2.78	**97.41**	**99.47**	**97.18**

face matching. To simulate a multi-probe face recognition paradigm, the $1 : N$ protocol is revised. In every split, one image is randomly sampled for each subject to form the new "search-gallery". The rest of the samples are split into sets (with three images from the same subject in each) to form the new "search-probe". This dataset is used to assess the performance of our approach "in the wild".

UHDB-31: The UHDB-31 [24] dataset comprises 77 subjects. For each subject, a still image is captured from 21 poses, under three different illumination conditions. The original protocol is designed for point-to-point face recognition. To simulate a multi-probe face recognition paradigm, we enroll the frontal face of each subject into the gallery. A set of three images from different poses are sampled and used as a probe. Details about the set sampling rules are provided in Figure 3 and Table II. Since the size of this dataset is small, we do not perform any training. This dataset is used to assess the performance of our approach in a "controlled" environment.

B. Baselines

We select the center loss for face recognition (CLFR) of Wen *et al.* [19] and the quality-aware network (QAN) of Liu *et al.* [16] as baselines. For CLFR we used the model pretrained on the CASIA WebFace database [22] and fine-tuned it on IJB-A. Since CLFR is designed for point-to-point matching, average fusion at a score level is performed for multi-probe face identification. QAN is designed for set-to-set matching. However, it can be easily adapted to point-to-set matching, by applying the quality scores only to the set data. Since the pre-trained model of QAN was not available, we used the code provided by the authors to train QAN from scratch on WebFace, and then fine-tune it on IJB-A. To conduct a fair comparison, CDN is trained using exactly the same protocol.

C. Experimental Results

Experiment 1: The objective of this experiment is to evaluate the identification performance of CDN against state-of-the-art approaches. For the IJB-A dataset, we report average results over ten splits. For UHDB-31, tests are conducted under three different illuminations independently (i.e., I01, I03, and I05 which correspond to lighting originating from the left, the central and the right side, respectively). The corresponding rank-1 accuracy (%) results are reported in Table I. CDN achieved higher performance over the baseline algorithms in both datasets. The performance of QAN reported in this paper is lower than the original paper. One possible reason for this is that the model used in the original QAN paper was trained with additional commercial data. To better understand where the performance gain is originating from we conducted additional experiments.

TABLE II: Rank-1 rate for sets comprising different poses.

Set	Rank-1 Rate (%)		
	QAN [16]	CLFR [19]	CDN
[1, 4, 7]	75.95	94.81	**97.40**
[4, 7, 10]	86.25	**98.70**	**98.70**
[7, 10, 13]	90.00	**98.70**	**98.70**
[10, 13, 16]	91.25	**98.70**	**98.70**
[13, 16, 19]	83.75	96.10	**97.40**
[2, 5, 8]	94.94	**97.40**	**97.40**
[5, 8, NaN]	94.94	**97.40**	**97.40**
[8, NaN, 14]	95.00	**98.70**	**98.70**
[NaN, 14, 17]	95.00	97.40	**98.70**
[14, 17, 20]	95.00	97.40	**98.70**
[3, 6, 9]	86.25	88.31	**94.81**
[6, 9, 12]	91.25	**98.70**	**98.70**
[9, 12, 15]	91.25	**98.70**	**98.70**
[12, 15, 18]	91.25	**98.70**	**98.70**
[15, 18, 21]	85.00	93.51	**96.10**

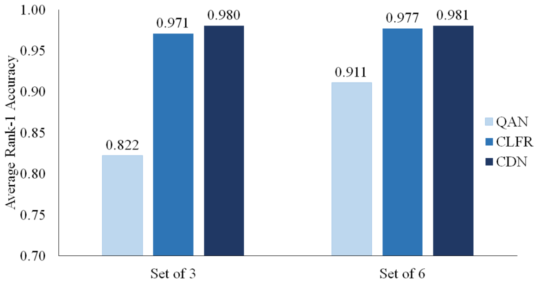


Fig. 4: Impact of image set size on the rank-1 accuracy.

Experiment 2: The objective of this experiment is to assess the performance improvement across sets comprising different poses. The UHDB-31 dataset is used to form sets with three or two different poses as depicted in Figure 3. Results obtained from the central illumination are provided in Table II. In all 15 sets, CDN performed better (or equally) than the other methods and demonstrated superior performance in sets that include extreme poses (i.e., ±90° in yaw and ±30° in pitch).

Experiment 3: The objective of this experiment is to investigate how CDN performs when sets contain images of varying illumination conditions. Each set is selected such that it contains three images of the same pose but with one image per illumination. Rank-1 identification rate for each pose is reported in Table IV. CDN achieved equal or higher accuracy in 17 out of 21 cases.

Experiment 4: The objective of this experiment is to assess the impact of the size of the image set. We evaluate all

TABLE III: Confidence score estimates of the performance predictor under three different illuminations when tested on the UHDB-31 database.

Illumination	Confidence Score
Left Side	0.1000 ± 0.0016
Center	0.1400 ± 0.0027
Right Side	0.0900 ± 0.0014

TABLE IV: Rank-1 rate (%) accuracy results when the set comprises images of the same pose but different illuminations. For example: $[(P1, I01), (P1, I03), (P1, I05)]$.

Method	P1	P4	P7	P10	P13	P16	P19
QAN	22.67	69.23	83.33	89.87	84.81	67.09	16.67
CLFR	**58.33**	**90.67**	100.00	100.00	100.00	92.11	73.33
CDN	52.78	88.00	100.00	100.00	100.00	92.11	**77.33**
	P2	P5	P8	P11	P14	P17	P20
QAN	56.41	91.03	96.20	96.20	96.20	89.87	63.29
CLFR	84.00	**96.00**	98.68	100.00	100.00	96.05	85.53
CDN	84.00	94.67	98.68	100.00	100.00	**97.37**	**86.84**
	P3	P6	P9	P12	P15	P18	P21
QAN	25.97	64.10	86.08	94.94	88.61	70.51	30.77
CLFR	**54.05**	85.33	100.00	100.00	98.68	85.33	42.67
CDN	47.30	**86.67**	100.00	100.00	98.68	**89.33**	**46.67**

three approaches when the set comprises three or six images and report the obtained results in Figure 4. For a set of six images, CDN's performance is still superior to the rest of the methods but not as much as with sets of three. A reason for this is that when sets comprise six images, it is likely that more informative images will be included and so a weighting scheme is less important.

D. How is the confidence score distributed across poses and illuminations?

In Figure 5, we present the confidence score predictions of the PPN for the UHDB-31 dataset, averaged for each pose, along with their standard error. We observe that the performance predictor provides on average higher confidence to images with near-frontal poses. The standard error for near-frontal poses (i.e., Pose IDs 10, 11, 12) is at least twice as much compared to the larger poses (i.e., Pose IDs 1, 2, 3, 19, 20, 21). The reason for this is that there are other factors besides the pose (such as skin color or illumination) that affect the identification performance. For images with extreme poses, the pose is the main reason for the performance drop. Thus, the standard deviation is smaller.

We now focus on the three different sources of lighting for which we provide our results in Table III. The performance predictor favors center lighting on average, and performs similarly in the other two illumination conditions.

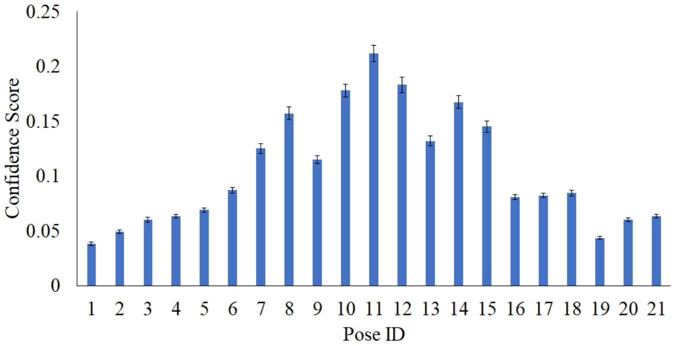
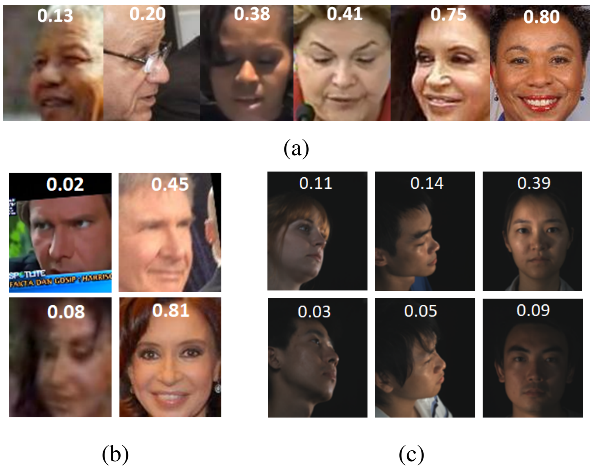


Fig. 5: Confidence score estimates of the performance predictor under different poses when tested on the UHDB-31 database.



(a)

(b)

(c)

Fig. 6: (a): Ranking of randomly selected images from low to high confidence from the IJB-A. (b): For the same subject CDN places more emphasis on samples that do not suffer from occlusions or image blur. (c): For the same pose and the same illumination condition, images from different subjects attain significantly different confidence scores.

E. What is the Performance Predictor Learning?

Looking solely at aggregated pose and illumination results does not provide a full picture of what the performance predictor is learning. Towards this direction, we provide qualitative results in Fig. 6. In Figures (b,c), we observe that: (i) for the same subject CDN assigns significantly less confidence to images with occlusions or blur, and (ii) when both pose and illumination conditions are kept constant, different subjects can have up to three times higher confidence than others.

VI. Conclusions

In this paper, we proposed the Confidence Driven Network (CDN): a framework that jointly learns a feature vector and a confidence score for each image in a multi-probe face identification setup. To learn the confidence score, a single sample-test mechanism was introduced to quantify the discriminative level of the template. CDN improves the rank-1 identification rate for multi-probe face identification in the selected datasets. We observed that CDN exhibits superior performance when image sets contain large pose variations, whereas for large image sets, the improvements are not significant. We identified several visual properties of the original image (e.g., pose, illumination, skin color) that affect the confidence score and provided quantitative and qualitative results to support our claims. Which exactly are these properties and to what extent they contribute to the identification accuracy is a topic for future research.

Acknowledgement: This material is based upon work supported by the U.S. Department of Homeland Security under Grant Award Number 2015-ST-061-BSH001. This grant is awarded to the Borders, Trade, and Immigration (BTI) Institute: A DHS Center of Excellence led by the University of Houston, and includes support for the project Image and Video Person Identification in an Operational Environment: Phase I awarded to the University of Houston. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security

References

[1] O. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. British Machine Vision Conference*, Swansea, UK, 2015, pp. 1–12. 1, 2

[2] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, 2015, pp. 815–823. 1, 2

[3] X. Zhu, X.-Y. Jing, F. Wu, Y. Wang, W. Zuo, and W.-S. Zheng, "Learning heterogeneous dictionary pair with feature projection matrix for pedestrian video retrieval via single query image," in *Proc. AAAI Conference on Artificial Intelligence*, San Francisco, CA, 2017, pp. 4341–4347. 1

[4] Y. Zhu, Z. Zheng, Y. Li, G. Mu, S. Shan, and G. Guo, "Still to video face recognition using a heterogeneous matching approach," in *Proc. International Conference on Biometrics: Theory, Applications and Systems*, Arlington, VA, 2015, pp. 1–6. 1

[5] H. Cevikalp and B. Triggs, "Face recognition based on image sets," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, 2010, pp. 2567–2573. 1

[6] M. Yang, P. Zhu, L. Van Gool, and L. Zhang, "Face recognition based on regularized nearest points between image sets," in *Proc. IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, Shanghai, China, 2013, pp. 1–7. 1

[7] P. Zhu, L. Zhang, W. Zuo, and D. Zhang, "From point to set: Extend the learning of distance metrics," in *Proc. IEEE International Conference on Computer Vision*, Sydney, Australia, 2013, pp. 2664–2671. 1

[8] M. Leng, P. Moutafis, and I.A. Kakadiaris, "Joint prototype and metric learning for set-to-set matching: Application to biometrics," in *Proc. International Conference on Biometrics: Theory, Applications and Systems*, Arlington, VA, 2015, pp. 1–8. 1

[9] J.-T. Chien and C.-C. Wu, "Discriminant waveletfaces and nearest feature classifiers for face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 12, pp. 1644–1649, 2002. 1

[10] J. Hamm and D. Lee, "Grassmann discriminant analysis: A unifying view on subspace-based learning," in *Proc. International Conference on Machine learning*, Helsinki, Finland, 2008, pp. 376–383. 1

[11] M. Harandi, R. Hartley, C. Shen, B. Lovell, and C. Sanderson, "Extrinsic methods for coding and dictionary learning on grassmann manifolds," *International Journal of Computer Vision*, vol. 114, no. 2, pp. 113–136, 2015. 1

[12] Z. Huang, R. Wang, S. Shan, and X. Chen, "Learning Euclidean-to-Riemannian metric for point-to-set classification," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, 2014, pp. 1677–1684. 1

[13] Y. Li, R. Wang, Z. Huang, S. Shan, and X. Chen, "Face video retrieval with image query via hashing across Euclidean space and Riemannian manifold," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, 2015, pp. 4758–4767. 1

[14] W. Wang, R. Wang, S. Shan, and X. Chen, "Discriminative covariance oriented representation learning for face recognition with image sets," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, Hawaii, 2017, pp. 5749–5758. 1

[15] T. Hassner, I. Masi, J. Kim, J. Choi, S. Harel, P. Natarajan, and G. Medioni, "Pooling faces: Template based face recognition with pooled face images," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Las Vegas, NV, 2016, pp. 127–135. 2

[16] Y. Liu, J. Yan, and W. Ouyang, "Quality aware network for set to set recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, 2017, pp. 4694–4703. 2, 5

[17] J. Yang, P. Ren, D. Zhang, D. Chen, F. Wen, H. Li, and G. Hua, "Neural Aggregation Network for Video Face Recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, Hawaii, 2017, pp. 1–8. 2

[18] S. Zhou, J. Wang, J. Wang, Y. Gong, and N. Zheng, "Point to set similarity based deep feature learning for person re-identification," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, 2017, pp. 5028–5037. 2

[19] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. European Conference on Computer Vision*, Amsterdam, The Netherlands, 2016, pp. 499–515. 2, 5

[20] S. Zhou, J. Wang, R. Shi, Q. Hou, Y. Gong, and N. Zheng, "Large margin learning in set to set similarity comparison for person re-identification," *IEEE Transactions on Multimedia*, vol. 20, no. 3, pp. 593–604, 2017. 2

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, 2016, pp. 770–778. 2

[22] D. Yi, Z. Lei, S. Liao, and S. Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014. 4, 5

[23] B. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, M. Burge, and A. Jain, "Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, 2015, pp. 1931–1939. 4

[24] H. Le and I.A. Kakadiaris, "UHDB31: A dataset for better understanding face recognition across pose and illumination variation," in *Proc. IEEE International Conference on Computer Vision Workshops*, Venice, Italy, 2017, pp. 2555–2563. 4, 5