Towards Universal Representation Learning for Deep Face Recognition (Supplementary Material)

Yichun Shi[1,2] Xiang Yu[2] Kihyuk Sohn[2] Manmohan Chandraker[2] Anil K. Jain[1]
[1]Michigan State University [2]NEC Labs America

A. Proofs

A.1. Confidence-aware Identification Loss

A.1.1 Single Embedding

Let $\mathcal{Z}$ denotes the latent embedding space and $\mathbf{z}$ a variable from the $\mathcal{Z}$. Different $\mathbf{z}$ represents different facial appearance. Given a face image $\mathbf{x}$, the network θ estimates the encoded appearance $p_\theta(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}; \mathbf{f}_i, \sigma_i^2\mathbf{I})$ where $\mathbf{f}_i$ is the embedded feature vector while σ_i^2 is the uncertainty of the representation. Let y denotes the identity label and C the number of identities. For each class $j \in \{1, 2, \dots, C\}$, we maintain a prototype vector $\mathbf{w}_j$, which represents the intrinsic appearance of the j^{th} identity in the latent space. In other words, $p(\mathbf{z}|y = j) = \delta(\mathbf{z} - \mathbf{w}_j)$, where δ is the Dirac delta function. Assuming an non-informative prior $p(\mathbf{z})$, the likelihood of $\mathbf{x}_i$ being a sample of j^{th} class is given by:

$$\begin{aligned} p(\mathbf{x}_i|y = j) &= \int p(\mathbf{x}_i|\mathbf{z})p(\mathbf{z}|y = j)d\mathbf{z} \\ &= \int \frac{p_\theta(\mathbf{z}|\mathbf{x}_i)p(\mathbf{x}_i)}{p(\mathbf{z})}\delta(\mathbf{z} - \mathbf{w}_j)d\mathbf{z} \qquad (15) \\ &= p_\theta(\mathbf{w}_j|\mathbf{x}_i)p(\mathbf{x}_i) \\ &\propto p_\theta(\mathbf{w}_j|\mathbf{x}_i) \end{aligned}$$

where

$$p_\theta(\mathbf{w}_j|\mathbf{x}_i) = \frac{1}{(2\pi\sigma_i^2)^{\frac{D}{2}}} \exp(-\frac{\|\mathbf{f}_i - \mathbf{w}_j\|^2}{2\sigma_i^2}). \qquad (16)$$

Therefore, the posterior probability of $\mathbf{x}_i$ belonging to the j^{th} class is:

$$\begin{aligned} p(y = j|\mathbf{x}_i) =& \frac{p(\mathbf{x}_i|y = j)p(y = j)}{\sum_{c=1}^{N} p(\mathbf{x}_i|y = c)p(y = c)} \qquad (17) \\ =& \frac{p_\theta(\mathbf{w}_j|\mathbf{x}_i)}{\sum_{c=1}^{N} p_\theta(\mathbf{w}_c|\mathbf{x}_i)} \qquad (18) \\ =& \frac{\exp(-\frac{\|\mathbf{f}_i - \mathbf{w}_j\|^2}{2\sigma_i^2})}{\sum_{c=1}^{N} \exp(-\frac{\|\mathbf{f}_i - \mathbf{w}_c\|^2}{2\sigma_i^2})}, \qquad (19) \end{aligned}$$

which is the Equation (4) in the main paper.

A.1.2 Multiple Sub-embeddings

For a sub-embedding network, the likelihood function becomes:

$$p_\theta(\mathbf{w}_j|\mathbf{x}_i) = \prod_{k=1}^{K} \frac{1}{(2\pi\sigma_i^{(k)2})^{\frac{D}{2K}}} \exp(-\frac{\left\|\mathbf{f}_i^{(k)} - \mathbf{w}_j^{(k)}\right\|^2}{2\sigma_i^{(k)2}}). \qquad (20)$$

And therefore, the posterior classification probability is:

$$\begin{aligned} p(y = j|\mathbf{x}_i) =& \frac{p_\theta(\mathbf{w}_j|\mathbf{x}_i)}{\sum_{c=1}^{N} p_\theta(\mathbf{w}_c|\mathbf{x}_i)} \qquad (21) \\ =& \frac{\exp(\mathbf{a}'_{i,j})}{\sum_{c=1}^{N} \exp(\mathbf{a}'_{i,j})}, \qquad (22) \end{aligned}$$

where

$$\mathbf{a}'_{i,j} = -\sum_{k=1}^{K} \frac{\left\|\mathbf{f}_i^{(k)} - \mathbf{w}_j^{(k)}\right\|^2}{2\sigma_i^{(k)2}}. \qquad (23)$$

Given that $\left\|\mathbf{f}_i^{(k)}\right\|^2 = \left\|\mathbf{w}_j^{(k)}\right\|^2 = 1$ and $s_i^{(k)} = \frac{1}{\sigma_i^{(k)2}}$, Equation (23) becomes:

$$\mathbf{a}'_{i,j} = \sum_{k=1}^{K} s_i^{(k)} \mathbf{f}_i^{(k)T} \mathbf{w}_j^{(k)} - \sum_{k=1}^{K} s_i^{(k)}. \qquad (24)$$

The second term is cancelled out when computing the probability. By further incorporating the margin m in to Equation (24) and taking the average score instead of the sum, one could derive the Equation (8) in the main paper.

A.2. Gradient of the sub-embeddings

Here, we try to understand how the confidence helps the training by looking at the gradient of the sub-embeddings in Equation (8) in the main paper. Notice that we have

$$\frac{\partial \mathcal{L}_{idt}}{\mathbf{a}_{i,j}} = p_{i,j} - \delta_{y_i,j}, \qquad (25)$$

where $\delta_{y_i,j}$ is 1 if $y_i = j$ and 0 otherwise. $p_{i,j} = p(y = j|\mathbf{x}_i)$ is the posterior classification probability. Since $\frac{\partial \mathbf{a}_{i,j}}{\partial \mathbf{w}_j^{(k)}} = s_i^{(k)} \mathbf{f}_i^{(k)}$ and $\frac{\partial \mathbf{a}_{i,j}}{\partial \mathbf{f}_j^{(k)}} = s_i^{(k)} \mathbf{w}_j^{(k)}$, we have

$$\begin{aligned} \frac{\partial \mathcal{L}_{idt}}{\partial \mathbf{w}_j^{(k)}} &= s_i^{(k)} (p_{i,j} - \delta_{y_i,j}) \mathbf{f}_i^{(k)} \\ \frac{\partial \mathcal{L}_{idt}}{\partial \mathbf{f}_j^{(k)}} &= s_i^{(k)} ((p_{i,y_i} - 1) \mathbf{w}_{y_i}^{(k)} + \sum_{j \neq y_i} p_{i,j} \mathbf{w}_j^{(k)}) \end{aligned} \tag{26}$$

From Equation (26), it can be seen that the gradient of the prototypes and sub-embeddings depend on both the confidence value and the classification probability. In particular, confidence value $s_i^{(k)}$ serves as a gating parameter during the back-propagation. In such a way, the prototypes would be affected more by the confident samples than the not confident ones. Similarly, the confident sub-embedding would also have a larger impact on the prototype.

B. Additional Implementation Details

The backbone of our embedding network θ is a modified 100-layer ResNet in [1]. The network is split into two different branches after the last convolution layer, each of which includes one fully connected layer. The first branch outputs a 512-D vector, which is further split into 16 sub-embeddings. The other branch outputs a 16-D vector, which are confidence values for the sub-embeddings. The $\exp$ function is used to guarantee all the confidence values $s_i^{(k)}$ are positive. The model θ_A that we used for mining additional variations is a four layer CNN. The four layers have 64, 128, 256 and 512 kernels, respectively, all of which are 3×3.

C. Ablation Study on Variation Decorrelation Loss

In Table 1 we show the results of training with different number of variations for the variation decorrelation loss. The base model in the first line is a model trained with all the modules proposed in the paper except variation decorrelation loss. The second to fourth line show the results of using different number augmentable variations (blur, occlusion and pose) and additional variations (gender, age and smiling). It can be seen that with more variation added into the training, the decorrelation becomes more effective and leads to a better performance.

D. Additional Results on IJB-S

Table 2 shows more results of our models as well as state-of-the-art methods on the IJB-S dataset. The "Surveillance-to-Single" protocol uses one single image in the gallery templates while the "surveillance-to-booking" use a set of face images with different poses in the gallery templates. Compared with our own baseline, significant performance boost can be observed on all the metrics, which proves the efficacy of the proposed method. Compared with the state-of-the-art methods, our final model achieves better performance on most of the metrics.

E. Visualization of Sub-embedding Confidence

Figure 1 shows the distributions of confidence values during training. It can be observed that the confidences of different sub-embeddings not only have different distributions, but also vary in terms of which kind of images have high/low confidence. Since the confidence guides the training signal of the corresponding features, this reflects that the sub-embeddings learn different features complementary to each other for better identification performance.

F. Visualization of Uncertainty

In Figure 2 and Figure 3, we show more results of uncertainty heatmaps. Overall, we can see that distinguishable face images have low uncertainty on most sub-embeddings. Faces with larger variations have some sub-embeddings with low uncertainty, depending on which kind of variation is present. For images with extremely large variations, high uncertainty is observed on all the sub-embeddings.

G. Visualization of Face Representations

In Figure 4, we show the t-SNE visualization of the embeddings from the baseline (with augmentation) method as well as the proposed method. The original training samples and the augmented ones are shown in circle and triangle, respectively. Notice that some augmented samples are hard to recognize and are close to be noises. Thus, by assuming an equal confidence for all the samples, the baseline method fails to converge to a good local minimum and many augmented samples cluster together in a small area. In comparison, by focusing more on the high-quality samples, the proposed method learns a more discriminative feature space. Although noisy outliers still exist in the proposed method, they are usually close to their own identities' samples.

H. Image Examples From the Testing Datasets

Figure 5 shows more image examples from different types of the dataset. The images in the LFW (Type I) dataset are mostly high quality face images with limited variations. Therefore, different models in our experiment all achieve similar performance on this dataset. The images in the IJB-A (Type II) show more variations, some of which are extremely challenging. This requires the representation model to be able to perform a cross-domain matching between images of high quality and low quality. Further, the TinyFace and IJB-S (Type III) datasets are mostly composed of low-quality

Method		TinyFace		IJB-S	
Augmentable Variations	Additional Variations	Rank1	Rank5	Rank1	Rank5
0	0	55.04	60.97	59.71	66.32
3	0	54.99	61.32	62.22	67.03
3	1	61.80	67.94	62.30	67.51
3	3	63.89	68.67	61.98	67.12

Table 1: Gradually adding more variations into variation decorrelation loss. All of the models use all the other modules proposed in the paper.

faces. This requires the face representation to be invariant to large variations that can hardly be found in the public training datasets.

References

[1] Jiankang Deng, Jia Guo, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *CVPR*, 2019. 2, 4

[2] Nathan D. Kalka, Brianna Maze, James A. Duncan, Kevin J. OConnor, Stephen Elliott, Kaleb Hebert, Julia Bryan, and Anil K. Jain. IJB-S : IARPA Janus Surveillance Video Benchmark . In *BTAS*, 2018. 4

[3] Yichun Shi and Anil K Jain. Probabilistic face embeddings. In *ICCV*, 2019. 4

Method	Training Data	Surveillance-to-Single					Surveillance-to-Booking				
		Rank-1	Rank-5	Rank-10	1%	10%	Rank-1	Rank-5	Rank-10	1%	10%
PFE [3]	4.4M	50.16	58.33	62.28	31.88	35.33	53.60	61.75	64.97	35.99	39.82
ArcFace [1]+	5.8M	57.35	64.42	68.36	41.85	50.12	57.36	64.95	68.57	41.23	49.18
Ours (Basline)	4.8M	47.94	55.40	59.37	25.60	36.03	37.14	46.75	51.59	24.75	31.10
Ours (Baseline + VA)	4.8M	60.61	66.53	68.57	31.97	44.25	51.27	58.94	63.25	31.19	44.22
Ours (all)	4.8M	58.94	65.48	68.31	37.57	50.17	60.74	66.59	68.92	37.11	51.00
Ours (all) + PA	4.8M	59.79	65.78	68.20	41.06	53.23	61.98	67.12	69.10	42.73	53.48

Table 2: Performance comparison on the IJB-S dataset. The performance is reported in terms of rank retrieval (closed-set) and TPIR@FPIR (open-set) instead of the media-normalized version [2]. The numbers "1%" and "10%" in the second row refer to the FPIR. "+" indicates the testing performance by using the released models from corresponding authors.

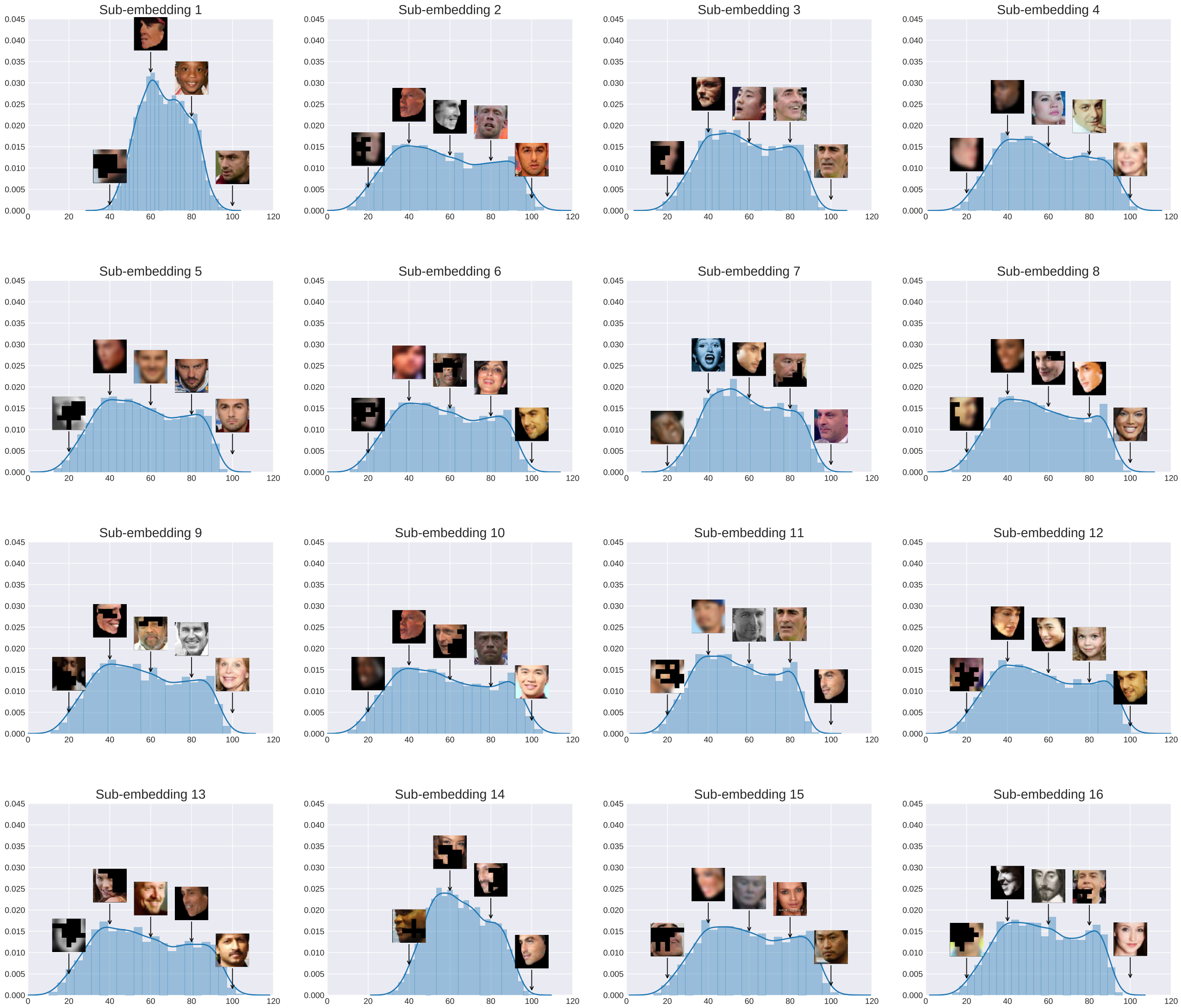


Figure 1: Visualization of sub-embedding confidence on training samples.

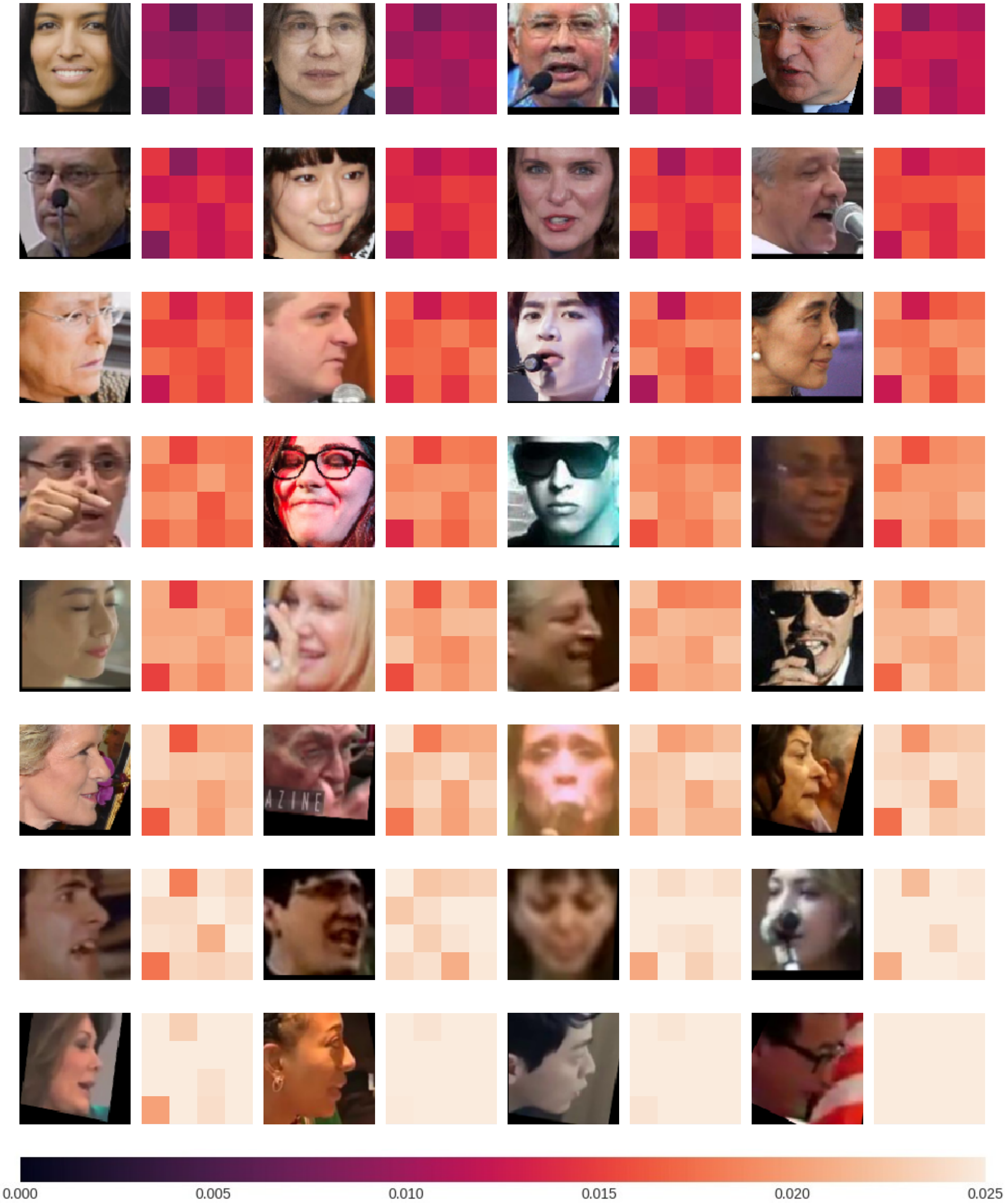


Figure 2: Visualization of sub-embedding uncertainty on testing images.

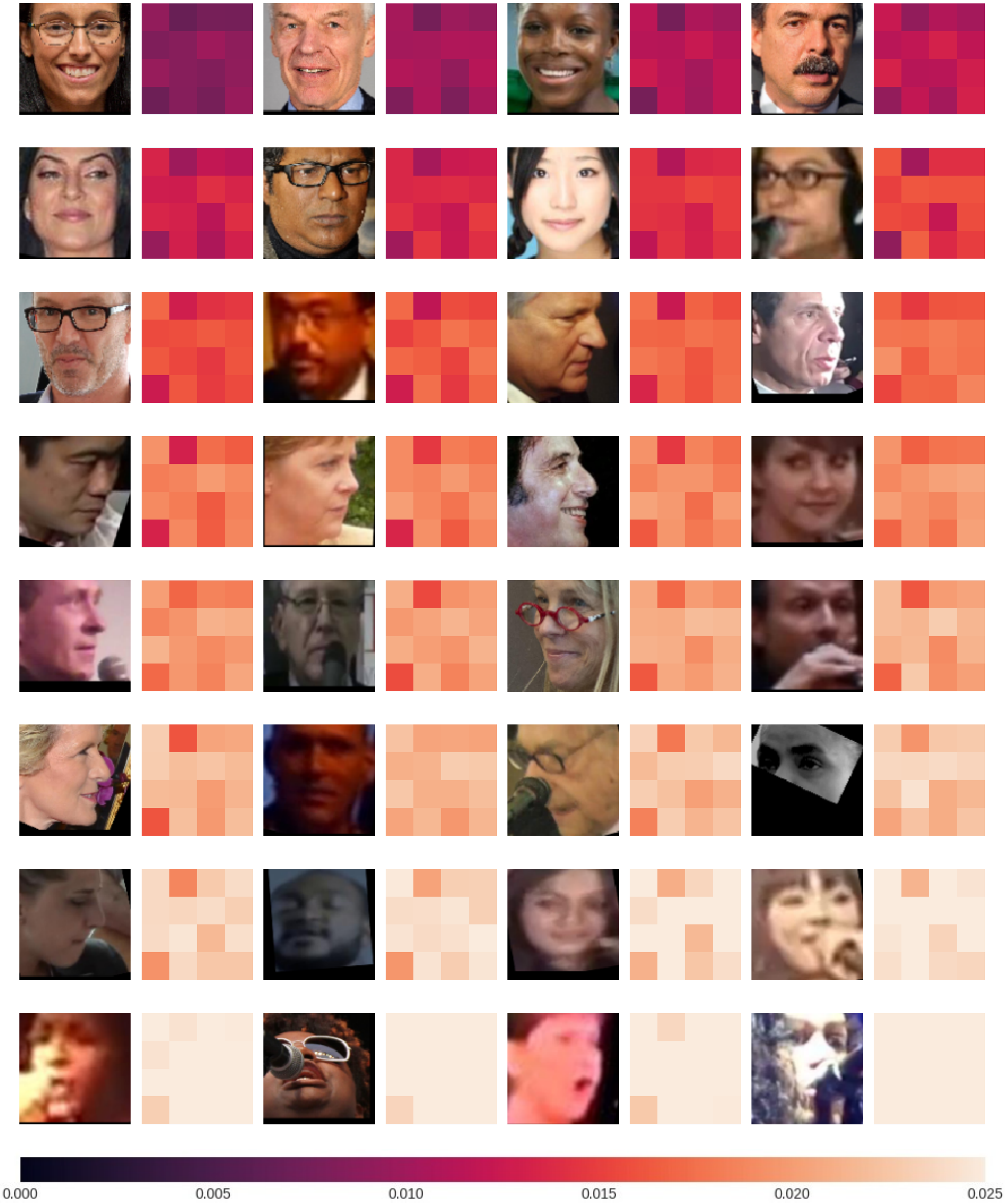
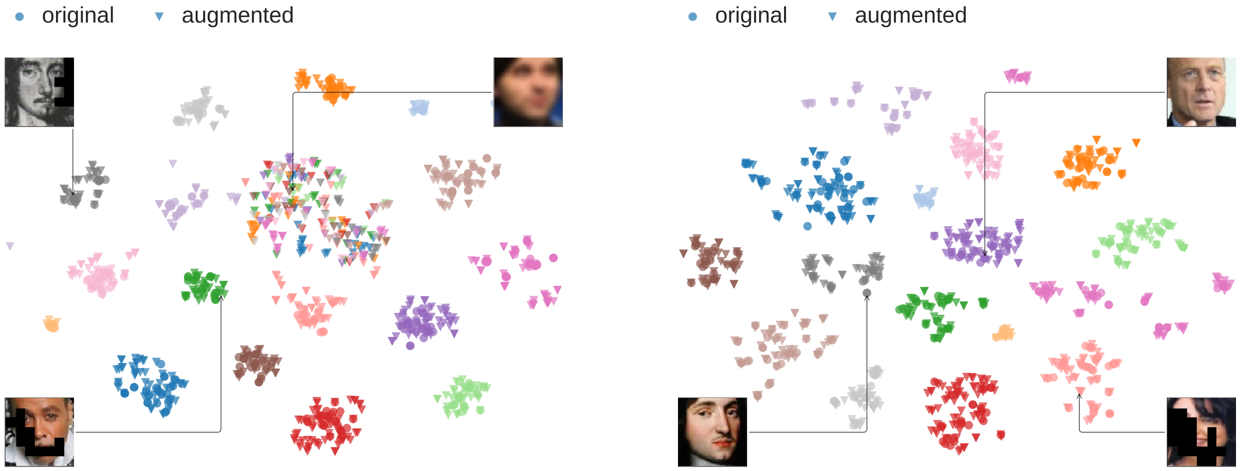


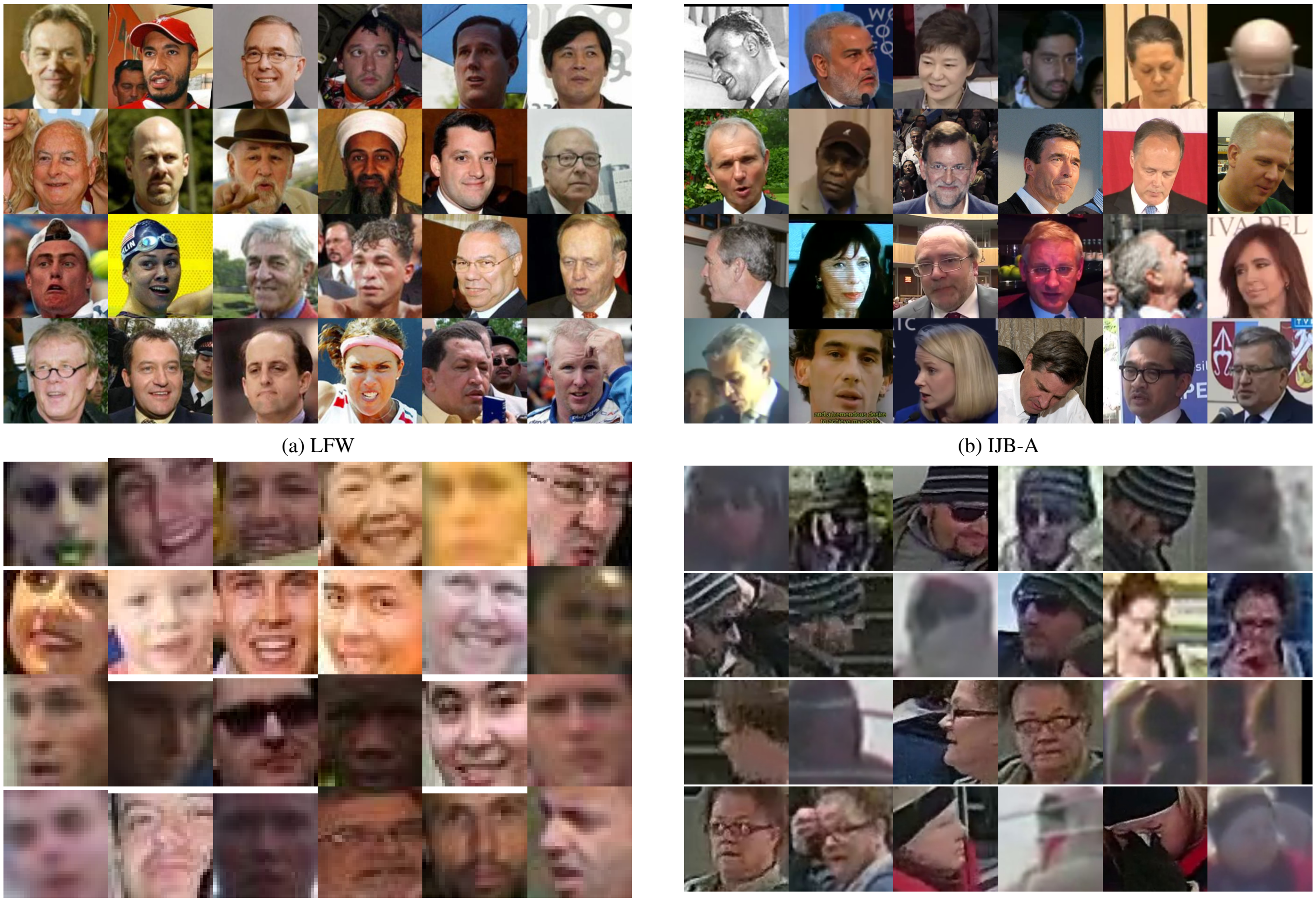
Figure 3: Visualization of sub-embedding uncertainty on more testing images.



(a) Baseline

(b) Proposed

Figure 4: t-SNE visualization of the features in a 2D space. Colors indicate the identities. Original training samples and augmented training samples are shown in circle and triangle, respectively.



(a) LFW

(b) IJB-A

(c) TinyFace

(d) IJB-S

Figure 5: Examples images from the testing datasets.