Harnessing Unrecognizable Faces for Improving Face Recognition

Siqi Deng Yuanjun Xiong[‡] Meng Wang Wei Xia[‡] Stefano Soatto
AWS AI Labs

Abstract

The common implementation of face recognition systems as a cascade of a detection stage and a recognition or verification stage can cause problems beyond failures of the detector. When the detector succeeds, it can detect faces that cannot be recognized, no matter how capable the recognition system is. Recognizability, a latent variable, should therefore be factored into the design and implementation of face recognition systems. We propose a measure of recognizability of a face image that leverages a key empirical observation: An embedding of face images, implemented by a deep neural network trained using mostly recognizable identities, induces a partition of the hypersphere whereby unrecognizable identities cluster together. *This occurs regardless of the phenomenon that causes a face to be unrecognizable, be it optical or motion blur, partial occlusion, spatial quantization, or poor illumination. Therefore, we use the distance from such an "unrecognizable identity" as a measure of recognizability, and incorporate it into the design of the overall system. We show that accounting for recognizability reduces the error rate of single-image face recognition by 58% at FAR=1e-5 on the IJB-C Covariate Verification benchmark, and reduces the verification error rate by 24% at FAR=1e-5 in set-based recognition on the IJB-C benchmark.*

1. Introduction

We aim at making face recognition systems easier to use responsibly. This requires not just reducing the error rate, but also producing interpretable performance metrics, with estimates of when recognition can be performed reliably, or otherwise should not be attempted, or deemed unreliable.

In most face recognition systems[8, 10, 38], each image is first fed to a face detector (FD), that returns the location and shape of a number of bounding boxes likely to portray faces. Those, together with the image, are then fed to a downstream face recognition (FR) module that returns either one of K labels corresponding to identities in a database (search), or a binary label corresponding to whether or not the bounding box matches a given identity (verification). FD and FR are typically non-interacting modules trained on different datasets: The FD is tasked with finding faces no matter whether they are recognizable. The FR is tasked with mapping each detection onto one of K identities. An obvious failure mode of such cascaded systems is when the FD is *wrong*: If a face is not detected, obviously it cannot be correctly recognized. If a detected bounding box does not show a face, the FR system will nonetheless map it to one of the known identities, unless post-processing steps are in place, typically involving a threshold on some confidence measure.

But even when the FD is *right*, there remain the following problems:

First, while an image may contain enough information to decide that there *is* a face, it may not contain enough to determine *whose* face it is, regardless of how good the FR system is. This creates a gap between the FD, tasked to determine that there is a face regardless of whether it is recognizable, and the FR, tasked to recognize it. An FR system should not try to recognize a face that is not recognizable. [‡] Accordingly, *how can we measure and account for the recognizability of a face image in face recognition?*

Second, failure to take into consideration recognizability can lead to misleading results in face recognition benchmarks. Unrecognizable faces due to optical, atmospheric, or motion blur, spatial quantization, poor illumination, partial occlusion, etc, are typically used to train and score FD systems, but FR systems are trained using recognizable faces, lest one could not establish ground truth. Accordingly, *how can we balance the reward in detecting unrecognizable faces with the risk of failure to recognize them?*

Third, failure to account for the recognizability of a face can have consequences beyond the outcome of FR on that face. Consider the problem of set-based face recognition where the identity is to be assigned *not* to a single image, but to a small collection of images known to come from the same identity, some of which unrecognizable. These may

[‡]Work done when at Amazon.

[‡]An optimal (Bayesian) FR system would forgo the FD and marginalize over all possible locations and shapes, which is obviously intractable. But conditioning on the presence of a face, rather than marginalizing it, is not the only problem: There is another latent variable, *"recognizability"* that is unaccounted for and instead assumed to be true by the FR.

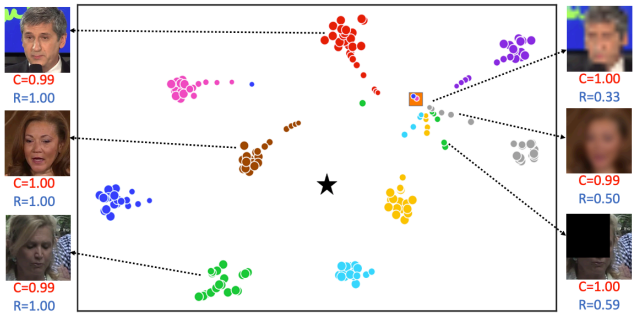

Figure 1. Hypersphere embeddings [38] of different faces from the IJB-C dataset (colored clusters) visualized as circles with area proportional to recognizability, using t-SNE [27]. As the images are perturbed artificially, becoming increasingly unrecognizable, their embeddings migrate to join a common cluster (orange square). Such an "unrecognizable identity" (UI) is described in Sec. 2.2.1 and distinct from the centroid of the recognizable embeddings (black pentagram). Note the difference between *face detection confidence (C)* and *embedding based recognizability score (R)*. The former is the output of a face detector and measures the likelihood that the image contains a face while the latter measures if the face can be recognized.

be frames of a video, some of which affected by motion blur or partial occlusions. It may seem that using all the available data can only improve the quality of the decision. However, uniform averaging does not factor in the differences with the set to guarantee optimal performance. Accordingly, *how should one combine images in set-based face recognition, assuming we have available a measure of recognizability?*

1.1. Main Hypothesis and Empirical Observation

The three questions above point to the need to explicitly represent "recognizability" as a latent variable. As was the case for the FD, marginalization is impractical. We instead hypothesize that recognizability can be quantified inferentially. A measure of recognizability could then be used to complete the hypothesis space for FR, effectively adding an additional class for *unrecognizable identities*, akin to open-set classification. An estimate of recognizability would also allow us to correctly weigh the influence of the detector in the overall FR results. Finally, it would allow proper weighting of different samples in set-based face recognition.

So, recognizability and the addition of an unrecognizable identity (UI) class would address the issues set forth in the introduction. But *what is the unrecognizable identity?* Is it an actual identity or just a moniker for the interstitial space around the decision boundaries among all known identities?

If we represent each face via an embedding in a compact metric space, such as the hypersphere, perturbing the image until it is unrecognizable moves the corresponding embedding close to a decision boundary. So, it is reasonable to expect that the embeddings of UIs distribute along the boundaries of decision regions with no particular relation to each other: Distant identities, when perturbed, would become distant UIs. Instead, we observe the following phenomenon: *When training an FR system without any UIs, and using the resulting embedding to cluster identities (both recognizable and not), UIs cluster together in representation space*, despite being unrecognizable versions of different identities that may otherwise be far in representation space.

Figure 2. During face clustering, faces with low recognizability scores are grouped into one cluster by distance-based clustering method (left), compared with clusters of faces with known identities (right). We refer to the former as the unrecognizable identity cluster (UI) cluster. The UI cluster includes images subject to occlusion, optical or motion blur, low resolution, poor illumination, etc. So the clustering of UIs is not simply due to visual similarity.

This phenomenon, illustrated in Fig. 1, is counter-intuitive at many levels: First, UIs do not distribute near the boundary between different identities, but rather close to each other and far from the corresponding identities. This happens without imposing any loss on the distance among UIs – for they are not even included in the training set for the FR – and is likely made possible by the geometry of the high-dimensional hypersphere.[‡] Second, this phenomenon is not only due "low-quality" images of UI being *visually similar to each other*. Unlike other domains where motion-blurred images form their own cluster, distinct from the low-resolution cluster, here the UI cluster is *highly heterogeneous,* with images that exhibit occlusion, optical or motion blur, low resolution, poor illumination, etc. So, the clustering of UIs is not simply due to visual similarity. Samples of these phenomena are illustrated in Fig. 2 and Sec. 2.2.1. We conjecture that this behavior is specific to FR, which is a fine-grained categorization task, where nuisance variability can cause coarse-grained perturbations that move the corresponding samples out of domain.

We are now ready to tackle the main goal: Leverage the key empirical observation above to *harness unrecognizable faces to improve face recognition.*

[‡]This phenomenon is unrelated to the known collapse of representations of low-quality images towards the origin of a linear embedding space [29], as here the embedding is constrained to be on the hypersphere.

1.2. Related Work and Contributions

The face recognition literature is gargantuan; by necessity, we limit our review to the most closely related methods, cognizant that we may omit otherwise important work. Like most, we use a deep neural network (DNN) model to compute our embedding function mapping an image x and a bounding box b to a (pseudo-)posterior score $\phi(x_b) \propto \log P(y|x,b)$ where $y \in \{1,\ldots,K\}$ denotes the identity for the case of search, and $K = 2$ for verification [9, 11, 20, 43, 30, 33, 25, 39, 49, 26, 41, 19]. Our work relates to efforts to understand the effect of image quality on face recognition: Some explicitly modeled image quality in face aggregation [13, 44, 45, 47, 23], others utilized correlation among images from the same person to improve set-based [21, 22] or video-based [32] face recognition. More recently, [15, 16] developed face quality assessment tools and methods [36], probabilistic representation to model data uncertainty [35, 2, 34], as well as explicit handling of quality variations with sub-embeddings [35].

Our contribution is three-fold, corresponding to the three questions posed in the introduction:

1) We propose a measure of recognizability that leverages the existence of a single UI cluster in the learned embedding. The *"embedding recognizability score"* (ERS) is simply the Euclidean (cordal) distance of the embedding of an image from the UI cluster center in the hypersphere.

2) We use the ERS to mitigate the detrimental effect of using different datasets for FD and FR. This results in 58% error reduction at FAR=1e-5 (Table 1) in single-image face recognition, without impacting the performance of the detector.

3) We propose an aggregation method for set-based face recognition, by simply calculating the weighted average relative to the ERS, and report 24% error reduction (at FAR=1e-5 on IJB-C) compared with uniform averaging.

We emphasize that unrecognizable faces are still included in the evaluation and the improvement is due to the proposed matching method using ERS (Sec. 2.2.2).

1.3. Implications on Bias and Fairness

As with any data-driven algorithmic method, a trained FR system is subject to statistical bias due to the distribution of the training set, and more general algorithmic bias engendered by the choice of models, inference criteria, optimization method, and interface. In addition to those externalities, there can be intrinsic sources of bias that exist before any algorithm is implemented or dataset collected: Light interacts with matter in a way that depends on the sub-surface scattering properties of materials. Therefore suboptimal calibration of the imaging tools could affect the recognizability of *any* vision system. Moreover, clothing, makeup, and accessories can impact recognizability by causing occlusions of facial features. We, therefore, expect that the UI will *not* represent an unbiased sampling of the population and reflect physical and phenomenological characteristics of the scene, the illuminant, the sensor, and the capture method, regardless of the algorithm used to infer recognizability. *In addition*, recognizability is based on the statistical properties of the embedding, which is subject to the usual data and algorithmic biases. It is not clear how to balance different populations in the UI. We defer these important and delicate questions to the many studies on bias and fairness[1, 40, 46, 4], and focus on the orthogonal question of how to deal with recognizability irrespective of what algorithm or system is used for recognition.

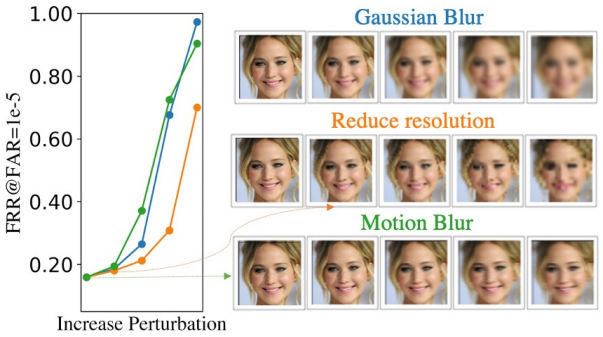

Figure 3. Face verification accuracy can be heavily impacted by image perturbation, evidenced by IJB-C Template-based Face Verification benchmark results: the left x-axis shows perturbation level low to high and the y-axis FRR@FAR=1e-5, with corresponding images on the right. Perturbation types are color-coded.

2. Face Recognizability in Face Recognition

To consider face recognizability in the context of face recognition systems, we first study the impact of face recognizability degradation on FR accuracy and reveal why it is important to account for input data recognizability. Then we propose a measure of face recognizability depending on the face embedding model. By factoring in the recognizability measure for FR prediction decisions, we propose methods to mitigate the detriment of unrecognizable faces to FR systems.

2.1. Observing Recognizability

The observation that **UIs cluster together** can be illustrated using face embeddings [38] on images of 8 randomly sampled celebrities from the IJB-C dataset. We synthesize UI images by perturbing images with increasing Gaussian blur, motion blur, occlusions, etc. The t-SNE visualization of the embeddings is shown in Fig. 1, where recognizable faces form separate clusters corresponding to their identities, but as they become increasingly unrecognizable they do not distribute around the boundary between identities, nor around their centroid, but rather around a distinct cluster, the UI cluster.

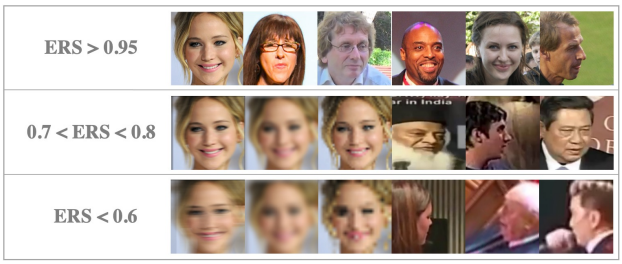

Figure 4. Sample images from IJB-C dataset, grouped by embedding based recognizability scores (ERS): high (>0.95), middle (0.7−0.8), and low (<0.6). Images with low ERS are hard to recognize even for human viewers.

To validate this result, we perform large-scale face clustering [6] on datasets where there are artificial or natural low recognizability (LR) images of faces or non-faces. To exemplify artificial LR images, we take a subset of 10K faces from DeepGlint [7], a face embedding training dataset. We randomly apply Gaussian blur, motion blur, resolution reduction, rotation, affine transformation, or occlusion to corrupt 1K of the faces. After running face clustering with the HAC algorithm [6] on the normalized features, a UI cluster emerges. The cluster size is larger than all the other clusters by more than two orders of magnitude. For natural LR images, we run clustering on face crops detected by a state-of-the-art FD model on the WIDERFace [48] dataset. Again, we obtain a heterogeneous cluster comprising solely unrecognizable and visually dissimilar faces. The identity-agnostic UI images distribute closely in one cluster and away from identity-based clusters. This corroborates two typical types of face quality-related errors in FR reported in several prior works [2, 34]: false positive matches between low-quality faces of different ids and false negatives between high and low-quality faces of the same id.

The clustering results on face datasets are surprising as low-quality images (for instance, a quality degraded ImageNet database) ordinarily tend to cluster by appearance: blurred images form a cluster that is distinct from that of dark images, and that of motion-blurred images, etc. We conjecture that the existence of UI has to do with the fine-grained nature of the face domain because we discover similar phenomena also emerge in other fine-grained recognition tasks like person re-identification and fashion retrieval.

What impact could the UIs have on the downstream FR task? To quantify this, we add recognizability corruption (Gaussian blur, reduced resolution, and linear motion blur) to images in the IJB-C face verification benchmark and show the change of error rates in Fig. 3. We observe a clear trend of error increase as more corruption is added. This shows the risk of not considering the recognizability of the images in recognition and calls for a mitigation solution. Luckily, the fact that *the UI is distant from recognizable identities in the embedding space*, also suggests a way of *measuring recognizability by measuring their distances to the UI* and factor recognizability into face recognition prediction, as we detail in Sec. 2.2.1.

2.2. Accounting for Recognizability

Based on the hypothesis that distance to the UI cluster centroid could serve as a measure of the face recognizability, we use the distance as an *embedding recognizability score* (ERS), which requires no additional training nor annotation and can be easily incorporated into both single-image and set-based face recognition.

2.2.1 The Embedding Recognizability Score

We define Embedding Recognizability Score (ERS) to be the distance between an embedding vector and the average embedding of UI images. UI images can be obtained either by artificially degrading the recognizability of regular face images such as those from face embedding model training dataset [‡], or clustering in-the-wild face detection datasets consisting naturally unrecognizable faces such as WIDERFace [48] and taking the resultant UI cluster (see details in Sec. 3.3). The normalized average feature of the UI images, $\mathbf{f}_{UI}$, which we call the UI centroid (UIC), is used to represent the UI. And ERS e_i of an embedding $\mathbf{f}_i$, is given by

$$e_i = 1 - \langle \mathbf{f}_{UI}, \mathbf{f}_i \rangle. \tag{1}$$

We illustrate the correlation between ERS and recognizability in Fig. 4 by taking images from the IJB-C dataset and grouping them by their ERS scores. The ERS decrease is accompanied by face quality variations such as occlusion distortion, larger poses, and increased image blurriness.

2.2.2 ERS in Single Image-Based Recognition

Face verification aims to determine whether two face images, x_1 and x_2, belong to the same person. They have corresponding groundtruth identity labels y_1 and y_2. The face embedding model represents an input face image x_i as the feature vector $\mathbf{f}_i \in \mathcal{R}^d$. Without considering recognizability, the estimated probability of two images being from the same person is usually

$$p(\hat{y}_1 = \hat{y}_2 | x_1, x_2) = s(\mathbf{f}_1, \mathbf{f}_2). \tag{2}$$

Here $s(\cdot, \cdot)$ is the cosine similarity function, and $\hat{y}_i$ is the estimated identity label of x_i, which in practice is not explicitly computed. Then a binary prediction is made based

[‡]The recognizability in our context is observer-dependent, where the observer could be a human or an algorithm. Although it could be debated that a face unrecognizable to the human eye may still be recognizable to an algorithm when we degrade the face images, we empirically found adding extreme noise as per manual check to work across the benchmarks

on whether the probability is higher than an empirically set threshold τ.

To take ERS into account, we allow the system to predict "unsure" instead of "same" or "different" when either e_1 or e_2 is below a threshold γ. When the ERS of x_1 or x_2 is low, we observe the recognizability to be low. The similarity between them cannot be reliably determined to predict if they belong to the same id. In applications, one can choose further actions on the unsure cases. For example, in our experiments on the IJB-C Covariate Test, we choose to predict all unsure cases as not belonging to the same person due to the empirical risk of false matches.

Face identification aims to tell whether one query image x_i, represented as $\mathbf{f}_i$ belongs to one of N indexed identities in the gallery, represented as $\{\mathbf{g}_j\}_{j=1}^{N}$. [‡] Here we assume that the gallery has mostly recognizable images. Without considering recognizability, the decision function $S(\mathcal{R}^d) \rightarrow [0, \ldots, N]$ is

$$S(\mathbf{f}_i; \{\mathbf{g}_j\}) = \mathbb{1}[\max_j s(\mathbf{f}_i, \mathbf{g}_j) \geq \tau] \cdot \arg\max_{j=1,\ldots,N} s(\mathbf{f}_i, \mathbf{g}_j) \quad (3)$$

where $\max_j s(\mathbf{f}_i, \mathbf{g}_j)$ indicates the maximal similarity score or the search top retrieval. The query has a positive match when the maximal similarity score to any of the gallery images is above the threshold τ. With ERS, the decision function becomes

$$S'(\mathbf{f}_i; \{\mathbf{g}_j\}) = S(\mathbf{f}_i; \{\mathbf{g}_j\}) \mathbb{1}(e_i \geq \gamma). \quad (4)$$

Again, when e_i is lower than γ, it becomes an unsure case, we predict there is no positive match. If there is no guarantee that the gallery images are mostly recognizable, ERS could be applied to reject retrieved gallery images below γ.

2.2.3 ERS in Image Set-based Face Recognition

In set-based face recognition, we have prior knowledge that each set or template [28], contains one or multiple face images belonging to a single person. Set-based face recognition also usually consists of face verification and face identification. We first extract feature vectors using the embedding model for every image in each set θ_i containing images $\{x_i^l\}$ as $\{\mathbf{f}_i^l\}_{l=1}^{|\theta_i|}$, where $|\theta_i|$ is the cardinality of θ_i. Then the feature vectors are aggregated into one feature vector $\mathbf{f}_i$. After aggregation, the processing is the same as in the single image case. We design an aggregation function weighted by the ERS of each image as

$$\mathbf{f}_i = \sum_{l=1}^{|\theta_i|} \frac{w(e_i^l)\mathbf{f}_i^l}{\sum_l e_i^l}, \quad e_i = \frac{\sum_l e_i^l}{|\theta_i|}, \quad (5)$$

[‡]In this work we only deal with the "open-set" setting [28] where predicting a query has no match in the gallery is allowed, as it is the most common usage in real-world FR systems.

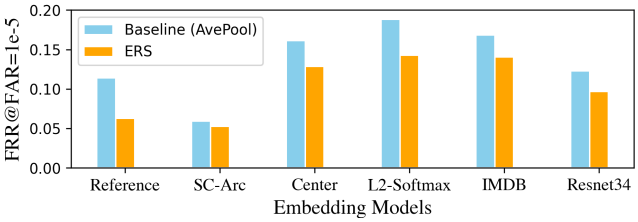

Figure 5. Pairwise performance comparison between baseline (average pooling) and our method (ERS) on IJB-C template across different settings, our method achieves consistent error reduction. Training settings: Reference (R101, CosFace, DeepGlint), SC-Arc(R101, Sub-center Arcface, DeepGlint), Center (R101, Softmax+Center, DeepGlint), L2-Softmax (R101, ℓ_2-Softmax, DeepGlint), IMDB (R101, CosFace, IMDB), ResNet34 (R34, CosFace Loss DeepGlint).

where $\mathbf{f}_i$ and e_i denote the aggregated feature vector and ERS for the set θ_i, $w : \mathcal{R} \rightarrow \mathcal{R}^+$ denotes the weighting function based on ERS. The aggregated feature vectors can then be used as in the single image cases. We discuss choices of weighting function in the ablation study.

3. Experiments

We examine the effectiveness of ERS in face recognition on multiple benchmarks. We consider two face image quality assessment methods as the baselines: FaceQnet [16] and SER-FIQ [36]. In set-based face recognition, we compare the ERS based aggregation function with other set-based recognition methods: NAN [47], Multicolumn [45], DCN [44], Iconicity [12], PFE [34] and DUL [2]. Due to the limitation of space, we share our preliminary bias analysis of the ERS in the Supplemental Material .

Implementation details We use the deep learning framework MXNet [3] in our model training and evaluation. We train a face embedding model using CosFace [38] loss, ResNet-101 (R101) [14] backbone and DeepGlint-Face dataset (including MS1M-DeepGlint and Asian-DeepGlint) [7]. HAC algorithm [6] is used to cluster extracted embeddings and generate UI clusters. We select threshold $\gamma = 0.60$ for ERS via cross-validation on the TinyFace [42] benchmark.

Evaluation data Single image and set-based face recognition experiments are run on the IARPA Janus Benchmark-C (IJB-C) [28] [‡]. We run evaluations on the IARPA Janus Benchmark-C (IJB-C) [28]. IJB-C dataset suite contains in-the-wild celebrity media, including photos and videos, and multiple predefined benchmark protocols. We employ four protocols from IJB-C: 1) *IJB-C Covariate Face Verifica-*

[‡]This paper contains or makes use of the following data made available by the Intelligence Advanced Research Projects Activity (IARPA): Benchmark C (IJB-C) data detailed at Face Challenges homepage. For more information see https://nigos.nist.gov/datasets/ijbc/request.

Mehod	Verification FRR@FAR			
	1e-6	1e-5	1e-4	1e-3
Baseline	0.6976	0.4540	0.1747	0.0714
FaceQnet [16] ($\gamma = 1.00$)	0.6976	0.4540	0.1747	0.0714
SER-FIQ [36] ($\gamma = 0.83$)	0.4027	0.2023	0.1164	0.0717
ERS ($\gamma = 0.60$)	**0.3819**	**0.1885**	**0.1113**	**0.0673**

Table 1. Comparison of recognizability conditioned face verification on IJBC Covariate Verification benchmark. We compare with FaceQnet [16] and SER-FIQ [36] as the alternatives for face recognizability measures.

tion Test for single image-based face verification, 2) *IJB-C Template-based Face Verification* for set-based face verification, 3) *IJB-C Template-based Face Search* for set-based face search, and 4) *IJB-C Test10: Wild Probe with Full Motion Video Face Search* for set-based video face search. It is worth noting that the IJB-C Test10 contains video frames with strong motion blur, forming an FR testbed with unrecognizable video faces.

Evaluation Metric Methods performance is measured on two standard face recognition test cases: 1:1 verification and 1:N search. For face verification, we measure False Reject Rates (FRRs, also known as $1-$ True Acceptance Rate), at a set of False Acceptance Rates (FARs). For face search, we measure False Negative Identification Rates (FNIRs, or $1-$ True Positive Identification Rate) at different False Positive Identification Rates (FPIR), as well as top-K accuracy.

3.1. ERS in Single Image-Based Face Verification

The IJB-C Covariate Test protocol benchmarks single image-based face verification. In Table 1, we evaluate our approach on this benchmark, and compare with FaceQnet [16] and SER-FIQ [36] as the alternatives for recognizability measurement. With the same embedding model, using ERS as the recognizability measure leads to 58% error reduction at FAR=1e-5 compared to that of not considering recognizability. The alternative recognizability measure FaceQnet cannot improve the baseline at all thresholds, and SER-FIQ can help reduce errors, but with less effect.

3.2. ERS in Set-Based Face Recognition

We evaluate set-based face recognition using ERS as described in Sec. 2.2.3. Results on the IJB-C Template-Based Face Verification and Search are illustrated in Table 2 (top). We adapt media-pooling [5] in set-based benchmarks. We compare ERS based method to the baseline of simple averaging without considering recognizability (AvePool). We also compare our approach with other methods developed for set-based face recognition. It can be seen that our approach significantly reduces recognition errors of the baseline. Our results are comparable or better than other complex methods developed for set-based recognition.

We also evaluate ERS on the IJB-C Test10 benchmark to test our algorithms on videos in the wild and implement an average pooling baseline and learning-based method NAN [47] for comparison. The results are summarized in Table 3. It can be seen that although obtained from images, ERS is also able to improve recognition accuracy in challenging video data.

Method	Backbone	Train Data	IJB-C Veri FRR@FAR		
			1e-5	1e-4	1e-3
Multi. [45]	R50	VGGFace2	0.2290	0.1380	0.0730
DCN [44]	R50	VGGFace2	-	0.1150	0.0530
Icon. [12]	CNN [43] L2 [31]	MS1M+UMD	0.1270	0.0770	0.0470
PFE [34]	64-Layer CNN	MS1M	0.1036	0.0675	0.0451
NAN [47]	R64 Cosface	DeepGlint	0.1023	0.0582	0.0347
DUL [2]	R64	MS1M	0.0977	**0.0539**	0.0530
AvePool	R64 Cosface	DeepGlint	0.1262	0.0639	0.0332
ERS			**0.0929**	0.0547	**0.0312**
AvePool	R101	DeepGlint	0.0592	0.0406	0.0271
ERS	SubCenterArc		**0.0528**	**0.0363**	**0.0245**

Table 2. Benchmark results on IJB-C Template-Based Verification. Top: we compare using ERS for aggregation with the baseline of averaging embedding (AvePool) and other face aggregation methods. The results are reported at different FAR levels. We report error rates to better illustrate the difference between methods. Bottom: ERS application on the latest Sub-center ArcFace model.

Improving State-of-the-art Face Embedding Models. In Fig. 1 we illustrate the UI cluster using the Cosface [38] for embedding. The UI clustering phenomenon is not limited to that embedding model. We empirically find that multiple other face embedding models have similar behaviors, despite their different loss functions, size of training datasets, and backbone architectures. We conjecture that the existence of UI clusters may be attributed to the nature of face recognition as a fine-grained categorization task. Since ERS is easy to obtain for any face embedding model without extra training or annotation, we test applying ERS to multiple state-of-the-art face embedding models trained under different settings, including (1) loss function designs: Sub-center ArcFace [9], Softmax+Center Loss [43] and ℓ_2-Softmax Loss [31] ; (2) training dataset: IMDB [37] and DeepGlintFace; (3) backbone architectures: ResNet-101, ResNet-50 and ResNet-34 [14].

Results on the IJB-C face verification benchmark are illustrated in Fig. 5. We observe consistent error reduction on all tested models, including 10% error reduction at FAR=1e-5 on a strong baseline from Sub-center ArcFace [9]. Full results can be found in Table 2 (bottom). This suggests that the ERS can be an easy plug-in to existing face recognition systems to reduce recognition errors.

3.3. Ablation Study

Generation of the UI Centroid. The UI centroid (UIC) is a key component of our ERS-based methods. We explore two approaches for obtaining the UI images. The first is a direct approach. As introduced in Sec. 2.1, we heavily perturbed the recognizability of 1K randomly sampled faces from the training dataset DeepGlint, using techniques exemplified by

Method	Identification FNIR@FPIR				Rank-N Error	
	1e-4	1e-3	1e-2	1e-1	1	5
AvePool	0.8607	0.6840	0.4129	0.2029	0.1564	0.1137
NAN [47]	0.8566	0.6697	0.3726	0.1956	0.1518	0.1100
ERS	**0.8299**	**0.6096**	**0.3054**	**0.1807**	**0.1457**	**0.1055**

Table 3. IJB-C Test 10: Wild Probe with Full Motion Video Face Search results, tested using backbone ResNet101 trained on DeepGlintFace with CosFace [38] loss. In comparison with baseline average pooling and NAN [47] trained on top of the same backbone model, our ERS-based aggregation achieves the best performance.

Clustering dataset	DeepGlint 1K	WIDERFace	FDDB
UI cluster average distance	0.3907	0.3213	0.4344
UIC distance to FDDB	0.0560	0.0526	0.0000
w/ ERS FRR@FAR=1e-5	0.0623	0.0627	0.0625

Table 4. Comparison between UI generated from different datasets and ERS results on IJB-C Template-based Face Verification. Without ERS the FRR@FAR=1e-5 is 0.1140. PDS indicates a perturbed DeepGlint subset.

Fig. 3. In the second approach, we automatically select UI images from face detection datasets, such WIDERFace and FDDB [18] datasets, where low recognizability faces are naturally present. After face clustering, we found the largest resultant clusters to be UI clusters for both datasets, and this is consistent across all examined models (Sec.3.2). We compare UICs generated from DeepGlint 1K, WIDERFace and FDDB [18] in table 4, where we list the average cosine distances of each UI cluster, a comparison between generated UICs, and associated aggregation results. It can be seen that our method is not sensitive to the UI image source, artificial or natural. This is also consistent with our observation that heterogeneous UI images gather in one cluster. We conclude that it is possible to obtain UIC from different data distributions where there are low recognizability images.

When clustering across different embedding models, we find resultant images in the UI clusters to have significant overlap. Further experiments prove a fixed set of UI images generated from one embedding model can be reused by other models for obtaining UIC. In fact, we obtained a UI set from a Res101-based Cosface model, and reused it to generate UIC for all other models in Fig. 5 and Table 3. Either using the direct or clustering-based approach, the application of our method is simple and efficient.

Choice of the Weighting Function w. Using ERS in set-based face recognition requires a choice of the weighting function w. We compare different choices of w, including identity, exponential, and square on the IJB-C Template-Based Face Verification benchmark. We also compare with two special choices that average the images with top-1% and top-10% ERSs within a set. From table 5 we can see square function achieves the best results. We use it in other experiments without special notes.

w	e_i	$softmax(e_i)$	e_i^2	e_i	e_i
Set Selection	N/A	N/A	N/A	top 1	top 10%
FRR@FAR 1e-5	0.0684	0.1393	**0.0627**	0.2196	0.2025

Table 5. Comparison between different ERS-based aggregation methods on IJBC Templated-based Face Verification benchmark, basenet ResNet101 trained with Cosface [38].

	FRR@FAR			
	1e-5	1e-4	1e-3	1e-2
Baseline AvePool	0.1140	0.0468	0.0264	0.0141
ERS (WeightedPool)	**0.0627**	**0.0393**	**0.0243**	**0.0140**
ERS-Enhanced AvePool	0.0644	0.0410	0.0255	0.0144

Table 6. Comparison among average pooling, ERS weighted pooling and average pooling with ERS enhanced features on IJB-C Template-based Face Verification benchmark, basenet is R101 trained with Cosface loss.

3.4. Exploration on Increasing ERS

Higher ERSs usually associate with faces with better recognizability. It is interesting to explore whether we can increase the ERS of a given image. We present the results of a naive method for this: enhance the face feature by removing its projection on the direction of the UI representation to yield high ERS ($e = 1$) features. Formally, we take the raw feature embedding $\mathbf{f}$, unit vector $\mathbf{f}_{UI}$ and calculate the feature

$$\mathbf{v}^{id} = \mathbf{f} - \langle \mathbf{f}, \mathbf{f}_{UI} \rangle \mathbf{f}_{UI}. \tag{6}$$

After ℓ_2 normalization we get the ERS-enhanced features $\mathbf{f}^{id}$. From table 6 we can see average pooling with the ERS-enhanced features surpasses the baseline by a large margin and achieves comparable results to ERS weighted aggregation. This suggests increasing embedding ERS can make a meaningful difference in benchmark results, and achieving higher ERS through more advanced techniques may lead to a further increase in recognition accuracy.

3.5. Emergence of the UI Cluster

Inspired by recent works in out-of-distribution detection [17], we provide an exploratory study on the evolution of UI image embedding during training. We analyze the pair-wise cosine distance of embedding vectors within and between three sets of images: (a) UI images from known UI clusters, (b) recognizable faces from the same identities, (c) recognizable faces from different identities. During training, we compute UIC on set (a) after each epoch. We measure the distance of these intermediate UICs to the UIC obtained after the final epoch. We average results from 10 independent training runs and show them in Fig. 6. We can see that as the training number of epochs increases: (1) The UI cluster emerges early on during training. Its centroid shifts significantly during model training. (2) The distance between UI images and recognizable ids stays high, similar to that among faces from the different ids (we call these the

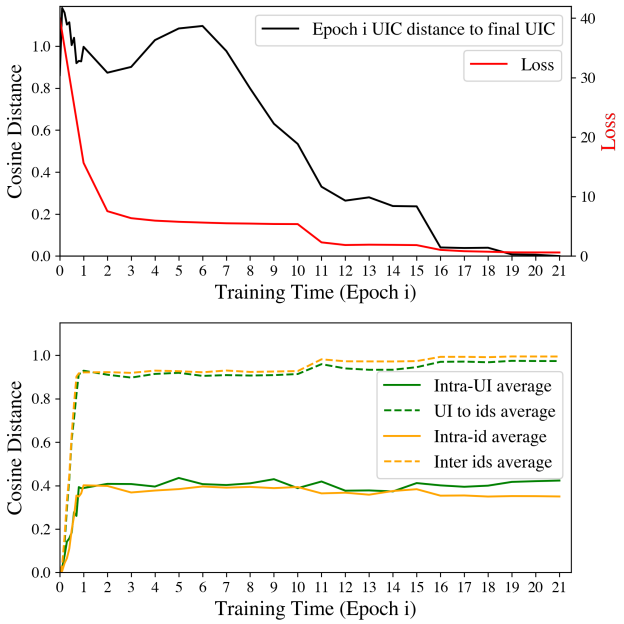

Figure 6. Top: Trajectory of the UI centroid during training w.r.t. the final model UIC. Bottom: Distances comparison between UI images, recognizable faces from the same identity and recognizable faces from different identities.

Figure 7. Images from Market1501 grouped by ERSs. A positive correlation between recognizability and ERS can be observed.

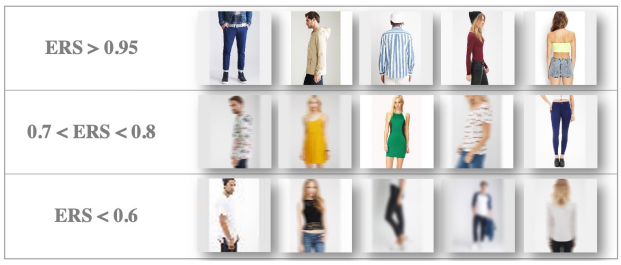

Figure 8. Images from Deepfashion with synthetic corruption grouped ERSs. A positive correlation between recognizability and ERS can be observed.

High Distance Groups). (3) Distances among UI images remain at low values, similar to that of faces from the same id (Low Distance Groups) (4) Between the High and Low Distance Groups, a clear gap could be observed.

3.6. Exploration on Other Vision Tasks

As an extended study, we explore clustering on person re-identification (re-id) and image retrieval to see whether the low recognizability clustering phenomenon exists and whether our method can be applied accordingly. Similar to Fig. 3, we perform re-id embedding clustering on Market1501 [50] dataset, and likewise for partially perturbed Deepfashion In-Shop dataset [24]. We observe low recognizability of miscellaneous samples gather in one cluster similar to those of faces. After devising the associated ERS measures, it can be observed from Fig. 7 and Fig. 8 that consistent with our findings on the face, the ERS also correlates with the input image recognizability. We show more results in the Supplemental Material due to the limitation of space.

4. Conclusion and Discussion

Face recognition is subject to a vast array of issues that can affect the quality of its result. While we acknowledge the presence and importance of statistical and algorithmic biases, in this paper we focus on additional issues that exist even before any dataset is collected, and affect systems that are not trained on data. Simply put, some images contain sufficient information to ascertain that there *is* a face, but insufficient to determine *whose* face it is. This gap is captured by the concept of "recognizability", which is affected by the physical properties of the subject (sub-surface scattering and pigmentation, occlusions from hairdo, accessories, makeup), but also extrinsic properties of the scene such as the nature and quality of the illuminant, physical properties of the sensor and pre-processing algorithms performed by the camera software, imaging conditions such as large aperture/high capture time resulting in motion blur, finite depth of field and resulting optical blur, etc.

We also acknowledge that the principled solution to both cascading failures of the detection and recognizability of the detected face, is to properly marginalize the corresponding latent variables. Since that is highly impractical, we settle for the intermediate inference of recognizability, through the proposal of an admittedly ad-hoc measure, suggested by the manifest clustering of unrecognizable identities. Because of the issues mentioned above, and also verified empirically, the UI is a highly heterogeneous cluster. It includes images subject to wildly varying nuisances, unlike other domains such as large-scale image classification where optical-blurred images form a cluster separate from motion-blurred images or low-resolution images.

With all due caveats in mind, we observe that explicitly accounting for recognizability through the admittedly unprincipled method we have proposed, we still achieve significant error reduction in face recognition on standard public benchmarks and effectively allow a system to operate in an open-set setting without the complications of full-fledged open universe training.

References

[1] Guha Balakrishnan, Yuanjun Xiong, Wei Xia, and Pietro Perona. Towards causal benchmarking of biasin face analysis algorithms. In *Deep Learning-Based Face Analytics*, pages 327–359. Springer, 2021.

[2] Jie Chang, Zhonghao Lan, Changmao Cheng, and Yichen Wei. Data uncertainty learning in face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5710–5719, 2020.

[3] Tianqi Chen, Mu Li, Yutian Li, Min Lin, Naiyan Wang, Minjie Wang, Tianjun Xiao, Bing Xu, Chiyuan Zhang, and Zheng Zhang. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. *arXiv preprint arXiv:1512.01274*, 2015.

[4] Alexandra Chouldechova, Siqi Deng, Yongxin Wang, Wei Xia, and Pietro Perona. Unsupervised and semi-supervised bias benchmarking in face recognition. 2022.

[5] Nate Crosswhite, Jeffrey Byrne, Chris Stauffer, Omkar Parkhi, Qiong Cao, and Andrew Zisserman. Template adaptation for face verification and identification. *Image and Vision Computing*, 79:35–48, 2018.

[6] William HE Day and Herbert Edelsbrunner. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of classification*, 1(1):7–24, 1984.

[7] Deepglint. http://trillionpairs.deepglint.com/overview.

[8] Jiankang Deng, Jia Guo, Tongliang Liu, Mingming Gong, and Stefanos Zafeiriou. Sub-center arcface: Boosting face recognition by large-scale noisy web faces. In *Proceedings of the IEEE Conference on European Conference on Computer Vision*, 2020.

[9] Jiankang Deng, Jia Guo, Tongliang Liu, Mingming Gong, and Stefanos Zafeiriou. Sub-center arcface: Boosting face recognition by large-scale noisy web faces. In *Proceedings of the IEEE Conference on European Conference on Computer Vision*, 2020.

[10] Jiankang Deng, Jia Guo, Xue Niannan, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019.

[11] Jiankang Deng, Jia Guo, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4685–4694, 2018.

[12] Prithviraj Dhar, Carlos Castillo, and Rama Chellappa. On measuring the iconicity of a face. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2137–2145. IEEE, 2019.

[13] Sixue Gong, Yichun Shi, and Anil K Jain. Video face recognition: Component-wise feature aggregation network (c-fan). *arXiv preprint arXiv:1902.07327*, 2019.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[15] Javier Hernandez-Ortega, Javier Galbally, Julian Fierrez, and Laurent Beslay. Biometric quality: Review and application to face recognition with faceqnet. *arXiv preprint arXiv:2006.03298*, 2020.

[16] Javier Hernandez-Ortega, Javier Galbally, Julian Fierrez, Rudolf Haraksim, and Laurent Beslay. Faceqnet: Quality assessment for face recognition based on deep learning. In *2019 International Conference on Biometrics (ICB)*, pages 1–8. IEEE, 2019.

[17] Haiwen Huang, Zhihan Li, Lulu Wang, Sishuo Chen, Bin Dong, and Xinyu Zhou. Feature space singularity for out-of-distribution detection. *arXiv preprint arXiv:2011.14654*, 2020.

[18] Vidit Jain and Erik Learned-Miller. Fddb: A benchmark for face detection in unconstrained settings. Technical report, UMass Amherst technical report, 2010.

[19] Muwei Jian and Kin-Man Lam. Simultaneous hallucination and recognition of low-resolution faces based on singular value decomposition. *IEEE Transactions on Circuits and Systems for Video Technology*, 25:1761–1772, 2015.

[20] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6738–6746, 2017.

[21] Xiaofeng Liu, Zhenhua Guo, Site Li, Lingsheng Kong, Ping Jia, Jane You, and BVK Kumar. Permutation-invariant feature restructuring for correlation-aware image set-based recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4986–4996, 2019.

[22] Xiaofeng Liu, BVK Kumar, Chao Yang, Qingming Tang, and Jane You. Dependency-aware attention control for unconstrained face recognition with image sets. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 548–565, 2018.

[23] Yu Liu, Junjie Yan, and Wanli Ouyang. Quality aware network for set to set recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5790–5799, 2017.

[24] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[25] Boyu Lu, Jun-Cheng Chen, and Rama Chellappa. Unsupervised domain-specific deblurring via disentangled representations. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10217–10226, 2019.

[26] Ze Lu, Xudong Jiang, and Alex Chichung Kot. Deep coupled resnet for low-resolution face recognition. *IEEE Signal Processing Letters*, 25:526–530, 2018.

[27] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[28] Brianna Maze, Jocelyn Adams, James A Duncan, Nathan Kalka, Tim Miller, Charles Otto, Anil K Jain, W Tyler Niggel, Janet Anderson, Jordan Cheney, et al. Iarpa janus benchmark-c: Face dataset and protocol. In *2018 International Conference on Biometrics (ICB)*, pages 158–165. IEEE, 2018.

[29] Alice J O'Toole, Carlos D Castillo, Connor J Parde, Matthew Q Hill, and Rama Chellappa. Face space representations in deep convolutional neural networks. *Trends in cognitive sciences*, 22(9):794–809, 2018.

[30] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *BMVC*, 2015.

[31] Rajeev Ranjan, Carlos D Castillo, and Rama Chellappa. L2-constrained softmax loss for discriminative face verification. *arXiv preprint arXiv:1703.09507*, 2017.

[32] Yongming Rao, Ji Lin, Jiwen Lu, and Jie Zhou. Learning discriminative aggregation network for video-based face recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 3781–3790, 2017.

[33] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015.

[34] Yichun Shi, Anil K Jain, and Nathan D Kalka. Probabilistic face embeddings. *arXiv preprint arXiv:1904.09658*, 2019.

[35] Yichun Shi, Xiang Yu, Kihyuk Sohn, Manmohan Chandraker, and Anil K Jain. Towards universal representation learning for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6817–6826, 2020.

[36] Philipp Terhorst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Ser-fiq: Unsupervised estimation of face image quality based on stochastic embedding robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5651–5660, 2020.

[37] Fei Wang, Liren Chen, Cheng Li, Shiyao Huang, Yanjie Chen, Chen Qian, and Chen Change Loy. The devil of face recognition is in the noise. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 765–780, 2018.

[38] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5265–5274, 2018.

[39] Mei Wang and Weihong Deng. Deep face recognition: A survey. *ArXiv*, abs/1804.06655, 2018.

[40] Mei Wang and Weihong Deng. Mitigating bias in face recognition using skewness-aware reinforcement learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9319–9328, 2020.

[41] Zhangyang Wang, Shiyu Chang, Yingzhen Yang, Ding Liu, and Thomas S. Huang. Studying very low resolution recognition using deep networks. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4792–4800, 2016.

[42] Zhifei Wang, Zhenjiang Miao, QM Jonathan Wu, Yanli Wan, and Zhen Tang. Low-resolution face recognition: a review. *The Visual Computer*, 30(4):359–386, 2014.

[43] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pages 499–515. Springer, 2016.

[44] Weidi Xie, Li Shen, and Andrew Zisserman. Comparator networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 782–797, 2018.

[45] Weidi Xie and Andrew Zisserman. Multicolumn networks for face recognition. *arXiv preprint arXiv:1807.09192*, 2018.

[46] Xingkun Xu, Yuge Huang, Pengcheng Shen, Shaoxin Li, Jilin Li, Feiyue Huang, Yong Li, and Zhen Cui. Consistent instance false positive improves fairness in face recognition. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 578–586, 2021.

[47] Jiaolong Yang, Peiran Ren, Dongqing Zhang, Dong Chen, Fang Wen, Hongdong Li, and Gang Hua. Neural aggregation network for video face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4362–4371, 2017.

[48] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[49] Jian Zhao, Yu Cheng, Yan Xu, Lin Xiong, Jianshu Li, Fang Zhao, Karlekar Jayashree, Sugiri Pranata, Shengmei Shen, Junliang Xing, et al. Towards pose invariant face recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2207–2216, 2018.

[50] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015.