Recognizing Tiny Faces

Siva Chaitanya Mynepalli
Carnegie Mellon University
smynepal@cs.cmu.edu

Peiyun Hu
Carnegie Mellon University
peiyunh@cs.cmu.edu

Deva Ramanan
Carnegie Mellon University
deva@cs.cmu.edu

Abstract

Objects are naturally captured over a continuous range of distances, causing dramatic changes in appearance, especially at low resolutions. Recognizing such small objects at range is an open challenge in object recognition. In this paper, we explore solutions to this problem by tackling the fine-grained task of face recognition. State-of-the-art embeddings aim to be scale-invariant *by extracting representations in a canonical coordinate frame (by resizing a face window to a resolution of say, 224x224 pixels). However, it is well known in the psychophysics literature that human vision is decidedly scale* variant*: humans are much less accurate at lower resolutions. Motivated by this, we explore* scale-variant *multiresolution embeddings that explicitly disentangle factors of variation across resolution and scale. Importantly, multiresolution embeddings can adapt in size and complexity to the resolution of input image* on-the-fly *(e.g., high resolution input images produce more detailed representations that result in better recognition performance). Compared to state-of-the-art "one-size-fits-all" approaches, our embeddings dramatically reduce error for small faces by at least* **70%** *on standard benchmarks (i.e. IJBC, LFW and MegaFace).*

1. Introduction

Objects are visually captured at a continuous range of distances in the real world. One of the remaining open challenges in object recognition is recognition of small objects at range [19]. We focus on the illustrative task of recognizing faces across a wide range of scales, a crucial task in surveillance [6]. This is a well-known challenge because distinctive features (such as eyebrows [27]) may not be resolvable in low resolution. Contemporary face recognition systems, which now outperform the average forensic examiner on high quality images [15], perform dramatically worse for lower resolutions (Fig. 2 and 3).

Scale: Recognition is often cast as an image retrieval task, where the central challenge is learning an embedding for matching image queries (probes) to a stored library (gallery).

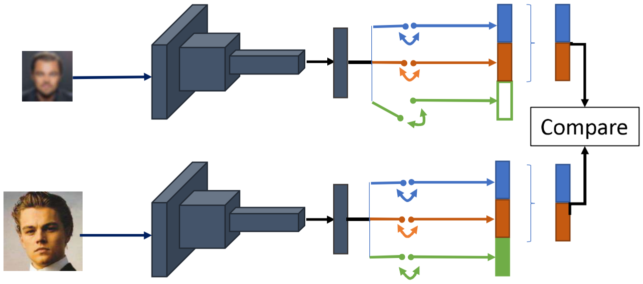

Figure 1: Traditional approaches for matching compare embedding vectors of a query and reference image. We introduce multi-resolution embeddings with several desirable properties (1) they adapt in complexity to the resolution of the input, such that larger embeddings are produced when additional high-res information is available (**bottom**). (2) they produce disentangled representations where frequency-specific components can be "switched off" when not present in the input (**top**). (3) they can adapted *on-the-fly* to any desired resolution by "‘zero’ing out" certain frequencies (the **bottom-right** embedding).

Virtually all contemporary retrieval systems learn a scale-*in*variant embedding, by first canonicalizing a given image crop to a standard resolution (of say, 224x224 pixels) before feature extraction [17]. However, recognition accuracy for human vision is decidedly scale *variant*. Humans are much more accurate at higher resolutions, and moreover, tend to rely on resolution-specific features to make inferences at particular resolutions [30]. Fig. 2 shows a reference image and candidate probe matches at varying resolutions. At low resolutions, coarse features such as the hairline and jaw shape seem to reveal the identity. At high resolutions, subtle features such as the eyebrow and nose shape appear to play an important role. Such resolution-specific identity cues cannot be captured by a scale-invariant embedding.

Mulitresolution embeddings: We begin by showing that a conceptually simple solution is to train *multiple* fixed-resolution embeddings, and use the appropriate one depending on the resolution of the query (probe) and reference (gallery) face to be compared. Moreover, one can significantly improve accuracy by combining these resolution-specific embeddings into a *single* multiresolution representation that explicitly disentangles factors of identity into frequency-specific components. For example, certain dimensions of the embedding vector are trained to encode low-frequency cues such as hairlines, while other dimensions are trained to encode high-frequency cues such as nose shape. In the limit, one can interpret our embeddings as a "fourier" decomposition of identity into frequency-specific components. Importantly, because the resolution of an input image is known, missing frequencies for low-res inputs can be "switched off". Moreever, even when present in high-res input, they can be "zero'd out" on-the-fly to facilitate comparisons to low-res images (Fig. 1).

Disentangled representations: We illustrate two applications that specifically exploit disentangled embeddings. The first is *adapation*: given a probe at a particular resolution, we adapt the gallery embedding *on-the-fly* by selecting the appropriate frequency-specific components in the embedding (Fig. 1).The second is *aggregation*: practical face recognition methods often match *sets* of faces (say, extracted from a video sequence). Such methods typically produce an aggregate template representation by pooling embeddings from faces in the set [26, 7]. We show that multiresolution pooling, that uses only high-resolution faces to produce the high-frequency components in the final embedding, is considerably more accurate.

Evaluation: Evaluating our method is hard because most benchmarks provide faces only at high-resolution. This reveals the inherent bias of the community for scale invariance! It is tempting to create artificial scale variation by resizing such images [16]. In fact, we do so for diagnostic experiments, resizing the well-known LFW datset [13] into different resolutions. However, recent work has shown that downsampling is not a good model for natural scale degradation [5]. As such, we present final results on the IJBC [23] benchmark, which is unique in that it includes the raw images on which faces were extracted, and so contains natural scale variation. Our results show that multiresolution embeddings can naturally cope with the various factors that influence real low resolution faces like jpeg artefacts, motion blur etc. by adapting the embedding *on-the-fly* to a lower resolution.

We also compare our algorithm on *resized* versions of the popular Megaface dataset to showcase our algorithm on a larger scale. Additionally, we compare the performance of our approach with more recent face recognition networks in the supplement.

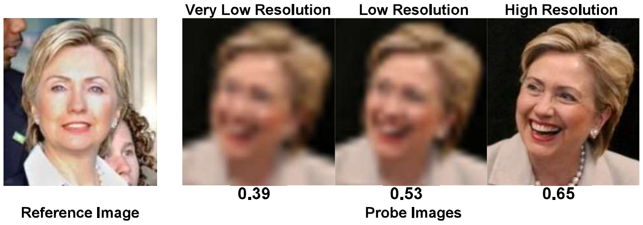

Figure 2: We illustrate the drop in recognition performance with resolution. The numbers at the bottom of each probe image is the similarity score obtained by comparing a probe of specified resolution with the reference image using a state-of-the-art face recognition model [7]. However, humans can make accurate inferences on these pairs of images by comparing resolution-specific features. For example, we rely on hairstyle, face shape etc. to accurately compare the very low resolution probe image with the reference image, and on finer details like eyebrows when verifying high res images.

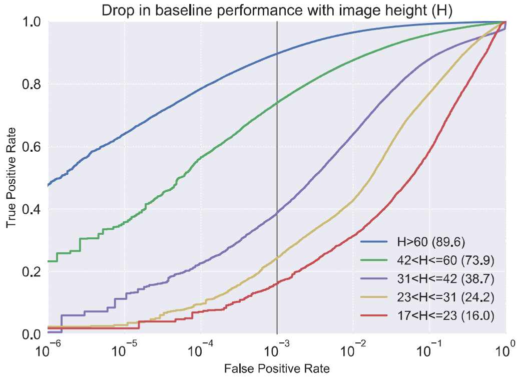

Figure 3: To explore how resolution affects recognition performance, we evaluate a state-of-the-art face embedding (VGGFace2 [7]) on resolution-constrained subsets of a standard face recognition dataset IJBC[23]. Note the significant drop in performance as resolution decreases (i.e. 20 pixels). At a false-positive rate of 10^{-3}, the true positive rate for small (20 pixel) faces drops by 60%.

2. Related work

CNN based face recognition: Recent methods for face recognition aim to learn an nonlinear embedding through a variety of loss functions, including triplet loss [28], softmax loss [24], and angular softmax loss [22]. We use the well-known VGG face network [7] as our backbone for fine-tuning. Instead of learning an "one-size-fits-all" embedding, we learn a multiresolution representation that can be adapted to different resolutions. Our approach to scale-invariance is inspired by previous work on pose invariance [14], which learns separate models for frontal and profile faces.

Human vision: Extensive studies on human vision show that human are surprisingly good at recognizing low-res faces [30]. [9] shows that human accurately recognize familiar faces even as small as 16x16. [6] points out the familiarity is the key – the more human are familiar with the face subject the more they can tolerate the poor quality of imagery. Perhaps the closest analogy to familiarity is learning-based recognition methods. Contemporary face recognition approaches train face embeddings on millions of images for many iterations. In some sense, given any new face image,

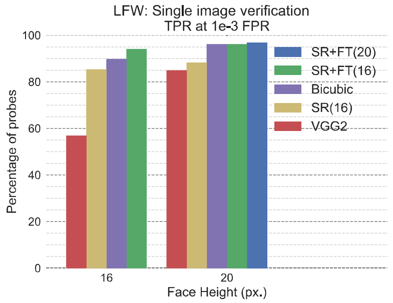

Figure 4: **Impact of resolution- specific models** We demonstrate the massive improvement in the performance of our resolution-specific models compared to the baseline **VGG2** embedding (trained for 224x224) on the task of low-res face verification. On the left, we test our resolution-specific model tuned for images of height 16 (LFW-16), **SR+FT(16)**. On the right, we test a resolution-specific model tuned for images of height 20 (LFW-20), **SR+FT(20)**. We show that super-resolving the low res image back to 224x224 (SR+FT) performs better than basic bicubic upsampling (**Bicubic**), and **VGG2**. We also show that **SR+FT(20)** performs better than **SR+FT(16)** on LFW-20. It shows that we need to train resolution-specific models at multiple resolutions for best performance. Full plots shown in supp. material.

it must have seen faces that feel familiar.

Multi scale representations in neural networks: Using representations drawn from multiple scales has been integral to computer vision tasks ever since the seminal work on gaussian pyramids [1]. More recently, researchers have been using deep representations drawn from multiple scales to include greater context for Semantic Segmentation [34], Object Detection [18] and other vision tasks. Our work is inspired by such approaches, but differs in its execution because the *dimensionality* of our underlying embedding depends on the image resolution.

Low resolution face recognition: Recent works on low-resolution face recognition can be classified into two categories [32]. The first category can be referred to as super-resolution based [2, 3, 21, 20, 11, 12, 35, 33, 16] approaches. Given a low-res probe, these methods first hallucinate the high-res version, and then verify/classify the high-res version. Alternatively, one might learn a feature representation that is designed to work at low resolutions [8, 4]. Such representations are often based on handcrafted features (such as color). In our approach, we learn resolution-specific features instead of hand-crafting them. Additionally, we employ super-resolution networks as a pre-processing stage that is trained end-to-end with the resolution-specific embedding.

Perhaps the most relevant work to ours is [33], which learns a *fixed-resolution* deep network to regress a high-res embedding from low-res images using a L2 loss. In comparison, we learn *multi-resolution* embeddings that are directly

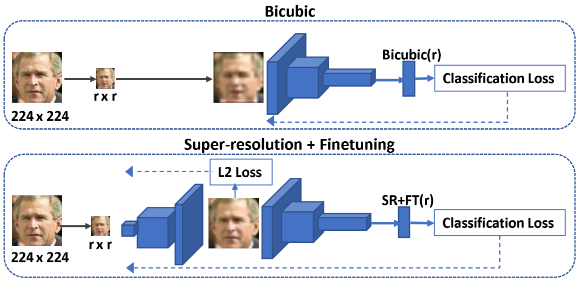

Figure 5: We describe different strategies for learning embedding networks tuned for a particular resolution of r pixels, that make use of pre-training. **Bicubic** interpolates training images of size r to a canonical resolution (224x224), and then fine-tunes a pre-trained embedding network. **Super-resolution(SR)** replaces bicubic interpolation with an off-the-shelf super-resolution network (not shown in figure). **SR+Finetuning(SR+FT)** fine-tunes both the front-end super-res network and the embedding network.

trained to minimize (categorical) identity mis-classifications.

3. Method

As argued above traditional face recognition models suffer a massive drop in performance on low-resolution images (Fig. 3). In this section, we explore various simple strategies to remedy this. We make use of an artificially-resized LFW dataset where all images are sized to a target resolution of X pixels (denoted as LFW-X) to support design decisions.

3.1. Resolution-specific models

The most intuitive way to alleviate the impact of resolution is to train separate models for specific resolutions. But, how does one train an embedding for say, a 16x16 image?

Training images: Ideally, we should train these models with *real* low-resolution images of size 16x16, but in general, there may not be enough in a given training set. An attractive alternative is to augment the training set with resized images, a common practice in multi-scale training. We find that upsampling images may introduce blurry artifacts, but downsampling is a relatively benign form of augmentation (even given the caveats of [5]). In practice, we downsample images from VGGFace2 to the resolution of interest to train resolution-specific models for all resolutions < 60.

Pre-training: Armed with a training set of 16x16 images, which network architecture do we use to learn an embedding? One option is training a custom architecture from scratch for that resolution. But this makes it hard to take advantage of pretrained backbone networks trained on faces resized to a fixed input size (finetuning networks pretrained on high-res images was shown to perform better than training them from scratch on low-res images [25, 29]). So, we *upsample* the downsampled images *back* to 224x224 with **Bicubic** interpolation, and fine-tune a ResNet-50 (pretrained

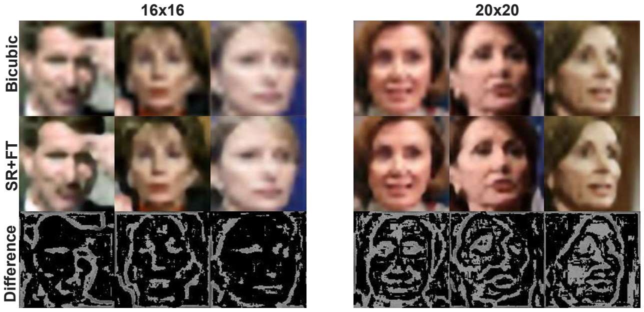

Figure 6: We illustrate the difference in upsampling strategies, on images of height 16 (left), and images of height 20 (right). **Bicubic** interpolated images are shown in the top row, while **SR+FT** upsampled images are shown in the central row. We can observe that the **SR+FT** upsampled images are sharper near the edges from the difference images in the bottom row. Zoom in to contrast between the sets of images.

on VGGFace2 at full-resolution) on such training images. To evaluate this approach, we train and test a face verification model on LFW-X. Fig. 4 demonstrates that simple resolution-specific models results in a dramatic relative improvement over an off-the-shelf embedding (**VGG2**): 60% for LFW-16 and 15% for LFW-20.

Super-resolution (SR): We posit that the specific method for upsampling the input image might have a large effect on recognition performance. Fig. 5 replaces the bicubic upsampler with a (lightweight) super-resolution (**SR**) network. Interestingly, Fig. 4 demonstrates that super-resolution networks may lose identity relevant information (also observed in [16]). In supplementary material, we show that this effect is even more pronounced with deeper state-of-the-art super-res networks operating on *real* images.

Super-resolution with Fine-tuning (SR+FT): Finally, we finetune the lightweight super-resolution network along with the backbone face embedding model with categorical cross-entropy loss, to guide the SR model to retain identity information. Fig. 4 shows that **SR+FT** outperforms bicubic interpolation. Fig. 6 visualizes images generated by the finetuned super-resolution network, which are sharper than the bicubic result.

Multiple resolution-specific embeddings: Fig. 4 suggests models tuned for particular resolutions (16px) might outperform models tuned for similar but distinct sizes (20px). To avoid training an exorbitant number of models, we choose a fixed number of 'anchor resolutions' r spaced along a linear scale of 16px, 35px, and 50px. We found this to provide a good tradeoff of memory and performance. Please see the Experiments section for additional details.

3.2. Multi-resolution (MR) embeddings

The above results suggest that one should train a set of resolution-specific models to improve recognition performance. It is natural to ask if these different resolution-specific embeddings could be *ensembled* together to improve performance. In order to apply a different network to a given input image, we would need to upsample or downsample it. As previously argued, downsampling an image is less prone to introducing artifacts, unlike upsampling. This suggests that given an image at a fixed resolution, one can ensemble together embeddings tuned for lower resolutions by downsampling.

Independent MR (MR-I): A reasonable solution is to concatenate these embeddings together to produce a 'multi-resolution' embedding.

$$\Phi(x) = \begin{bmatrix} \frac{\phi_1(x_1)}{\|\phi_1(x_1)\|} & \frac{\phi_2(x_2)}{\|\phi_2(x_2)\|} & \dots & \frac{\phi_\mathrm{n}(x_n)}{\|\phi_\mathrm{n}(x_n)\|} \end{bmatrix} \quad (1)$$

where, x_i is a lower resolution version of a given image x resized to anchor height r_i, and ϕ_i denotes a resolution-specific model tuned for that specific resolution. We find that normalizing each resolution-specific embedding is necessary to match the relative scales of the embeddings.

MR-I inference: Given an input image at a particular resolution, we create its downsampled versions corresponding to anchor resolutions of equal or smaller size. This collection of blurred images are processed with resolution-specific streams that produce embeddings that are concatenated together to produce the final multi-resolution vector given by Equation 1 for **MR-I** models. With such a representation, the similarity score between an image pair (x, y) downsampled to the same anchor resolution is evaluated as follows,

$$s(\Phi(x), \Phi(y)) = \frac{\Phi(x)^T \Phi(y)}{\|\Phi(x)\| \, \|\Phi(y)\|} \quad (2)$$

The similarity score is equal to the mean cosine similarity of the resolution-specific embeddings. Qualitatively, this is equivalent to comparing probe and reference images at multiple scales.

Jointly-trained MR (**MR-J**): Because the above approach naively concatenates together independently-trained embeddings, they might contain redundant information. To truly *disentangle* features across scale, we would like to jointly train all constituent resolution-specific embedding "streams" of a network. Following the grand tradition of *residual network*s [10], joint training would force the resolution-specific streams tuned for higher resolutions to learn residual complementary information.

Fig.7 demonstrates the operation of a joint multi-resolution model. It shows that certain parts of an MR-J network are designed to only operate on inputs of certain resolutions, while other parameters are shared. For example, given a low resolution image ($\mathrm{r_1 x r_1}$), the network outputs only a part of the overall embedding (blue), while it outputs the full embedding for a higher resolution image($\mathrm{r_3 x r_3}$).

Method	Embed dim.	TPR at 1e-3 FPR
MR-J	128	61.2
MR-I	128	60.7
SR+FT	128	54.3
VGG2	2048	38.7

Table 1: **Given a fixed embedding dimension (say 128), does MR embedding perform better than its fixed counterparts?** The table shows that **MR** embeddings, both joint and independent, composed of two 64 dimensional embeddings perform much better than a single resolution embeddings of same size **SR+FT**, and also the 2048 dimensional baseline model **VGG2** on real low resolution images (height < 40px.). We use real images to better visualize the difference between the models. Full plots are shown in the supp. material.

Given an input image, the outputs of these resolution specific streams are concatenated together to output a true multi-resolution embedding as discussed earlier. We show in the supplementary material that joint training forces higher resolution streams to learn to ignore low resolution features like gender [31] etc., demonstrating that they encode disentangled features.

Parameter sharing: What is the optimal policy to share parameters between the resolution-specific streams? We experiment with two extreme strategies to help us identify the ideal approach. (a) we test a model in which no parameters are shared across different resolutions, i.e. each stream operates independently till the final output stage. We refer to this model as **MR-J(W)** or MR-J(Wide). (b) at the other extreme, we test another model in which small 3-layered resolution-specific streams operate on an embedding output by a fully shared network. We refer to this model as **MR-J**. As a consequence of aggressively sharing parameters across different resolutions, **MR-J** is much more efficient than **MR-J(W)**. Its memory footprint and computational complexity are comparable to a single ResNet-50 model (25M vs 23M params). We direct the reader to the supplementary material for a detailed description of the training scheme.

In the Experiments section, we show that multi-resolution embeddings significantly outperform **VGG2**, and also our resolution-specific models **SR+FT**.

Embedding dimension: We would like embeddings with small memory footprints. Our multi-res embeddings might generate large memory footprints if implemented naively. Table 1 asks the salient question: given a target dimension for an embedding (of say, 128d), do multi-resolution embeddings outperform their fixed counterpart? The answer is yes! Multi-resolution embeddings composed of two 64 dimensional embeddings (**MR-I, 128-dim** and **MR-J, 128-dim**) outperform single-res embeddings of equal size (**SR+FT,128-dim**) which are trained on the same data with the same loss function.

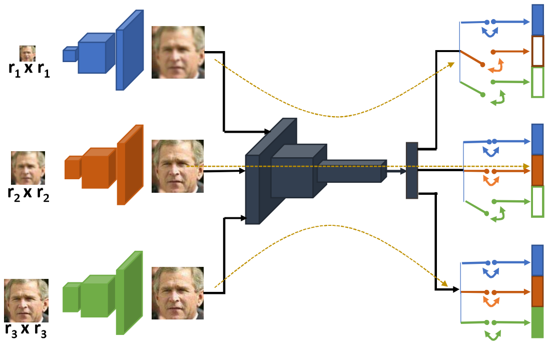

Figure 7: Jointly trained multi-resolution embedding **MR-J**. Each low-res image is super-resolved by **SR**. The figure shows that certain parts of the network are designed to only operate on images of specific resolution. These parts output embeddings tuned to images of those resolutions. As discussed earlier, (1)they adapt in complexity to the resolution of the input, such that larger embeddings are produced when additional high-res information is available (bottom). (2)they produce disentangled representations where frequency-specific components can be "switched off" when not presenting the input (top/centre). (3) they can be adapted on-the-fly to any desired resolution

TPR at FPR 1e-3					
LFW-16 vs LFW-25			LFW-20 vs LFW-25		
SR+FT(16)	SR+FT(25)	VGG2	SR+FT(20)	SR+FT(25)	VGG2
94.1	89.0	74.5	96.7	96.7	88.5

Table 2: **Given a probe and gallery image pair of different resolutions, what should be the resolution of the embeddings used to compare them?** The table shows that in case of a large mismatch in resolution of the probe and the gallery image: the best performance is achieved by resizing the higher resolution image (25 px) to the lower resolution (16 px), and employing lower-resolution (16 px) embedding (left). If the mismatch is not large, we can use either representation (right). Full plots are shown in the supp. material.

3.3. Adaptive inference

Choosing the ideal representation: Thus far, our results indicate that when comparing two images at a particular resolution, we should use **MR** embeddings tuned for that resolution. Now, what about comparing two faces at *different* resolutions? Two natural options are (a) downsample the larger image to the smaller size, and use a model tuned for the smaller resolution or (b) upsample the smaller image and use a model tuned for the larger image. We analyze these strategies along with the baseline approach on dissimilarly resized LFW datasets for a clean evaluation. Table. 2 shows that when the two resolutions are similar (20px vs 25px), it doesn't quite matter. But for a large mismatch (16px vs 25px), (a) using a representation tuned for the lower

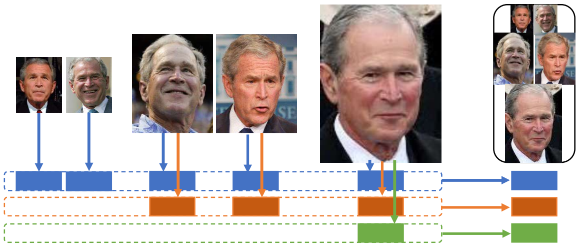

Figure 8: **Multi-resolution pooling.** The embeddings corresponding to each scale are pooled separately to generate the final multi-resolution representation of a template, as described by Eqn.4

resolution image is more effective.

Adaptive multi-resolution inference: Assume we are given a gallery of high-resolution face images. Our model produces a multi-resolution embedding that is stored for all gallery images. Given a probe of a particular size r, our prior experiments suggest that we should tune the gallery to the closest-anchor resolution, r_i. This is trivial to do with a *disentangled* multi-resolution embedding. Simply tune the gallery embeddings "on-the-fly" with array indexing:

$$\Phi(x)[1:i] \tag{3}$$

Multi-resolution pooling: Practical face recognition methods often operate on *sets* of faces (say, extracted from a video sequence). Such methods generate an aggregate template representation by pooling embeddings of face images in the set. The templates are then used to efficiently compare these sets with a single image or with an other set of faces. In our supplementary material, we show that naive pooling of our multi-resolution embeddings is not optimal. Intuitively, naive pooling mixes information across scales. Rather, we should use only high-resolution faces to construct the pooled high-frequency feature. We operate on the i^{th} anchor resolution as follows:

$$\bar{\phi}_i = \frac{1}{|S_i|} \sum_{x_i \in S_i} \phi_i(x_i) \tag{4}$$

where S_i is the set of images in the set that are of at least the resolution of r_i, and $\bar{\phi}_i$ is the pooled feature for anchor resolution r_i (Fig. 8). These features are concatenated to output a multi-resolution template embedding, as done earlier.

4. Experiments

As argued earlier, we focus our final results on the IJB-C dataset because it includes *real* low resolution images. We create 4 resolution constrained subsets of low resolution faces (height < 60) from the IJBC dataset to test the effectiveness of our algorithm at various scales. Each of these subsets, named **IJBC-**X, contains faces of height close to $X \in \{20, 25, 35, 50\}$. For example, a face of height 28 px is placed in the IJBC-25 subset. We will make these subsets publicly available.

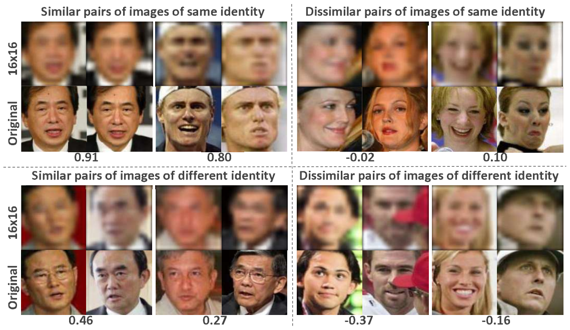
Similar pairs of images of same identity
Dissimilar pairs of images of same identity
16x16
Original
0.91
0.80
-0.02
0.10
Similar pairs of images of different identity
Dissimilar pairs of images of different identity
16x16
Original
0.46
0.27
-0.37
-0.16

Figure 9: We visualize the salient features captured by a 16px embedding by plotting both the low-res image pairs and their high-res counterparts. The top-left quadrant show face pairs of the same identity with high cosine similarity (shown at the bottom of each pair). The top right shows face of same identity with low similarity (due to expression or makeup changes). The bottom left mistakes suggest that the low res model relies heavily on racial and face shape cues, as different people from the same race are predicted to have high similarity. The bottom right suggests that gender appears to be an easily distinguishable feature at low resolution, previously observed in [31]

In the following subsections, we discuss the results of our algorithms on probe images drawn from these splits when tested under various protocols of the IJB-C dataset and compare them with the baseline VGG2. Additionally, we compare our results with a VGG2 model finetuned with artificially downsampled images of *all* resolutions, **FT-all** and show that our models massively outperform it. This demonstrates that it is necessary to handle images of very different resolutions with resolution-specific streams.

4.1. Single image verification

Setup: The simplest IJB-C protocol is 1:1 covariate verification, where a single probe image is compared with a single reference image. The protocol specifies over 48M verification pairs from which we sample those pairs with at least one low resolution image (height < 60). We bin verification pairs into one of 4 groups, **IJBC Covariate-**X, when the lower resolution image in the pair belongs to IJBC-X.

Results: First, we begin with qualitative results for resolution-specific verification (Fig. 9). We refer the reader to the caption for a more detailed analysis. Fig. 10 shows the true positive rate (TPR) at 1e-3 false positive rate (FPR). The plot shows that a simple resolution-specific model tuned for images of height 16, (both MR-J, SR+FT) almost **doubles** the performance of VGG2 on both IJBC Covariate-20, IJBC Covariate-25. Note that for the lowest anchor resolution (16x16), MR-J is same as SR+FT. Similarly, resolution-

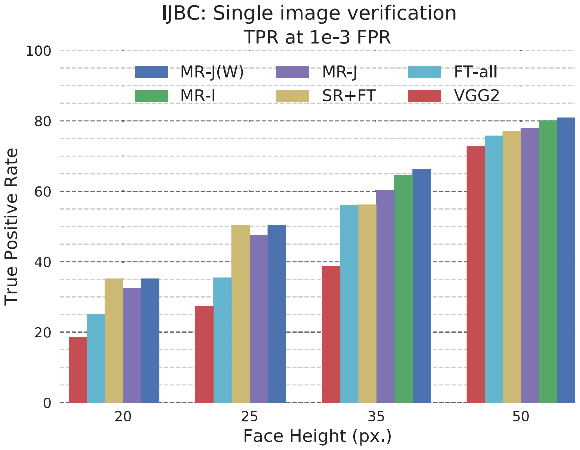

Figure 10: **Performance on Single image verification**. The plots show the TPR at 1e-3 FPR for various methods at different input resolutions. The plots are generated by verifying pairs from IJBC Covariate-X. We observe that MR-J(W) almost **doubles** VGG2's performance and easily outperforms FT-all for IJBC Covariate-20, and IJBC Covariate-25. SR+FT surpasses VGG2 by **45%** on IJBC Covariate-35, and **6%** on IJBC Covariate-50. Remarkably, MR models outperform VGG2 by **70%** on IJBC Covariate-35 and **11%** on IJBC Covariate-50. Notice that MR-I models outperform MR-J models at both these resolutions. It is interesting to observe that the difference between our best models and FT-all increases with decrease in probe resolution. Full ROC plots are presented in the supplementary material.

specific models SR+FT, exceed the baseline's performance by 45% on IJBC Covariate-35, and 6% on IJBC Covariate-50 respectively. More importantly, we draw attention to the remarkable performance of multi-resolution embeddings, MR-J(W), MR-J and MR-I. We find that the MR models outperform VGG2 by 70% on IJBC Covariate-35, and 11% on IJBC Covariate-50. They also easily surpass the resolution-specific models and FT-all. All relative improvements are reported at 10^{-3} False Positive Rate.

Discussion: (a)Why do MR models massively outperform other models? Disentangling resolution-specific features forces models to learn to encode scale-specific features which were ignored when trained on higher resolution images. Also, verifying faces by comparing them at multiple scales seems to help recognition.

(b)In particular, we demonstrate that although FT-all and MR-J are trained on same images, with the same loss, and similar size (25M vs 23M params.), the small resolution-specific streams operating at the top of MR-J greatly improve its recognition performance at all low resolutions. FT-all also allows us to show that an unmodified single ResNet model cannot optimally encode both low and high resolution features.

(c)MR-J(W) models slightly outperform MR-I models. This shows that joint training of multi-resolution models enjoys an advantage over training independently, as they do not encode redundant information. MR-J(W) also slightly outperform MR-J. We propose that, apart from model complexity (3 times larger), the inability of a single network to optimally model scale variation is also a contributing factor.

(d) In our experiments, we observed that a model tuned for images of height 16 alone performs better than tuning multiple resolution-specific models for images of height < 30. This is surprising, as we would expect appropriately tuned resolution-specific models to perform better! One probable reason is that the *effective resolution* of real images is influenced by other factors such as JPEG compression, motion blur etc., and the additional blur created by using a model tuned for a lower resolution assists in dealing with them. This observation suggests that a multi-resolution model can naturally handle these factors by adopting an embedding tuned to a lower resolution.

(e) The difference between our best models, and FT-all increases with a drop in resolution. Also the performance of our MR-J model which shares parameters across all resolution drops in comparison to MR-J(W). This observation validates our method, as it shows that lower resolution images need separate models for optimal performance.

4.2. Identification

Setup: Given a face image from IJBC-X, this protocol asks which one of N (3531) identities does it belong to? Each of the N subjects in the gallery is represented by a set of high quality images. It is an important protocol resembling the operational work of law enforcement [23]. Moreover, it allows us to test test *multi-resolution pooling*, and *adaptive inference* for multi-resolution embeddings.

Results: Fig. 11 presents the percentage of probe images which had the ground truth (GT) in one of their top-10 predictions for each of our models and the baseline over various IJBC-X. From the figure, we observe that the resolution-specific embeddings MR-J(W) **quadruples** the performance of VGG2 for probes from IJBC-20, and **double** the baseline's performance for probes from IJBC-25. Similar to earlier experiment, SR+FT surpasses VGG2's performance by 44% and 13.5% for IJBC-35, IJBC-50 respectively.

We can observe that MR-I, and MR-J again outperform the baseline by 66% on probes from IJBC-35, and 22% on IJBC-50. Also, MR models' significantly better performance validates adaptive multi-resolution inference.

4.3. Image set-based verification

Setup: This is the more common 1:1 verification protocol defined in IJB-C dataset[23]. In this setting, probe sets are compared with gallery sets. We sample relevant probe sets with more than 60% images of very low resolution

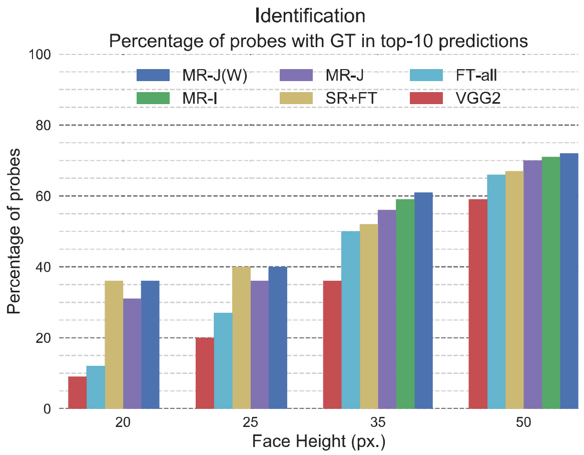

Figure 11: **Performance on Identification.** The plots show the percentage of probes with GT in their top 10 predictions. The plots were generated by classifying single image probes from IJBC-X, to one of a defined gallery of 3531 classes, each represented by a *set* of images. The plots for IJBC-20, and IJBC-25 show that MR-J(W) at least **doubles** VGG2's performance. The plots for IJBC-35, and IJBC-50 show that SR+FT models perform much better VGG2. They also demonstrate that MR models surpass VGG2's performance by **66%** and **22%** respectively. The full CMC curves are presented in the supplementary material.

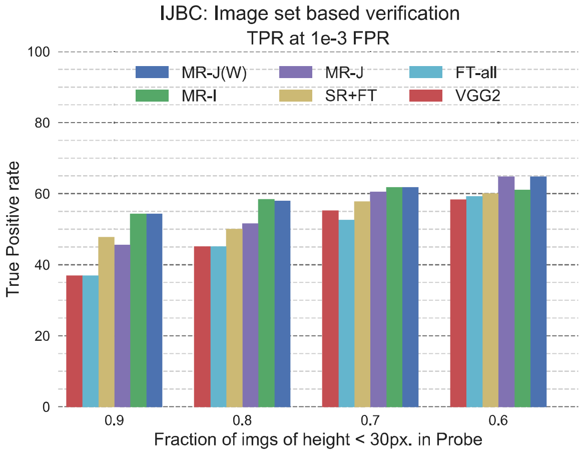

Figure 12: **Image set based verification.** These plots show TPR at 1e-3 FPR, for probe sets with varying ratios of low-resolution images. SR+FT outperforms VGG2 and FT-all at higher ratios (0.8, 0.9). MR models (particularly MR-J) outperform all other approaches with increasingly larger margins for higher ratios. The full ROC plots are presented in the supplementary material.

	Rank 1. Acc.				
Face height	MR-J(W)	MR-J	SR+FT	FT-all	VGG2
20	40.1	38.9	40.1	32.0	15.9
35	71.5	70.7	70.2	58.0	51.0
50	79.2	77.5	77.4	64.3	65.2

Table 3: Rank-1 accuracy on downsized images (height={20,35,50}) of the Megaface dataset (100K image disctractor set). The table shows that our multiresolution models continue to outperform the baseline models (VGG2, FT-all), and also the SR+FT models. However, note that the difference between SR+FT and MR-X is not high because the test images are artificially downsampled and the models may overfit to this downsampling method.

(height<30) to perform this experiment.

Results: In the plots of Fig. 12, we show our results with probes containing increasing fractions of very low resolution images. The figure shows that the SR+FT outperforms VGG2, and FT-all, by 11%, 30% respectively, on probe sets with larger fraction of very low res images (0.8, 0.9). Their performances are comparable for probe sets with lower fractions (0.6, 0.7) of low res images, as SR+FT is unable to capitalize on the additional high-res information in the probe set. We show that both MR models outperform all other approaches with increasingly larger margins on probe sets with increasing fractions of low resolution images. Particularly, the MR-J, MR-J(W) models beat the baseline by 11.1%, 11.9%, 28.8%, 47.1% for probe sets with fraction of low resolution images greater than 0.6, 0.7, 0.8, and 0.9 respectively, proving that the MR models optimally combine both high-resolution and low-resolution features of images in the probe and reference sets.

4.4. Megaface

Megaface is a popular large-scale testing benchmark for face recognition. However, the dataset does not contain images of low resolutions. To test our method at this large scale, we resize all images in the Megaface dataset to specific sizes before evaluating our methods on these resized images. Table 3 shows the Rank-1 accuracy obtained by our models and the baseline at various such sizes. All results are obtained by using a distractor set of 100K images.

5. Conclusion

We propose a simple yet effective approach for recognizing faces at low resolution. We first point out that state-of-the-art face recognizers, which use fixed-resolution embeddings, perform dramatically worse as face resolution drops below 30 pixels. We then show that by simply tuning resolution-specific embedding we can significantly improve the recognition accuracy. We further explore multi-resolution embedding that efficiently adapts in size and complexity to the resolution of test image *on-the-fly*. Finally, comparing to state-of-the-art fixed-resolution embeddings, our proposed embedding dramatically reduces recognition error on small faces on standard benchmarks.

References

[1] E. H. Adelson, C. H. Anderson, J. R. Bergen, P. J. Burt, and J. M. Ogden. Pyramid methods in image processing. *RCA engineer*, 1984.

[2] S. Baker and T. Kanade. Hallucinating faces. In *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pages 83–88. IEEE, 2000.

[3] S. Baker and T. Kanade. Limits on super-resolution and how to break them. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(9):1167–1183, 2002.

[4] S. Biswas, K. W. Bowyer, and P. J. Flynn. Multidimensional scaling for matching low-resolution face images. *IEEE transactions on pattern analysis and machine intelligence*, 34(10):2019–2030, 2012.

[5] A. Bulat, J. Yang, and G. Tzimiropoulos. To learn image super-resolution, use a gan to learn how to do image degradation first. *arXiv preprint arXiv:1807.11458*, 2018.

[6] A. M. Burton, S. Wilson, M. Cowan, and V. Bruce. Face recognition in poor-quality video: Evidence from security surveillance. *Psychological Science*, 10(3):243–248, 1999.

[7] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. *arXiv preprint arXiv:1710.08092*, 2017.

[8] J. Y. Choi, Y. M. Ro, and K. N. Plataniotis. Color face recognition for degraded face images. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(5):1217–1230, 2009.

[9] L. D. Harmon. The recognition of faces. *Scientific American*, 229(5):70–83, 1973.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[11] P. H. Hennings-Yeomans, S. Baker, and B. V. Kumar. Simultaneous super-resolution and feature extraction for recognition of low-resolution faces. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[12] P. H. Hennings-Yeomans, B. V. Kumar, and S. Baker. Robust low-resolution face identification and verification using high-resolution features. In *Image Processing (ICIP), 2009 16th IEEE International Conference on*, pages 33–36. IEEE, 2009.

[13] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[14] M. Iacopo, S. Rawls, G. Medioni, and P. Natarajan. Pose-aware face recognition in the wild. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[15] P. Jonathon, A. Yates, Y. Hu, C. Hahn, E. Noyes, K. Jackson, and J. Cavazos. Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences*, 2018.

[16] Z. Kaipeng, Z. Zhang, C.-W. Cheng, W. H. Hsu, Y. Qiao, W. Liu, and T. Zhang. Super-identity convolutional neural network for face hallucination. *ECCV*, 2018.

[17] G. Koch, R. Zemel, and R. Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML Deep Learning Workshop*, volume 2, 2015.

[18] T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie. Feature pyramid networks for object detection. *CVPR*, 2017.

[19] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[20] C. Liu, H.-Y. Shum, and W. T. Freeman. Face hallucination: Theory and practice. *International Journal of Computer Vision*, 75(1):115–134, 2007.

[21] C. Liu, H.-Y. Shum, and C.-S. Zhang. A two-step approach to hallucinating faces: global parametric model and local nonparametric model. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–I. IEEE, 2001.

[22] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. Sphereface: Deep hypersphere embedding for face recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, 2017.

[23] B. Maze, J. Adams, J. A. Duncan, N. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Anderson, J. Cheney, et al. Iarpa janus benchmark-c: Face dataset and protocol.

[24] O. M. Parkhi, A. Vedaldi, A. Zisserman, et al. Deep face recognition. In *BMVC*, volume 1, page 6, 2015.

[25] X. Peng, J. Hoffman, X. Y. Stella, and K. Saenko. Fine-to-coarse knowledge transfer for low-res image classification. In *Image Processing (ICIP), 2016 IEEE International Conference on*, pages 3683–3687. IEEE, 2016.

[26] R. Ranjan, C. Castillo, and R. Chellappa. L2-constrained softmax loss for discriminative face verification. *arXiv preprint arXiv:1703.09507*, 2017.

[27] J. Sadr, I. Jarudi, and P. Sinha. The role of eyebrows in face recognition. *Perception*, 32(3):285–293, 2003.

[28] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[29] B. Singh and L. S. Davis. An analysis of scale invariance in object detection snip. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3578–3587, 2018.

[30] P. Sinha, B. Balas, Y. Ostrovsky, and R. Russell. Face recognition by humans: Nineteen results all computer vision researchers should know about. *Proceedings of the IEEE*, 94(11):1948–1962, 2006.

[31] S. Tamura, H. Kawai, and H. Mitsumoto. Male/female identification from 8x6 very low resolution face images by neural network. *Pattern recognition*, pages 331–335, 1996.

[32] Z. Wang, Z. Miao, Q. J. Wu, Y. Wan, and Z. Tang. Low-resolution face recognition: a review. *The Visual Computer*, 30(4):359–386, 2014.

[33] E. Zangeneh, M. Rahmati, and Y. Mohsenzadeh. Low resolution face recognition using a two-branch deep convolutional neural network architecture. *arXiv preprint arXiv:1706.06247*, 2017.

[34] H. Zhao, S. Jianping, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. *CVPR*, 2017.

[35] W. W. Zou and P. C. Yuen. Very low resolution face recognition problem. *IEEE Transactions on Image Processing*, 21(1):327–340, 2012.