Optimizing Super Resolution for Face Recognition

Antonio Augusto Abello and R. Hirata Jr.
Email: {abello,hirata}@ime.usp.br
Instituto de Matemática e Estatística
Universidade de São Paulo
São Paulo, Brazil, 05508-090

***Abstract*—Face Super-Resolution is a subset of Super Resolution (SR) that aims to retrieve a high-resolution (HR) image of a face from a lower resolution input. Recently, Deep Learning (DL) methods have improved drastically the quality of SR generated images. However, these qualitative improvements are not always followed by quantitative improvements in the traditional metrics of the area, namely PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity Index). In some cases, models that perform better in opinion scores and qualitative evaluation have worse performance in these metrics, indicating they are not sufficiently informative. To address this issue we propose a task-based evaluation procedure based on the comparative performance of face recognition algorithms on HR and SR images to evaluate how well the models retrieve high-frequency and identity defining information. Furthermore, as our face recognition model is differentiable, this leads to a novel loss function that can be optimized to improve performance in these tasks. We successfully apply our evaluation method to validate this training method, yielding promising results**

I. Introduction

Single Image Super Resolution (SISR) is the task of retrieving a high-resolution (HR) image from a low-resolution (LR) input. It is an ill-posed problem, since a high-resolution image can generate various low-resolution counterparts and vice-versa. Evaluation and comparison of methods is thus a difficult task. The most commonly used metrics, the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) both presuppose a ground-truth example and act on a pixel-by-pixel or window-by-window basis, characteristics deemed problematic. It is also known that these metrics correlate poorly with human perception [1], [2].

An emblematic case of the insufficiency of current metrics is that of the use of Generative Adversarial Networks (GANs) [3] and Perceptual Losses [4]. While yielding results that were clearly superior qualitatively, these techniques had a lower quantitative performance on those metrics. Mean Opinion Scores (MOS) [5] were then used to confirm these methods generate more aesthetically pleasing and overall more credible images. However these experiments are hard to replicate, prone to biases and deviations due to sample size and selection and are ultimately still based purely on human subjectivity. It has become common to report new developments in SISR with two versions: one trained with GANs for visually compelling examples and qualitative evaluation and one trained alone for quantitative analysis [6]

In this context we study ways to use other Computer Vision (CV) tasks as proxies for the quality of generated images, a framework known as task-based evaluation. The benefits of this approach is twofold: it helps integrate SR with other fields of CV, approximating model evaluation to the practical, actual use cases of the model, and defines new hard quantitative metrics that may bring new insight and more powerful justification for present and future models.

Face Super Resolution, sometimes called Face Hallucination, is the specific subset of SR that deals with resolving LR images of human faces. As so, there are a number of applications under the Face Recognition umbrella that can be aided with SR [7], [8], such as face verification (defining whether two images belong to the same person) or face identification (attributing an identity to an image of a face) [9]. We then also study Face Recognition models and methods to build an evaluation procedure based on the performance of super-resolved images in these tasks.

As we have found that most of the current state-of-the-art methods for face recognition are based on differentiable models [10], [11], we are also able to optimize our SR models specifically to perform well on these tasks. Using a pre-trained Face Recognition (FR) model we build a "FR Loss" based on the distance between the super-resolved image and the ground-truth on the FR model's representation. Our loss function would express how well our super-resolution model is recovering identity-defining information. We evaluate it under our evaluation procedure and get motivating results.

The main contributions of this work are then:

- we develop a robust task-based evaluation protocol for Face Super-Resolution models using FR tasks and apply it to state-of-the-art models
- we develop a new method of training, involving minimizing a "FR Loss" that aids SR Models to recover identity-defining information

The rest of the paper is structured as follows: we perform a literature review in Section II, formalize our proposal in Section III, describe our experimental design in Section IV, present results and brief discussion in Section V and conclude in Section VI.

II. Related Work

A. Single Image Super-Resolution and Face Super-Resolution

Since Dong et al's introduced the Super-Resolution Convolutional Neural Network (SRCNN) [12], deep learning methods became the state-of-the-art for single image SR.

Further developments on upsampling techniques [13], network architectures [3] and others have continued to improve results both in quantitative and qualitative ways. Wang et. al. [14] presents a more in-depth review of the development of the area.

Recently, researchers have found that minimizing a pixel-by-pixel loss function alone may lead to over-smooth results, i.e., images that lack high-frequency details [3], [4]. To address this problem, more complex loss functions have been devised to take into account image quality in a more global way, such as the perceptual loss [4], the adversary loss [3], or losses based on the wavelet transform [15]. Most of the time, these innovations yield worse results quantitatively, in terms of PSNR and SSIM, but better qualitatively, and subjectively, through Mean Opinion Scores (MOS).

Although methods of general SR still work on face images, techniques exploiting unique properties of these images exist [6].

B. Deep Face Recognition

Taigman et al's work in 2014 [10] introduced a Deep Learning-based Face Recognition approach that beat and quickly became the state-of-the-art for various FR tasks. It consisted of a Deep Neural Network trained first on a closed-set scenario as a multiclass classifier, using the softmax activation function and minimizing cross-entropy. Since the classifier must have learned an useful representation of faces in order to separate the classes, the authors hypothesize that this representation may be useful for an open-set scenario.

The authors validate this hypothesis empirically by using an intermediate layer of the classifier as an embedding for general face images. Simple models for face verification were trained using the classifier's embedding as input and achieved results far greater than the state-of-the-art then, on datasets with different faces than the ones used in training. Impressive results were achieved even using simple methods such as the euclidean distance as a verification metric.

Further developments in Deep Face Recognition were made in order to take advantage of large-scale face datasets. Parkhi et al [11] develop a new loss function that can be used to train the embedding on an open-set scenario, based on the distance between positive and negative examples of generic identities. Developments in the softmax loss [16] were also made for increasing discriminability and also facilitating training.

C. Neural Networks as Kernel Functions

The idea of using pre-trained neural networks as "embeddings" or "kernel functions" is not new, specially the idea of using distance metrics in these embeddings as loss functions or evaluation metrics. Johnson et al [4] use the output of a VGG16 network pre-trained to classify examples on ImageNet to define a "Feature Reconstruction Loss" and a "Style Reconstruction Loss" that are then used for SR and style-transfer.

The output of an intermediate layer of an Inception-like network [17] also pre-trained on ImageNet is commonly used as an evaluation metric for generative models. It is generally refined into the Inception Score (IS) [18], or Frchet Inception Distance (FID) [19]

While these losses and metrics are intuitive and, more importantly, experimentally successful, there is no clear theoretical justification in using these determinate networks and not other ones for distance metrics.

The approach proposed in this paper can be thought of as a variant of these methods, but with a crucial difference. In our case, the embedding space and the distance chosen are already semantically meaningful, as they express differences or similarities in face characteristics.

D. Task-based evaluation and training

Dai et al [20] previously argued that SISR is mostly evaluated perceptually. They proceeded to do a review of the state-of-the-art methods and their effects on other CV tasks, with generally positive conclusions about the effect of SR in other CV tasks, and asking for further integration between SISR and other subfields of CV.

Since Face Super-Resolution has a natural use-case in surveillance applications, task-based evaluation seems to be more common in this area. Before the emergence of Deep Learning, Hu et al [21] investigate the effects of SR on surveillance applications. Rasti et al [7] train CNNs for super-resolution of faces and evaluate them using the performance of a Hidden Markov Model (HMM) model for face recognition. These works focus more on face verification tasks, while our work extends also into face identification, as described in Section III

There are previous works using information from other CV tasks to aid SR. This idea can also appear under the framework of multi-task learning. Bulat et al [22] train a network to perform both facial landmark estimation and SR at once. Haris et al [23] develops an approach similar to ours but in regards of general object-detection instead of Face Recognition. They train a SR Network to minimize both the reconstruction loss and the error of a pre-trained neural network for object detection on the super-resolved images. We instead focus on face images and Face Recognition and define our loss function in a different way, presented in Section III.

Zhang et al. [24] propose to jointly optimize separate Face Recognition and Super Resolution models and develop techniques for this joint training, that would result in FR models robust to differences in resolution and SISR models that can recover identity information. The joint training leads to some confusion in the experimental design, though, which overlooks the generated images in favor of evaluating the jointly trained FR model. They present three evaluation protocols: Visual Quality, quantitative and qualitative analysis of generated images, Identity Recovery, which measures the cosine similarity of super-resolved images and original images on the trained FR model's embedding, and Identity Recognizability, which trains a new FR Model on super-resolved images and test its performance on traditional FR benchmarks

As most of these evaluation protocols involve both the generated images and the trained embedding, there is little evidence about the quality of the super-resolved images. When they are considered on its own they use only the traditional SISR metrics and the image's distance on the jointly trained FR Model, which may be biased in favor of the network it was trained with. Our evaluation protocols, defined on Section III, produce a more fine-grained view of the amount of information present in the super-resolved images by considering them on their own.

III. Method

In this section we formally define our models, training and testing methods.

A. Single Image Super-Resolution Networks

A SISR Network is a neural network that aims to retrieve a high-resolution image from a low-resolution input. It can be thought of as a parametrized mapping, M,

$$I_{SR} = M(I_{LR}, \theta), \tag{1}$$

that produces a super-resolved image (I_{SR}) from a low-resolution image (I_{LR}), where θ represents the Neural Network's parameters. On a real-world scenario we generally do not have access to the high-resolution version of the image (I_{HR}). Therefore, for training we usually model a degradation process D that produces low-resolution images, I_{LR}, from high-resolution ones, I_{HR}, presented in the original image datasets:

$$I_{LR} = D(I_{HR}, \delta), \tag{2}$$

where δ represents the degradation parameters such as scale. In this work we use for degradation model a simple downsampling operation via interpolation alongside with an anti-aliasing blur kernel.

This degradation model is used to produce pairs of low-resolution and ground-truth high-resolution images. A SISR Network then receives the LR image and produces a super-resolved proposal.

Through comparison between the super-resolved image and the HR ground truth, we can then define a loss function that express the distance between the model's output and the desired output, that turns learning feasible.

The most common loss function is simply the normalized L2 norm between each image, also called MSE (Mean Squared Error) loss or pixel-loss:

$$\text{MSE} = ||I_{SR} - I_{HR}||_2 \tag{3}$$

B. Face Recognition Networks

A Deep Face Recognition Network (FR Network) is a CNN that produces a real-valued vector representation of face images. It can be thought as an embedding, ϕ, given by:

$$\phi : I \to R^n, \tag{4}$$

where I is a set of images.

This embedding is trained in such a way that proximity for a certain similarity measure means proximity of face characteristics, and can be used to determine whether two images are from a shared identity or to classify images according to different identities.

C. Face Recognition Loss

Using an FR model we can define then our novel FR Loss. Similarly with a perceptual loss, given an FR network ϕ, the FR Loss is defined as:

$$FR = ||\phi(I_{SR}) - \phi(I_{HR})||^2 \tag{5}$$

This FR Loss is different from the task-based loss of Haris et al [23] in the sense that it is not oriented towards an specific task, but to Face Recognition as a whole. We understand that by being more abstract our loss leads our models to recover facial characteristics in general, and not only characteristics relevant to a specific task.

D. Training for FR Loss

Experimentally, we have found that training SISR Networks exclusively on the FR Loss may lead to instability, poor local optima and overfitting that causes color aberrations, artifacts and other non-optimal behavior that seems to help minimize the FR Loss. In order to mitigate this we developed a training procedure that is illustrated in Fig. 1. First, we train a base network to minimize MSE exclusively. On a second phase, we fine-tune the base network to minimize a weighted sum of the FR Loss and MSE. We carefully define the weights so as to each one of the losses contribute approximately 50% of the total loss at the beginning of training. To provide fair comparison we also fine-tune the base network solely on MSE. We have used for network architecture the Super Resolution Residual Network (SRResNet). It is a version of the Residual Network architecture [25] adapted to perform Super Resolution by Ledig et al [3]. We have used 10 residual blocks.

During training we keep the weights for the FR model frozen. However we have found that it is necessary to let the Batch Normalization parameters update during training as not doing so leads to color distortions on the final results. We hypothesize that keeping those parameters frozen leads the network to trying to make the image intensity distribution match the one from the original dataset the FR model was trained on.

E. Evaluating Information Recovery of SISR Networks

Besides using classic metrics for evaluation of our method, we devise a testing procedure that is able to quantify how much identity-defining information the neural network is able to retrieve from the low-resolution image.

For a high-resolution test dataset we produce a degraded version through our degradation model and a super-resolved version through our SISR Networks. We then produce embeddings of these versions of the test dataset using an FR Model,

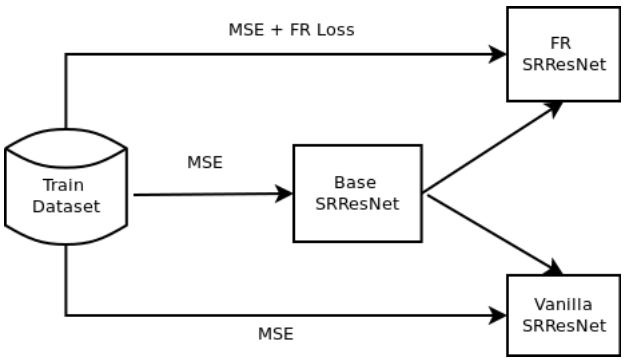

Fig. 1. Training Scheme for our Networks

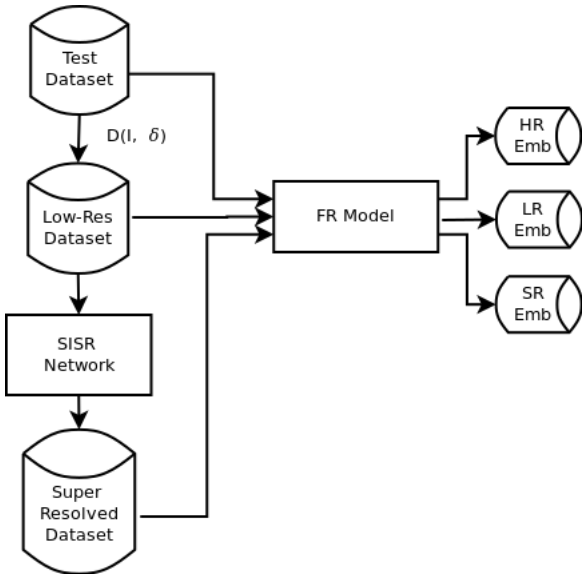

Fig. 2. Testing scheme for a single SISR Network

and these embeddings are then evaluated on classical FR scenarios, which we describe more in depth on subsection IV-B. This procedure is illustrated in Fig. 2.

A natural hypothesis for this procedure is that loss of resolution leads to loss of identity discriminability. The embedding produced by a good FR Model on high-resolution images should separate different identities on different clusters of the embedding space, in a way that allows the embedding to be used for identification and verification effectively.

If this hypothesis is correct, the embedding produced by the low-resolution version of the dataset should have a worse performance when used for the same tasks. Furthermore, the better a super-resolution model is on retrieving high-frequency and identity information, the closer the embedding of the super-resolved test dataset should act as the original high-resolution one.

IV. Experiments

In this section we present the datasets and the experimental design to assess the proposed method.

A. Datasets

1) CelebFaces Attributes Dataset: The CelebFaces Attributes Dataset (CelebA) [26] is the main dataset used in our work. It contains 202,599 face images from 10,177 distinct identities, the number of images per individual identity varies between one and thirty. The dataset is manually annotated to face landmarks and binary characteristics and there is a previous proposed partition into train, validation and test dataset containing strictly non-intersecting identities. We train our SISR Networks exclusively with the training partition of the dataset. For face identification testing and traditional SISR evaluation we select the identities of the test dataset which have exactly thirty image examples.

2) Labelled Faces in the Wild: The Labelled Faces in the Wild (LFW) Dataset [27] is a classical Face Recognition dataset that is comprised of 13,233 images pertaining to 5,749 different identities. It has become famous for providing a series of test protocols for diverse scenarios, some of which have become widely used benchmarks. In our work we follow the "unrestricted with labeled outside data" protocol for testing face verification. Besides the faces themselves, this protocol offers a list of pairs of images of faces available in the dataset alongside with a classification of whether they belong to the same person or not.

3) Datasets used Indirectly: We indirectly take advantage of the VGGFaces2 dataset [28], which is a large scale FR dataset that was used to train the pre-trained FR Model we used in this work.

B. Metrics

In this section we present some metrics used to assess the methods.

1) SISR Metrics: For intrinsic evaluation of the super-resolved images we used the two most common SISR metrics: Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM). PSNR is a metric based on the MSE measured in decibels. For a given MSE, the formula for the PSNR is:

$$\text{PSNR} = 10 * \log_{10}(\frac{\text{MAX}_I^2}{\text{MSE}}) \tag{6}$$

where MAX is the maximum pixel intensity possible for the image.

The higher the measure, the more similar the images are. When the images are equal, PSNR is infinite and when the images are such that the sum pixel by pixel is equal to the maximum value of a pixel can reach, then PSNR is zero. RGB images are generally transformed to the YCbCr format for calculation of the PSNR, which is then done exclusively on the luminance channel [12] [13].

The SSIM metric uses image moments to calculate statistical approximations for the difference in luminance, contrast and structure between two images [29]. The SSIM consists then on a weighted geometric mean of these statistics, and ranges from 0 to 1, where 1 means a perfect match. The SSIM is generally calculated on a series of small windows of both

images that is then averaged. As with the PSNR, SSIM is calculated exclusively on the luminance channel.

The FR Loss function we defined on Section III can also be used as a metric during test time. It is expected that our model trained to minimize the FR Loss will naturally present lower FR Loss scores on test images and this in itself is not a powerful argument for the effectiveness of our method. We still choose to report it to verify generalization to unseen images and to compare how it behaves on different degradation scales.

2) Face Verification Metrics: Face Verification is an FR task to evaluate whether a pair of images belongs to the same person or not. If the output of the FR model is a simple scalar metric, then different thresholds can be used as criteria for determining positive or negative matches. Furthermore, one can plot the relationship between false and true positives over variations in threshold in a Receiver Operating Characteristic (ROC) Curve. This gives a more fine-grained view of the behaviour of our model, since not always the most accurate threshold is the most desirable for most applications (specially those in which the damage of a false negative and false positive greatly differs). Common statistics based on the ROC Curve are the AUC (Area Under Curve) and the EER (Equal Error Rate), the value for which false acceptance and false rejections is equal. [30] [31].

A fixed threshold can also be determined by cross-validation. In this work we use 10-fold cross-validation to determine the best threshold as well as calculate mean accuracy and variance.

On most Deep FR applications [10] [16] the metric used for face verification is a distance metric between faces in the embedding space learned by the model. We use the simple Euclidean distance to produce a vector of distances for each pair and evaluate the embedding using the metrics described above. As discussed in Section III, the performance of the embedding can be used to gauge how much information the SISR Network could retrieve.

3) Face Identification Metrics: Face Identification is an FR task to associate an identity to an individual image based on an available existing group of images of diverse identities. If we have a closed-set of identities that are known to the model beforehand, this task simply reduces to a classification problem. On most real-world applications, though, the set of matching identities are more likely to be open and unknown. To simulate this we adopt a test protocol based on the identification task of the Face Recognition Vendor Test 2002 [9]. For an embedding of a test set with different identities associated to each point, we test a k Nearest Neighbor (kNN) model on leave-one-out cross-validation, which amounts to classifying each point using all the others.

As with face verification, there are more fine-grained metrics to understand a model's performance. We investigate not only if the nearest neighbor belongs to the same class of the data point, but also if the class is present at all on the nearest k points. If so, the probe is said to have a rank k. A graphic called Cumulative Match Characteristic (CMC) shows how many searches have rank k or lower.

C. Pre-processing and Post-processing

All face images are aligned using Multi-task CNN (MTCNN) [32]. For training SISR Networks we convert all pixel values to $[0, 1]$ range. Before passing images through the FR models we do a simple pre-whitening, which normalizes each image by their own mean and standard deviation. In the evaluation we convert the outputs of the network back to the $[0, 256]$ range.

To acommodate the low-resolution images to the FR Model we used, which has a fixed input size of 160x160, we upscale the low-resolution images using bi-cubic interpolation beforehand. As this is an up-scaling method that adds no new information to the image, it does not significantly compromise our hypothesis test that resolution loss implies loss of identity-defining information.

D. Training

We trained SISR Networks to retrieve high-resolution images from the CelebA training partition after degrading them on a 4x and 8x scale. We trained both the base and fine-tuned versions with Adam [33], and a learning rate of 10^{-4} and 10^{-5}, respectively. We compare the results between the network trained solely on MSE ("VanillaSRResNet") and our model ("FRSRResNet").

The Face Recognition Model we used to both calculate the embeddings at test time and to calculate the FR Loss was a pre-trained[1] model, trained on the VGGFace2 Dataset [28] using simple softmax loss.

E. Testing

We conduct standard SISR evaluation on the test set of the CelebA Dataset, reporting average MSE, SSIM and FR Loss between the super-resolved images and the original images.

We build also a face identification test on our test subset of the CelebA dataset to evaluate the performance of our SISR Networks in retrieving identity defining information. To further evaluate this we also perform a face verification test following LFW's "unrestricted with labeled outside data" protocol. This experiment allows us to test whether our proposed method leads our networks to retrieve more information from faces in general or if it is just overfitting to CelebA-style faces in particular.

V. Results and Discussion

In this section we show the results obtained by the tested models and discuss them considering the performance in terms of the retrieval metrics and classification.

A. SISR Evaluation

Table I shows that, considering only classical SISR evaluation metrics and methods, our model performs slightly worse when it is optimized for the FR Loss. The super-resolved images for the FRSRResNet are generally closer than the ground-truth on our embedding space but this should come

[1] Available at: https://github.com/davidsandberg/facenet/

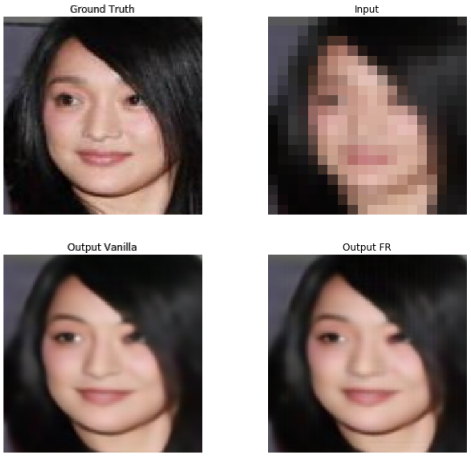

Fig. 3. Comparison of results for an example downsampled in 8x

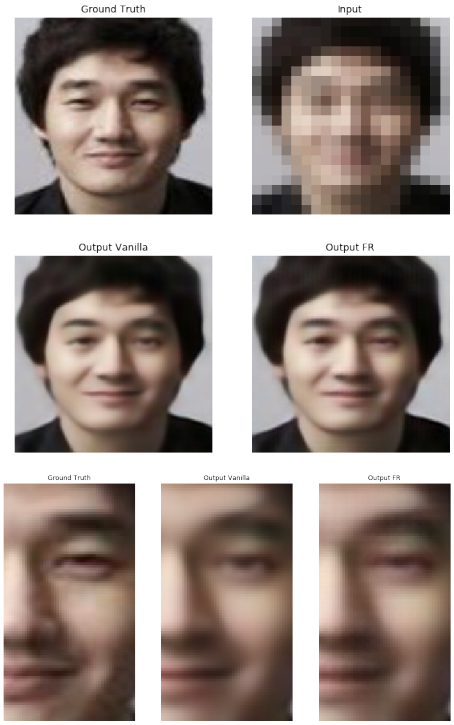

Fig. 4. Comparison of results and specific inset

as no surprise, as this is what the model was trained to do. Otherwise, the images have less PSNR and SSIM. It is very common to models trained on different losses to behave like this while yielding seemingly better-looking images [4] [3] [6]. What usually follows is a qualitative argument, or the use of opinion scores to justify the model.

Figure 3 presents a high-resolution image (top left) and its low-resolution downsizing (top right). It also presents the output of the VanillaSRResNet (botton left) and our FRSRResNet model (botton right). As expected, the VanillaSRResNet presents a more pleasant image than our FRSRResNet model, that preserves better the characterists of the person being

Fig. 5. Comparison of results and specific inset

	PSNR (dB)	SSIM	FR Loss
VanillaSRResNet (8x)	**27.49 +- 2.07**	**0.875 +- 0.04**	8.83 +- 1.53
FRSRResNet (8x)	27.30 +- 2.03	0.870 +- 0.04	**8.37 +- 1.49**
VanillaSRResNet (4x)	**32.82** +- 2.56	**0.956** +- 0.02	4.16 +- 0.95
FRSRResNet (4x)	32.57 +- 2.49	0.953 +- 0.02	**3.96 +- 0.91**

TABLE I
RESULTS FOR INTRINSIC SISR EVALUATION. BEST RESULTS FOR EACH SCALE BOLDED

imaged. Figure 4 presents another result of the same methods (in the same relative positions) and an inset where we can see that our model does a better job than the state-of-the-art at recovering characteristics associated with an Asian face structure, such as epicanthic folds on the eyes. Indeed this is an advantage that can be seen on numerous other examples omitted for the sake of brevity.

The advantages observed are not limited to geographical characteristics, though. We also call attention to the reconstruction of face contours in both previous examples and specifically on Figure 5 (using the same location pattern) and an inset showing a more accurate reconstruction of mouth and nose contours. Finally, one can notice some checkerboard-like artifacts that appear in the images generated by our method. This is something that was also reported by Johnson et al. [4] for their models trained with the Perceptual Loss, and is assumed to be the cause of the degradation of PSNR/SSIM performance.

B. Evaluation on Face Resolution Tasks

Beyond the qualitative argument, our evaluation procedure allows us to make quantitative arguments about the usefulness of our model despite the loss of performance on traditional

	1NN	5NN	10NN
High Resolution	0.9714	0.9736	0.9729
FRSRResNet (4x)	**0.9542**	**0.9595**	**0.9587**
VanillaSRResNet (4x)	0.9523	0.9569	0.9574
Low-Resolution (4x)	0.9106	0.9227	0.9212
FRSRresNet (8x)	**0.8087**	**0.8333**	**0.8346**
VanillaSRResNet (8x)	0.7779	0.8058	0.8157
Low-Resolution (8x)	0.4194	0.4381	0.4496

TABLE II

KNN RESULTS FOR EMBEDDING EVALUATION ON FACE IDENTIFICATION TASK. BEST RESULTS FOR EACH SCALE ARE BOLDED

Model	Accuracy	AUC	Equal Error Rate
High Resolution	0.988 +- 0.005	0.999	0.012
FRSRResNet (4x)	0.980 +- 0.005	**0.997**	**0.018**
VanillaSRResNet (4x)	**0.980 +- 0.003**	**0.997**	**0.018**
Low Resolution (4x)	0.969 +- 0.004	0.995	0.029
FRSRResNet (8x)	**0.934 +- 0.008**	**0.981**	**0.065**
VanillaSRResNet (8x)	0.922 +- 0.016	0.976	0.079
Low Resolution	0.826 +- 0.015	0.906	0.174

TABLE III

ACCURACY, AUC AND EER FOR EMBEDDING EVALUATION ON FACE VERIFICATION TASK. BEST RESULTS FOR EACH SCALE ARE BOLDED

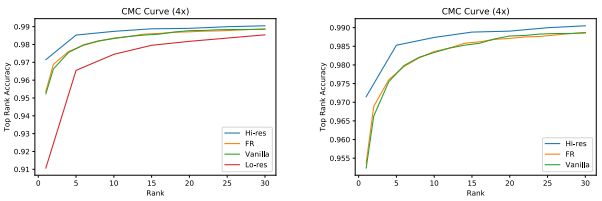

Fig. 6. Cumulative Match Characteristic for Face Identification Task at 4x Scale. In the right plot we remove the low-resolution baseline for increased detail

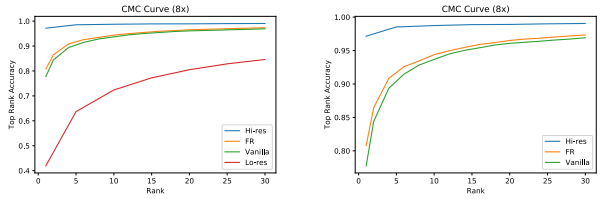

Fig. 7. Cumulative Match Characteristic for Face Identification Task at 8x Scale. In the right plot we remove the low-resolution baseline for increased detail

metrics. Observing Tables II and III we can see that our hypothesis is correct and indeed the loss of resolution hinders the embedding's performance on Face Resolution tasks. This effect is more visible on higher scales of degradation, though. In the case of the LFW face verification task, which seems to be all-around easier, the loss of performance in 4x scale is so little the results are not conclusive.

Our method of training gives better results in both tasks on all metrics reported. The amount of improvement seems to be related to the scale as well. While there are decisive improvements in 8x scale, these improvements are more timid on 4x scale. This may be indicative of the kind and scale of the information the FR Model uses to determine proximity. The CMC plot for the 8x scale presented on Fig. 7 also shows that our method is consistently better than the traditional MSE, and not only on average (such as AUC, Accuracy) or on special cases of hyperparameter selection (Accuracy of selected cases of k-Nearest-Neighbors).

There seems to be an overall correlation between lower FR Loss scores and higher scores on FR related tasks, as expected. This relation seems to be non-linear, with decreasing marginal gains. We can observe that the improvements yielded by our method in terms of FR Loss is relatively the same in 4x and 8x scale but these do not translate in gains of the same magnitude on other evaluated FR tasks.

C. Limitations and Future Work

Our evaluation method proved to be useful to quantify information recovery beyond both classical metrics and opinion scores. It could be applied to a variety of methods in the state-of-the-at that produce qualitatively and subjectively better results but lack quantitative justification that is not based on PSNR/SSIM. A more varied sample of different SISR Networks with distinct PSNR/SSIM results could also be studied on how much these metrics correlate with our proposed task-based ones, and ultimately whether one can be used as a proxy for another. A more in-depth comparison of our training method to the state-of-the-art could bring more credibility to it as well.

Our training method has shown significant improvements upon the standard training procedure for a common SISR Network in Face Recognition tasks. As our method consists of a loss function and a method to optimize it, it could be directly applied to a wider range of network architectures. This would be useful to investigate the relationship between a network's representational power and how much our method can improve its performance. It could be the case that networks with more representational power can improve more, as they learn to represent identity-defining characteristics, or it could be the case that they improve less, as they are able to learn these without our method.

Likewise, the use of different FR models could bring more evidence for the quality of our method or even information about which characteristics determinate FR Models take more into consideration. Defining a "training FR Model" exclusively for the FR Loss and a "test FR Model" exclusive for the task-based evaluation could also bring light to whether our method learns identity characteristics in abstract or only the specific characteristics used by a certain FR model.

The way the FR Loss was constructed can also be improved. We have defined the loss as the distance between the original and reconstructed image on the FR embedding. However, as we have seen, this distance is more informative on greater scales of resolution loss. This may not be the case if we use the distance between the reconstructed image and different images belonging to the same identity instead. Iteratively minimizing the distance between the reconstructed image and a random picture from the same identity or the centroid of all identities of the same person could then be a more effective optimization

strategy for greater improvements even in lesser degrees of degradation.

VI. Conclusions

In this work we have built an evaluation framework that can give more fine-grained information about a super-resolution model's performance and behavior and successfully applied it to argue in favor of a training method inspired by the same framework. Further investigation about our training method is necessary, while our testing framework can already be easily applied for other models.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001 and by FAPESP 15/22308-2, 15/01587-0.

References

[1] P. Hanhart, P. Korshunov, and T. Ebrahimi, "Benchmarking of quality metrics on ultra-high definition video sequences," in *2013 18th International Conference on Digital Signal Processing (DSP)*. IEEE, 2013, pp. 1–8.

[2] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of psnr in image/video quality assessment," *Electronics letters*, vol. 44, no. 13, pp. 800–801, 2008.

[3] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681–4690.

[4] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711.

[5] ITU-T, "Mean opinion score interpretation and reporting," *ITU Recommendation P.800.2*, 2013.

[6] Y. Chen, Y. Tai, X. Liu, C. Shen, and J. Yang, "Fsrnet: End-to-end learning face super-resolution with facial priors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 2492–2501.

[7] P. Rasti, T. Uiboupin, S. Escalera, and G. Anbarjafari, "Convolutional neural network super resolution for face recognition in surveillance monitoring," in *International conference on articulated motion and deformable objects*. Springer, 2016, pp. 175–184.

[8] P. Li, L. Prieto, D. Mery, and P. J. Flynn, "On low-resolution face recognition in the wild: Comparisons and new techniques," *IEEE Transactions on Information Forensics and Security*, 2019.

[9] P. J. Phillips, P. Grother, R. Micheals, D. M. Blackburn, E. Tabassi, and M. Bone, "Face recognition vendor test 2002," in *2003 IEEE International SOI Conference. Proceedings (Cat. No. 03CH37443)*. IEEE, 2003, p. 44.

[10] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1701–1708.

[11] O. M. Parkhi, A. Vedaldi, A. Zisserman *et al.*, "Deep face recognition." in *bmvc*, vol. 1, no. 3, 2015, p. 6.

[12] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *European conference on computer vision*. Springer, 2014, pp. 184–199.

[13] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1874–1883.

[14] Z. Wang, J. Chen, and S. C. H. Hoi, "Deep learning for image super-resolution: A survey," *CoRR*, vol. abs/1902.06068, 2019. [Online]. Available: http://arxiv.org/abs/1902.06068

[15] H. Huang, R. He, Z. Sun, and T. Tan, "Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1689–1697.

[16] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 212–220.

[17] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[18] T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," *CoRR*, vol. abs/1606.03498, 2016. [Online]. Available: http://arxiv.org/abs/1606.03498

[19] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems*, 2017, pp. 6626–6637.

[20] D. Dai, Y. Wang, Y. Chen, and L. Van Gool, "Is image super-resolution helpful for other vision tasks?" in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–9.

[21] S. Hu, R. Maschal, S. S. Young, T. H. Hong, and P. J. Phillips, "Face recognition performance with superresolution," *Applied optics*, vol. 51, no. 18, pp. 4250–4259, 2012.

[22] A. Bulat and G. Tzimiropoulos, "Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 109–117.

[23] M. Haris, G. Shakhnarovich, and N. Ukita, "Task-driven super resolution: Object detection in low-resolution images," *arXiv preprint arXiv:1803.11316*, 2018.

[24] K. Zhang, Z. Zhang, C.-W. Cheng, W. H. Hsu, Y. Qiao, W. Liu, and T. Zhang, "Super-identity convolutional neural network for face hallucination," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 183–198.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[26] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

[27] 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," University of Massachusetts, Amherst, Tech. Rep. 07-49, October 2007.

[28] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 67–74.

[29] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli *et al.*, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[30] J.-M. Cheng and H.-C. Wang, "A method of estimating the equal error rate for automatic speaker verification," in *2004 International Symposium on Chinese Spoken Language Processing*. IEEE, 2004, pp. 285–288.

[31] T. Fawcett, "An introduction to roc analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[32] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.