Frequency Band Encoding for Face Super-Resolution

Klemen Grm[1], **Vitomir Štruc**[1]

[1]*University of Ljubljana, Faculty of Electrical Engineering*
E-mail: klemen.grm@fe.uni-lj.si

Abstract

In this paper, we present a novel method for face super-resolution based on an encoder-decoder architecture. Unlike previous approaches, which focused primarily on directly reconstructing the high-resolution face appearance from low-resolution images, our method relies on a multi-stage approach where we learn a face representation in different frequency bands, followed by decoding the representation into a high-resolution image. Using quantitative experiments, we are able to demonstrate that this approach results in better face image reconstruction, as well as aiding in downstream semantic tasks such as face recognition and face verification.

1 Introduction

Face super-resolution is a subset of the general single-image super-resolution problem. By itself, single-image super-resolution is the inverse problem of recovering a high-resolution image given a low-resolution (e.g., sub-sampled) version of it. The single-image super-resolution problem is highly ill-posed, since a single low-resolution image can correspond to many different high-resolution images - for a given sub-sampling factor f, the ratio of information from the high-resolution image retained in the low resolution image is at most $\frac{1}{f^2}$ given perfect sampling and anti-aliasing strategies. Unlike the general single-image super-resolution problem, however, its domain specific subsets such as face super-resolution allow us to use the constraints on the high-resolution image appearance such as face identity information [2] or face component segmentation [8] to constrain the solution space and achieve better high-resolution reconstruction at higher magnification ratios than are typically considered in general single-image super-resolution applications.

Existing work on super-resolution methods based on machine learning techniques (e.g., [3, 12, 6, 7, 13]) typically consists of generating pairs of aligned low-resolution and high-resolution images $\mathbf{x}$ and $\mathbf{y}$, respectively, by starting with a high-resolution image dataset and downsampling the images through a process

$$\mathbf{x} = H\mathbf{y} \downarrow_d + N, \tag{1}$$

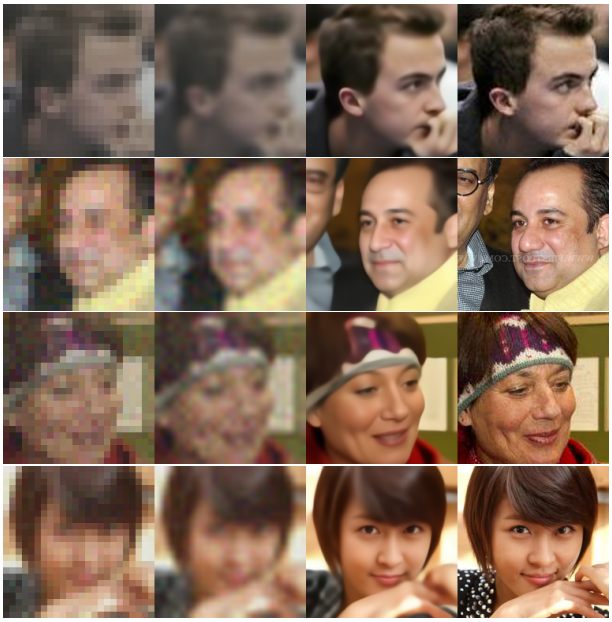
where H is an image filtering operator (e.g., the Gaussian or Lanczos filters), $\downarrow$ is the sub-sampling operator, d is the sub-sampling factor, and N is a noise component. Given the generated dataset of $(\mathbf{x}, \mathbf{y})$ pairs, a differentiable model m_θ with free parameters θ is then trained to approximate the inverse process, i.e., to predict an approximation of the high resolution image $\hat{\mathbf{y}}$ given a low-resolution image $\mathbf{x}$ as an input,

$$\hat{\mathbf{y}} = m_\theta(\mathbf{x}), \tag{2}$$

by setting the model parameters, θ through gradient descent on the loss function, typically a distortion measure between the model predictions and the ground-truth high-resolution images, e.g.,

$$\mathfrak{L}(\mathbf{x}, \mathbf{y}) = \|m_\theta(\mathbf{x}) - \mathbf{y}\|_p^p, \tag{3}$$

where p is the order of the norm used to measure the distortion between the model approximations and expected (i.e., groundtruth) high-resolution images.



(a) *(b)* *(c)* *(d)*

Figure 1: Output samples of the proposed face super-resolution method. The figure depicts the input low-resolution images *(a)*, interpolated to the output resolution using bicubic interpolation *(b)*, compared against the results of the proposed method *(c)* and the high-resolution groundtruth *(d)*.

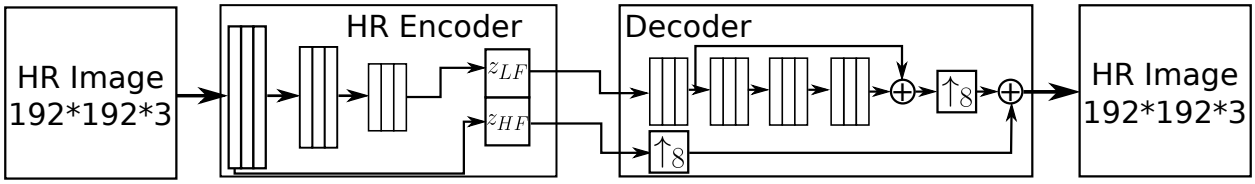


Figure 2: The architecture of the HR encoder and universal decoder segments of our model. The HR encoder produces the high- and low-frequency components of the face representation, where the high-frequency component is derived from the earlier layers of the network with smaller receptive fields, whereas the low-frequency component of the representation is derived from the final output of the convolutional backbone.

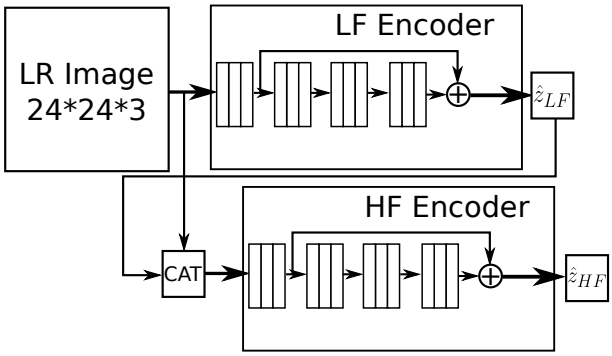


Figure 3: The architecture of the low-resolution encoder and segment of our model. The LR encoder produces the high- and low-frequency components of the face representation from an input LR image by using two sub-models in a cascaded residual setup.

Unlike previous approaches, our proposed method separates the training of the upsampling model, m_θ, into two steps, namely, *i)* learning an appropriate representation of face images for high-resolution reconstruction, and *ii)* performing the high-resolution reconstruction given the representation. This approach has several advantages over the end-to-end approach. Firstly, the representation can be derived only using high-resolution images and autoencoder training, without involving a synthetic downsampling process such as (1). Secondly, our chosen encoder architecture explicitly splits the representation of high- and low-frequency bands of the input image, allowing both to be retained in the representation, which in turn allows the decoder to learn sharper reconstructions. Thirdly, given a trained decoder that produces high-resolution images given a face image representation, we could then in principle train different encoders for images of different resolution or quality levels, as opposed to a universal model, allowing more fine-tuned results without needing different reconstruction models.

To summarize, the key contributions of this paper are

1. We propose a novel, encoder-decoder based architecture for the face super-resolution problem,
2. We present a training process that allows our model to learn face image representations with the explicit separation of information content from the high- and low- frequency bands of the input image,
3. We evaluate the proposed method on the standard image reconstruction task using distortion metrics, as well as on downstream semantic tasks, to demonstrate the effectiveness of our approach.

2 Methodology

2.1 Datasets

We use the VGGFace2 [1] dataset to train our proposed method. We use the dataset because of its relatively large size (3.31 million images of 9131 subjects) compared to datasets commonly used for face super-resolution training such as Casia WebFace [15] and CelebFacesA [9], as well as higher-quality images overall in terms of sharpness. A large dataset of diverse, high-quality images is needed to best make use of large model capacity. We use the VGGFace2 test set (with non-overlapping subjects with regards to the training set) to evaluate the reconstruction capability of our model. Furthermore, we also use the LFW dataset [5] to evaluate the utility of our proposed method to downstream semantic tasks, i.e., to face recognition.

In order to downsample the high-resolution images (to generate the corresponding low-resolution inputs), we perform the following steps:

1. Resize the image to $256px$ along the short side, maintaining the aspect ratio of the original
2. Extract a random $192px$ square crop from the image
3. Flip the image horizontally with probability 0.5 (at this step, the image represents the HR part of the (x, y) pair)
4. Perform Gaussian filtering on the HR image, with $\sigma \sim \mathcal{U}(3, 4)$ and kernel size $k = \lceil 4\sigma \rceil + 1$
5. Downsample the filtered image by a factor of 8 using bicubic resampling
6. Desaturate the image contrast by a random factor $f \sim \mathcal{U}(0.5, 1)$
7. Add Gaussian white noise with $\sigma \sim \mathcal{U}(0, \frac{10}{255})$ to each color channel of the low-resolution image.

2.2 Model architecture

The overall architecture of the proposed model is illustrated in Figures 2 and 3. The model is comprised of three main components, namely,

The **HR encoder**, which produces a band-separated face representation $z_{LF}||z_{HF}$ given a high-resolution input face image. The CNN backbone is based on the VGGFace [11] architecture, which has proven to be more useful for perceptual tasks and style transfer than newer image recognition models based on deep residual learning [10]. The high-frequency band of the representation is extracted from the earlier layers of the convolutional backbone, where the network has a limited receptive field and therefore lacks the capability to learn the low-frequency global structure of the image. In contrast, the low-frequency band of the face representation is derived from the final output of the convolutional backbone.

The **Universal decoder**, which is trained to reconstruct a high-resolution image given the corresponding face representation. In order to avoid having to encode spatial information into latent vectors and the decoder having to interpret the spatial information, we opt to retain the spatial relations of the entries in the face representation by keeping it in its tensor form, i.e., given an input image $I_{HR} \in \mathbb{R}^{192\times192\times3}$, it is the case for the face representations that $z_{HF} \in \mathbb{R}^{24\times24\times8}$ and $z_{LF} \in \mathbb{R}^{24\times24\times8}$. We decode the face representation into a high-resolution image using a modified EDSR [7] network where the first layer is changed to accept the 8-channel low-frequency band of the representation, whereas the high-frequency band of the representation is passed directly to a subpixel convolution [12] upsampling module, whose output is added to the final result of the decoder.

The **LR encoder**, which is trained to produce a face representation $\hat{z}_{LF}||\hat{z}_{HF}$ given an input low-resolution image. The encoder consists of two sub-models, which produce the respective frequency bands of the encoding. Both share the same architecture, i.e., an EDSR-like [7] convolutional backbone, without the upsampling module, since the spatial resolution of the face representation is already equal to the spatial resolution of the input low-resolution images. The low-frequency band of the approximated representation is derived directly from the LF encoder. Then, $\hat{z}_{LF}$ is concatenated to the input low-resolution image along the channel axis and used as an input to the HF encoder. This allows the HF encoder to use information about which parts of the image are already encoded in the low-frequency band of the representation and focus on refining the output image. The final approximation of the representation, $\hat{z}_{LF}||\hat{z}_{HF}$, can then be used as an input to the universal decoder.

2.3 Training procedure

We begin training our model with autoencoder training of the HR encoder and universal decoder, as illustrated in Figure 2. This combined model is trained to reconstruct high-resolution images presented to the HR encoder as input. We begin by only training the low-frequency band of the representation. Once the training converges, the high-frequency band path is activated to further refine the results. In Figure 4, we show an image reconstructed using the HR encoder and universal decoder. Using this architecture, we are able to achieve near-perfect reconstructions of high-resolution images ($SSIM > 0.99$),

(a) *(b)* *(c)* *(d)*

Figure 4: A reconstructed high-resolution image using our HR encoder and universal decoder. The figure depicts *(a)* the output of the decoder given only the low-frequency band of the face representation, *(b)*, the full output of the decoder, *(c)* the groundtruth high-resolution image, and *(d)* the difference between the outputs from low-frequency only and full representations.

which means the decoder is well-suited for the face super-resolution tasks. We use the L_1 loss to train the HR encoder and universal decoder with gradient descent, i.e.,

$$\begin{aligned}\mathfrak{L}(\mathbf{y},\hat{\mathbf{y}}) &= \mathbb{E}_{\mathbf{y}\sim\mathcal{T}}\left|\mathbf{y}-\hat{\mathbf{y}}\right| \\ &= \mathbb{E}_{\mathbf{y}\sim\mathcal{T}}\left|\mathbf{y}-m_{DEC}\left(m_{ENC_{HR}}\left(\mathbf{y}\right)\right)\right|,\end{aligned} \tag{4}$$

where the image $\mathbf{y}$ is considered as a sample of the training set $\mathcal{T}$.

Once the universal decoder is initialized through high-resolution autoencoder training, we train the low-resolution encoder model to produce approximations of the face representation from a low-resolution image, $\hat{z}_{LF}||\hat{z}_{HF}$. As above, the representations are passed to the universal decoder to produce a reconstruction of the high-resolution image. To train the low-resolution encoder, we use a combination of L_1 loss per-pixel on the reconstructed image, and a cycle consistency loss by passing the reconstructed image back to the high resolution encoder. The full loss is

$$\begin{aligned}\mathfrak{L}(\mathbf{x},\mathbf{y}) &= \lambda_1\left|\mathbf{y}-\hat{\mathbf{y}}\right| + \lambda_2\,\mathfrak{L}_{CYC}(\mathbf{x},\mathbf{y}) \\ &= \lambda_1\left|\mathbf{y}-m_{DEC}(m_{ENC_{LR}}(\mathbf{x}))\right| \\ &+ \lambda_2\,\mathfrak{L}_{CYC}(\mathbf{x},\mathbf{y}),\end{aligned} \tag{5}$$

where $\mathfrak{L}_{CYC}$ is the cycle consistency loss, i.e.,

$$\begin{aligned}\mathfrak{L}_{CYC}(\mathbf{x},\mathbf{y}) &= \left|m_{ENC_{HR}}(\mathbf{y}) - m_{ENC_{HR}}(\hat{\mathbf{y}})\right| \\ &= \left|m_{ENC_{HR}}(\mathbf{y})\right. \\ &\left.- m_{ENC_{HR}}(m_{DEC}(m_{ENC_{LR}}(\mathbf{x})))\right|,\end{aligned} \tag{6}$$

and λ_1 and λ_2 are weights of the pixel loss and the cycle consistency loss, respectively. Using a logarithmic grid search, we experimentally set $\lambda_1 = 10^{-3}$, $\lambda_2 = 1$. The model is trained using the Adam [4] gradient descent optimization method, using a learning rate of 10^{-4} and batch size of 64, until convergence.

3 Results

We evaluate the high-resolution face reconstruction capabilities of our model on the VGG2 test set, which contains 169 396 images of 500 subjects, disjoint from the

Method	SSIM
Bicubic interpolation	0.5921
EDSR [7]	0.6487
C-SRIP [2]	0.6991
FBE-FSR (proposed)	0.7531

Table 1: Image restoration results on the VGG2 test set. SSIM results closer to 1 indicate better face super-resolution performance.

Method	Verification accuracy ($\mu \pm \sigma$)
Bicubic interpolation	0.8417 ± 0.0101
EDSR [7]	0.9137 ± 0.0079
C-SRIP [2]	0.9341 ± 0.0052
FBE-FSR (proposed)	0.9465 ± 0.0059
Original HR Images	0.9936 ± 0.0042

Table 2: Face verification results on the LFW dataset. Results on the original high-resolution images are included for comparison against restored low-resolution images.

training set. We generate LR images using the same process outlined in section 2.1. We measure the distortion between the groundtruth high-resolution images and the outputs of our model using the Structural Similarity Index (SSIM [14]) metric, the literature standard for evaluating reference-based image restoration algorithms. We compare the restoration results against competing algorithms in Table 1. We note that in comparison to the C-SRIP [2] model, the proposed method is able to perform more convincing face image restoration on edge cases such as extreme pose (profile images), face occlusion, and face scale variation. We show some qualitative comparisons of the model output in the Figure 5.

We also evaluate the proposed method in terms of its utility to downstream semantic computer vision tasks, i.e., face recognition. We use the LFW [5] dataset for this experiment, because it is considered a "solved" dataset in the sense that modern face recognition models achieve near 100% verification accuracy, and we are mainly interested in the amount of performance degradation on downsampled and super-resolved images. We downsample the images using a similar algorithm as outlined above, except the $192px$ crop is always central, since the face images are already aligned. Next, we perform image restoration on the generated LR images, and perform the standard LFW face verification experiment using a pretrained VGG2-SENet-101[1] pre-trained face feature extraction model. The verification accuracy of the considered methods (in terms of mean and standard deviation over the 10-fold evaluation protocol) are compared in Table 2.

4 Conclusion

We have presented FBE-FSR, a novel method for face image super-resolution based on frequency component separation in the latent-space of an encoder-decoder architecture. We have shown that our proposed method is capable of outperforming our previous work and competing methods in terms of image restoration capability and utility for downstream vision tasks.

(a) *(b)* *(c)* *(d)* *(e)*

Figure 5: A qualitative comparison of our proposed method with the next best tested method. The figure depicts LR input images *(a)*, results of bicubic interpolation *(b)*, C-SRIP *(c)*, the proposed method *(d)*, and the HR image *(e)*.

In terms of future work, we plan on extending the method by adapting additional encoder methods for specific image degradation pipelines, and extending the training framework with generative adversarial objectives.

References

[1] 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)*, pages 67–74. IEEE, 2018.

[2] K. Grm, W. J. Scheirer, and V. Štruc. Face hallucination using cascaded super-resolution and identity priors. *IEEE Transactions on Image Processing*, 29:2150–2165, 2019.

[3] J. Kim, J. K. Lee, and K. M. Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016.

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

[5] G. Learned-Miller. Labeled faces in the wild: Updates and new reporting procedures. *University of Massachusetts, Amherst, Tech. Rep. UM-CS-2014-003*, 2014.

[6] 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*, pages 4681–4690, 2017.

[7] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017.

[8] L. Liu, S. Wang, and L. Wan. Component semantic prior guided generative adversarial network for face super-resolution. *IEEE Access*, 7:77027–77036, 2019.

[9] Z. Liu, P. Luo, X. Wang, and X. Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018):11, 2018.

[10] A. Mordvintsev, N. Pezzotti, L. Schubert, and C. Olah. Differentiable image parameterizations. *Distill*, 2018. https://distill.pub/2018/differentiable-parameterizations.

[11] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. 2015.

[12] 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*, pages 1874–1883, 2016.

[13] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018.

[14] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

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