Face Hallucination with Tiny Unaligned Images by Transformative Discriminative Neural Networks

Xin Yu, Fatih Porikli
Australian National University
{xin.yu, fatih.porikli}@anu.edu.au

Abstract

Conventional face hallucination methods rely heavily on accurate alignment of low-resolution (LR) faces before upsampling them. Misalignment often leads to deficient results and unnatural artifacts for large upscaling factors. However, due to the diverse range of poses and different facial expressions, aligning an LR input image, in particular when it is tiny, is severely difficult. To overcome this challenge, here we present an end-to-end transformative discriminative neural network (TDN) devised for super-resolving unaligned and very small face images with an extreme upscaling factor of 8. Our method employs an upsampling network where we embed spatial transformation layers to allow local receptive fields to line-up with similar spatial supports. Furthermore, we incorporate a class-specific loss in our objective through a successive discriminative network to improve the alignment and upsampling performance with semantic information. Extensive experiments on large face datasets show that the proposed method significantly outperforms the state-of-the-art.

Introduction

Face images provide vital information for visual perception and identity analysis. Nonetheless, when the resolution of the face image is very small (*e.g.* in typical surveillance videos), there is little information that can be inferred from it. Very low-resolution (LR) face images not only degrade the performance of the recognition systems but also impede human interpretation. This challenge motivates the reconstruction of high-resolution (HR) images from given LR counterparts, known as face hallucination, and attracts increasing interest in recent years.

Previously proposed face hallucination methods based on holistic appearance models (Liu, Shum, and Zhang 2001; Baker and Kanade 2002; Wang and Tang 2005; Liu, Shum, and Freeman 2007; Hennings-Yeomans, Baker, and Kumar 2008; Ma, Zhang, and Qi 2010; Yang et al. 2010; Li et al. 2014; Arandjelović 2014; Kolouri and Rohde 2015) demand LR faces to be precisely aligned beforehand. However, aligning LR faces to appearance models is not a straightforward task itself, and more often, it requires expert feedback when the input image is small. Pose and expression variations that naturally exist in LR face images hinder the accuracy of automatic alignment techniques, which usually assume facial landmarks are visible and detectable. As a result, the performance of face hallucination degrades severely. Such a broad spectrum of pose and expression variations also makes learning a comprehensive appearance model even harder. For instance, Principal Component Analysis (PCA) based schemes become critically ineffective to learn a reliable face model while aiming to capture different in- and out-plane rotations, scale changes, translational shifts, and facial expressions. As a result, these methods lead to unavoidable artifacts when LR faces are misaligned or depict different poses and facial expressions from the base appearance model.

Rather than learning holistic appearance models, many methods upsample facial components by transferring references from an HR training dataset and then blending them into an HR version (Tappen and Liu 2012; Yang, Liu, and Yang 2013; Zhou and Fan 2015). These methods expect the resolution of input faces to be sufficient enough for detecting the facial landmarks and parts. When the resolution is very low, they fail to localize the components accurately, thus producing non-realistic faces. In other words, the facial component based methods are unsuitable to upsample very LR faces.

In this paper, we present a new transformative discriminative neural network (TDN) to overcome the above issues and achieve super-resolving a tiny (*i.e.*16×16 pixels) and unaligned face image by a remarkable upscaling factor 8, where we reconstruct 64 pixels for each single pixel of the input LR image.

Our network consists of two components: an upsampling network that comprises deconvolutional and spatial transformation network (Jaderberg et al. 2015) layers, and a discriminative network. The upsampling network is designed to progressively improve the resolution of the latent feature maps at each deconvolutional layer. We do not assume the LR face is aligned in advance. Instead, we compensate for any misalignment and changes through the spatial transformation network layers that are embedded into the upsampling network. One can use the pixel-wise intensity similarity between the estimated and the ground-truth HR face images as the objective function in the training stage. However, when the upscaling factor becomes larger, employing only the pixel-wise intensity similarity causes over-smoothed outputs. Therefore, we incorporate class similarity information that is provided by a discriminative network to enforce the upsampled HR faces to be similar to real face images. We back-propagate the discriminative errors to the upsampling network. Our end-to-end solution allows fusing the pixel-wise and class-wise information in a manner robust to spatial transformations and obtaining a super-resolved output with much richer details.

Overall, our main contributions have four aspects:

- We present a novel end-to-end transformative discriminative network (TDN) to super-resolve very low-resolution (16×16 pixels) face images with an upscaling factor 8×.
- For tiny input images where landmark based methods inherently fail, our method is the first solution to hallucinate an unaligned LR face image without requiring precise alignment in advance, which makes our method practical.
- Fusion of pixel-wise appearance similarity and class-wise discriminative information allows the super-resolution process to take full advantage of class-specific cues for the alignment and detail enhancement tasks.
- Our method achieves almost 4 dB PSNR improvement over the state-of-the-art.

Related Work

Face hallucination aims to magnify an LR image to its HR version, which contains extra high-frequency details. State-of-the-art face hallucination methods can be grouped into two categories: appearance based methods and facial components based methods.

Appearance based methods employ PCA to build a holistic face model or apply reference HR patches to reconstruct the HR counterparts of the LR patches. Baker and Kanade (2002) construct high-frequency details of aligned frontal face images by searching the best mapping between LR and HR patches from the training dataset. Wang and Tang (2005) develop an eigen-tranformation to super-resolve face images by establishing a linear mapping between LR and HR face subspaces. Liu, Shum, and Freeman (2007) employ a PCA based global appearance model to upsample LR faces and a local non-parametric model to enhance the facial details. Kolouri and Rohde (2015) explore optimal transport and subspace learning to morph an HR output. Ma, Zhang, and Qi (2010) hallucinate an LR face image with position patches sampled from multiple aligned HR images, while Li et al. (2014) model the local face patches as a sparse coding problem. Since appearance based face hallucination methods require that the LR images are precisely aligned and have the same pose and expression as the HR references, these methods are sensitive to the misalignment of LR images. When misalignment or different poses and expressions exist, their performance may degrade dramatically.

Facial components based methods super-resolve facial parts rather than entire faces, and thus they can address various poses and expressions. Tappen and Liu (2012) use SIFT flow (Liu, Yuen, and Torralba 2011) to align LR images, and then restore the details of LR images by deforming the reference HR images. Yang, Liu, and Yang (2013) first detect facial components in the LR images and then transfer the most similar HR facial components in the dataset to the LR input. Since the facial components based methods require to extract facial components from LR inputs, the resolution of the input LR images cannot be very low. Otherwise, these methods may fail to localize facial components, thus generating non-realistic HR results.

Recently, convolutional neural network (CNN) based methods have been proposed and claimed the state-of-the-art performance (Dong, Loy, and He 2016; Kim, Lee, and Lee 2015; Wang et al. 2015; Bruna, Sprechmann, and LeCun 2016). Because these methods are designed to upsample generic patches and do not fully exploit class-specific information, they are not suitable to hallucinate tiny faces. Zhou and Fan (2015) present a bi-channel CNN to hallucinate blurry face images. They first use CNN to extract facial features and then feed the features to fully connected layers to generate high-frequency facial details. This method is restricted to the input image size as the other facial component based approaches.

Proposed Method: TDN

Our transformative discriminative neural network achieves the image alignment and super-resolution simultaneously. The entire processing pipeline is shown in Fig. 1.

Network Architecture

The transformative discriminative neural network consists of two parts: an upsampling network that combines spatial transformation network layers and deconvolutional layers, and a discriminative network.

Upsampling Network The parameters of our upsampling network are shown in Fig. 1 (red frame).

Deconvolutional Layers: The deconvolutional layer, also known back-convolutional layer, can be made of a cascade of an upsampling layer and a convolutional layer, or a convolutional layer with a fractional stride. Therefore, the resolution of the output of the deconvolutional layers is larger than the resolution of its input. We employ the ℓ_2 regression loss, also known as Euclidean distance loss, to constrain the similarity between the hallucinated HR faces and their original HR ground-truth versions. We notice that previous works also employ similar deconvolutional layers to upsample natural scenes (Long, Shelhamer, and Darrell 2015; Fischer et al. 2015). However, they only apply to generic images without exploiting any class-specific cues. Thus, their results tend to be smooth. In contrast, we train the network with face images and let it learn and memorize the facial parts for hallucination.

Spatial Transformation Layers: The spatial transformation network (STN) is recently proposed by Jaderberg et al. (2015). It can estimate the motion parameters of images, and warp images to the canonical view. In our architecture, the spatial transformation network layers are represented as the green boxes in Fig. 1. These layers contain three modules: a localization module, a grid generator module, and a sampler. The localization module consists of a number of hidden layers and outputs the transformation parameters of an input

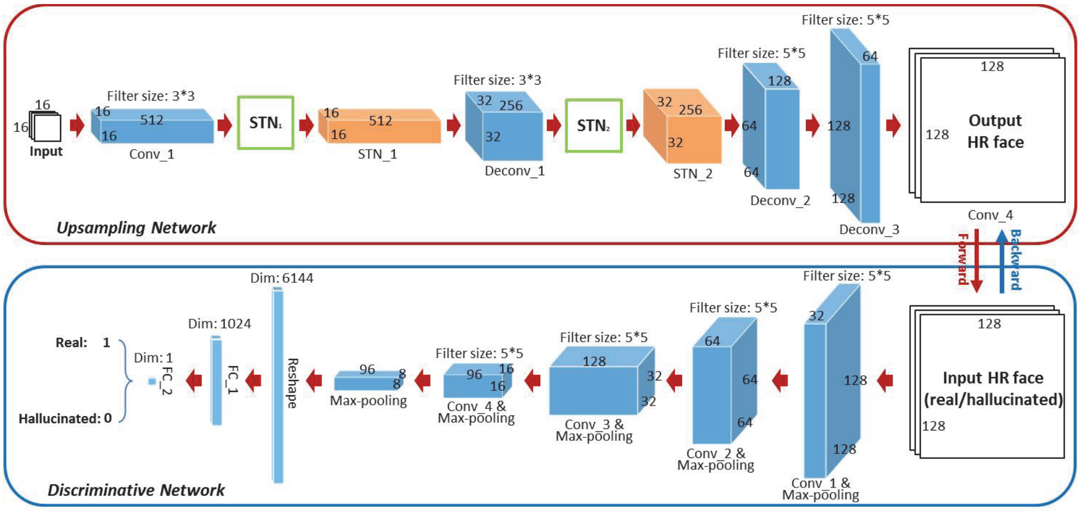

Figure 1: Our TDN consists of two parts: an upsampling network (in the red frame) and a discriminative network (in the blue frame).

relative to the canonical view. The grid generator module creates a sampling grid according to the estimated parameters. Finally, the sampler module maps the input onto the generated grid by bilinear interpolation.

Since we focus on in-plane rotations, translations, and scale changes without requiring a 3D face model, we employ the similarity transformation for face alignment. Although the STN can warp images, it is not straightforward to use them directly to align very LR face images. There are several factors needed to be considered: (i) After the alignment of LR images, facial patterns are blurred due to the resampling of the aligned faces by bilinear interpolation. (ii) Since the resolution is very low and a wide range of poses exists, spatial transformations lead to alignment errors. (iii) Due to the blur and alignment errors, the upsampling network may fail to generate realistic HR faces. These factors can be observed in Fig. 2(f), where simply employing an STN to align an LR image causes artifacts in the upsampled faces due to interpolation blur and alignment errors.

Instead of using a single STN to align LR face images, we employ multiple STN layers to line up the feature maps. Using multiple layers significantly reduces the load on each spatial transformation network. In addition, resampling feature maps by multiple STN layers prevents from damaging or blurring input LR facial patterns. Since STN layers and the upsampling network are interwoven together (rather than being two individual networks), the upsampling network can learn to eliminate the undesired effects of misalignment in the training stage. As shown in Fig. 2(e), our upsampling network can reconstruct more high-frequency details than the CNN based super-resolution method (SRCNN) (Dong, Loy, and He 2016), even when SRCNN is retrained with face patches.

Discriminative Network As seen in Fig. 2(e), the hallucinated faces are not sharp enough because the common parts learned by the upsampling network are averaged from similar components shared by different individuals. Thus, there is a quality gap between the real face images and the hallucinated faces. To bridge this gap, we inject class information. We integrate a discriminative network to distinguish whether the generated image is classified as an upright real face image or not. The parameters of the discriminative network are shown in the blue frame of Fig. 1. We employ a binary cross-entropy as the loss function. We backpropagate the discriminative error to revise the coefficients of the upsampling network, which enforces the facial parts learned by the deconvolutional layers to be as sharp and authentic as the real ones. A similar idea is employed in the generative adversarial networks (Goodfellow, Pouget-Abadie, and Mirza 2014; Denton et al. 2015; Radford, Metz, and Chintala 2015), which are designed to generate a new face. Furthermore, the use of class information also improves the performance of the STN layers for face alignment since only upright faces are classified as valid faces. Therefore, the discriminative network also determines whether the faces are upright or not. As shown in Fig. 2(g), with the help of the discriminative information, the hallucinated face embodies more authentic, much sharper and better aligned details.

Training Details of TDN

In the training stage of our TDN, we assemble LR and HR face image pairs $\{L_i, H_i\}$ as our training dataset. Notice that the LR image L_i is not directly downsampled from the HR image H_i. There are different rotations, translations, and scale changes applied in the LR images while the training HR images are kept upright.

For the upsampling network, we use a pixel-wise ℓ_2 regression loss. Our intuition here is that the hallucinated HR face image $\hat{H}_i$ should be similar to its corresponding reference HR image H_i. Since the STN layers are embedded in the upsampling network, the objective function $V(u, t)$ of

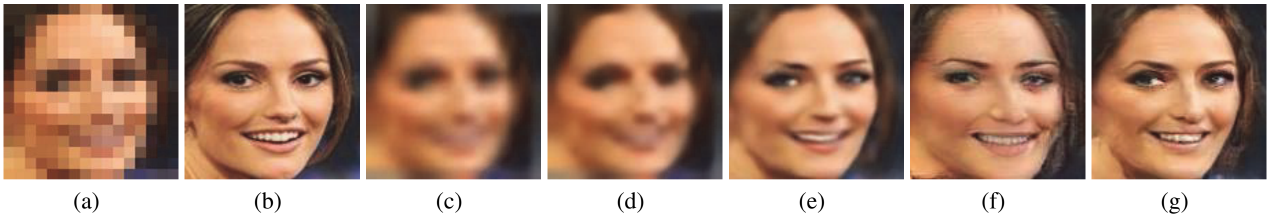

Figure 2: Illustration of TDN with different configurations. (a) Unaligned 16×16 LR image. (b) Original 128×128 HR image. (c) Bicubic interpolation. (d) Result of SRCNN (Dong, Loy, and He 2016) retrained with face patches. (e) Result of TDN without the discriminator network. (f) Result of TDN where an STN applied on the LR image directly. (g) Our full TDN.

the upsampling network is modeled as

$$\min_{u,t} V(u,t) = \mathbb{E}_{p(L_i,H_i)} \|\hat{H}_i - H_i\|_F^2, \qquad (1)$$

where u and t represent the parameters of the upsampling network and the STN layers that are updated jointly. The STN layers align the feature maps while the upsampling network super-resolves the LR images with the deconvolutional layers. Above, $p(L_i, H_i)$ represents the joint probability distribution of the LR and HR faces in the training dataset.

As we mentioned, we exploit the discriminative information to achieve high-quality super-resolution of face images. To this end, we employ a set of convolutional layers in our discriminative network. These layers assess whether the hallucinated face is real and upright, or not. If the upsampling network can hallucinate an HR face that can convince the discriminative network that it is an authentic face, our super-resolved face will be very similar to real face images. In other words, the discriminative network cannot differentiate upsampled faces from real faces. This objective is achieved by maximizing the cross entropy. Therefore, we optimize the loss function of the discriminative network D as follows:

$$\begin{aligned}\max_{d} D(d) &= \mathbb{E}\left[\log D(H_i) + \log(1 - D(\hat{H}_i))\right] \\ &= \mathbb{E}_{p(H_i)}[\log D(H_i)] + \mathbb{E}_{p(\hat{H}_i)}[\log(1 - D(\hat{H}_i))],\end{aligned} \qquad (2)$$

where d indicates the parameters of the discriminative network, and $p(H_i)$ and $p(\hat{H}_i)$ represent the distributions of real faces and the hallucinated faces from LR faces in the dataset. The above objective reaches the maximum when the network cannot distinguish H_i and $\hat{H}_i$. The loss D is back-propagated to the upsampling network to update the parameters u and t. By tuning u and t, the upsampling network not only can super-resolve the LR face images with appearance similarity, but also makes the hallucinated faces contain more class-specific details.

We use RMSprop (Hinton) to update the parameters u, t and d. In order to maximize D, the parameters d are updated by the stochastic gradient ascent,

$$\begin{aligned}\Delta^{j+1} &= \alpha\Delta^j + (1-\alpha)(\frac{\partial D}{\partial d})^2, \\ d^{j+1} &= d^j + \gamma\frac{\partial D}{\partial d}\frac{1}{\sqrt{\Delta^{j+1} + \epsilon}},\end{aligned} \qquad (3)$$

where γ and α are the learning rate and the decay rate, j represents the iteration index, Δ is an auxiliary variable, and ϵ is set to 10^{-8} to avoid division by zero. The parameters u and t are not only updated by the loss V but also D. For simplicity, let $T = (u, t)$, and the parameters are updated by the stochastic gradient descent,

$$\begin{aligned}\Delta^{j+1} &= \alpha\Delta^j + (1-\alpha)(\frac{\partial V}{\partial T} + \lambda\frac{\partial D}{\partial T})^2, \\ T^{j+1} &= T^j - \gamma(\frac{\partial V}{\partial T} + \lambda\frac{\partial D}{\partial T})\frac{1}{\sqrt{\Delta^{j+1} + \epsilon}},\end{aligned} \qquad (4)$$

where λ is used to trade off the appearance similarity constraint and the class-specific discriminative constraint. Since we aim to super-resolve an LR image, we put more constraint on appearance similarity. In our experiments, we set λ to 0.01. As the iterations progress, the upsampled faces become more similar to real faces, and thus we reduce the impact of the discriminative network gradually,

$$\lambda^i = \max\{\lambda \cdot 0.99^i, \lambda/2\}, \qquad (5)$$

where i indicates the index of the epochs. Eqn. 5 guarantees that the influence of the discriminative information is preserved in the upsampling network. In our algorithm, the learning rate γ is set to 0.001 and multiplied by 0.99 after each epoch, and the decay rate is set to 0.01.

Hallucinating a Very LR Face Image

The discriminative network is only used for training of the upsampling network. In the testing stage (super-resolving a given test image), we feed the LR image into the upsampling network to obtain its upright super-resolved HR version. Because the ground-truth HR face images are upright in the training stage of the entire network, the output of the upsampling network will be an upright face image. As a result, our method does not require alignment of the very low-resolution images in advance. Our network provides an end-to-end mapping from an unaligned LR face image to an upright HR version, which mitigates potential artifacts caused by misalignment.

Implementation Details

In Fig. 1, the STN layers are constructed by convolutional and ReLU layers (Conv+ReLU), max-pooling layers with a stride 2 (MP2) and fully connected layers (FC). In particular, STN$_1$ layer is cascaded by: MP2, Conv+ReLU (with

Table 1: Quantitative evaluation on the entire test dataset.

Methods	Bicubic	Yang (2010)	Dong (2016)	Liu (2007)	Yang (2013)	Ma (2010)	Ours
PSNR	18.41	18.21	18.28	18.00	18.40	18.34	**22.66**
SSIM	0.54	0.52	0.54	0.48	0.53	0.52	**0.66**

the filter size: 512×20×5×5), MP2, Conv+ReLU (with the filter size: 20×20×5×5), FC+ReLU (from 400 to 20 dimensions) and FC (from 20 to 4 dimensions). STN_2 is cascaded by: MP2, Conv+ReLU (with the filter size: 256×128×5×5), MP2, Conv+ReLU (with the filter size: 128×20×5×5), MP2, Conv+ReLU (with the filter size: 20×20×3×3), FC+ReLU (from 180 to 20 dimensions) and FC (from 20 to 4 dimensions). In the convolution operations, we do not use padding.

In the following experimental part, some algorithms require the alignments of LR inputs (Liu, Shum, and Freeman 2007; Ma, Zhang, and Qi 2010). Thus, we use STN_0 to align the LR inputs images for those methods. The only difference between STN_0 and STN_1 is that the first MP2 step in STN_1 is removed in STN_0.

Experiments

In this section, we compare our method with the state-of-the-art methods qualitatively and quantitatively. (More experimental results are given in the **supplementary material**.)

Dataset

Our network is trained on the Celebrity Face Attributes (CelebA) dataset (Liu et al. 2015). There are more than 200K face images in this dataset, and the images cover different pose variations and facial expressions. In training our network, we disregard these variations without grouping the face images into different pose and facial expression subcategories.

When generating the LR and HR face pairs, we randomly select 30K cropped face images from the CelebA dataset, and then resize them to 128×128 pixels as HR images. We manually transform the HR images while constraining the faces in the image region, and then downsample the HR images to generate their corresponding LR images. Note that, we do not explicitly change the scale of faces because in the CelebA the face sizes are different. (All protocol details, data, and code for this paper will be released.)

Comparison with the State-of-the-Art

Since we super-resolve an image with a substantial upscaling factor of 8×, for the methods that do not provide 8×, we apply the maximum upscaling factors recommended by the original papers multiple times (*e.g.*, twice 4× upscaling). For the face hallucination methods that assume very low-resolution faces are aligned beforehand, we use STN_0 to align LR faces. For fair comparisons and better illustration, we transform all the LR input images to the upright view as the inputs of the other methods.

In Tab. 1, we report the quantitative comparison results using the average PSNR and structural similarity scores (SSIM) on the entire test dataset. As indicated in Tab. 1, our TDN attains the best PSNR and SSIM results. We found that if we only use the upsampling network to super-resolve LR faces, we can gain an extra 0.18 dB improvement but produces over-smoothed results. Therefore, there is a trade-off between the upsampling and discriminative networks. Since we aim to hallucinate high-resolution realistic facial details, we incorporate our discriminative network, and our TDN achieves an impressive 4.25 dB PSNR improvement over the state-of-the-art.

As shown in Fig. 3(c), traditional upsampling methods, *i.e.*, bicubic interpolation, cannot hallucinate authentic facial details. Since the resolution of inputs is very small, little information is contained in the input images. Simply interpolating input LR images cannot recover extra high-frequency details. As seen in Fig. 3(c), the upsampled images by bicubic interpolation still have some skew effects rather than laying in the upright view. This implies that simply using STN_0 to align input images still suffers from misalignment. Since we apply multiple STNs on the feature maps, which improves the alignment of the LR inputs, our method outputs well-aligned faces. As shown in the last row of Fig. 3, STN_0 uses bilinear interpolation to resample images, which changes the intensities of the LR input and introduces extra blurriness as well. In contrast, with the help of the discriminator network, our method can achieve much sharper results.

As shown in Fig. 3(d), the sparse coding based super-resolution (SCSR) method (Yang et al. 2010) cannot reconstruct high-frequency details either when the scaling factor is very large (*e.g.*8×), because the SCSR method cannot find a consistent correspondence between LR and HR patches as the upscaling factor becomes larger.

Dong, Loy, and He (2016) propose a patch based convolutional network to super-resolve generic images, also known as SRCNN. This method is trained on generic patches and the maximum upscaling factor is 4. SRCNN, as a patch based method, cannot capture the whole face structure. However, training SRCNN with the whole face will introduce more ambiguity between LR and HR patches because the training patch size (*i.e.*128×128) is too large to learn a valid non-linear mapping. Hence, we retrain their model with face patches and an upscaling factor 8. As seen in Fig. 2(e), SRCNN cannot produce authentic high-frequency facial details. This also implies that our upsampling network is more suitable for the face hallucination task.

The face hallucination method based on appearance model (Liu, Shum, and Freeman 2007) can super-resolve

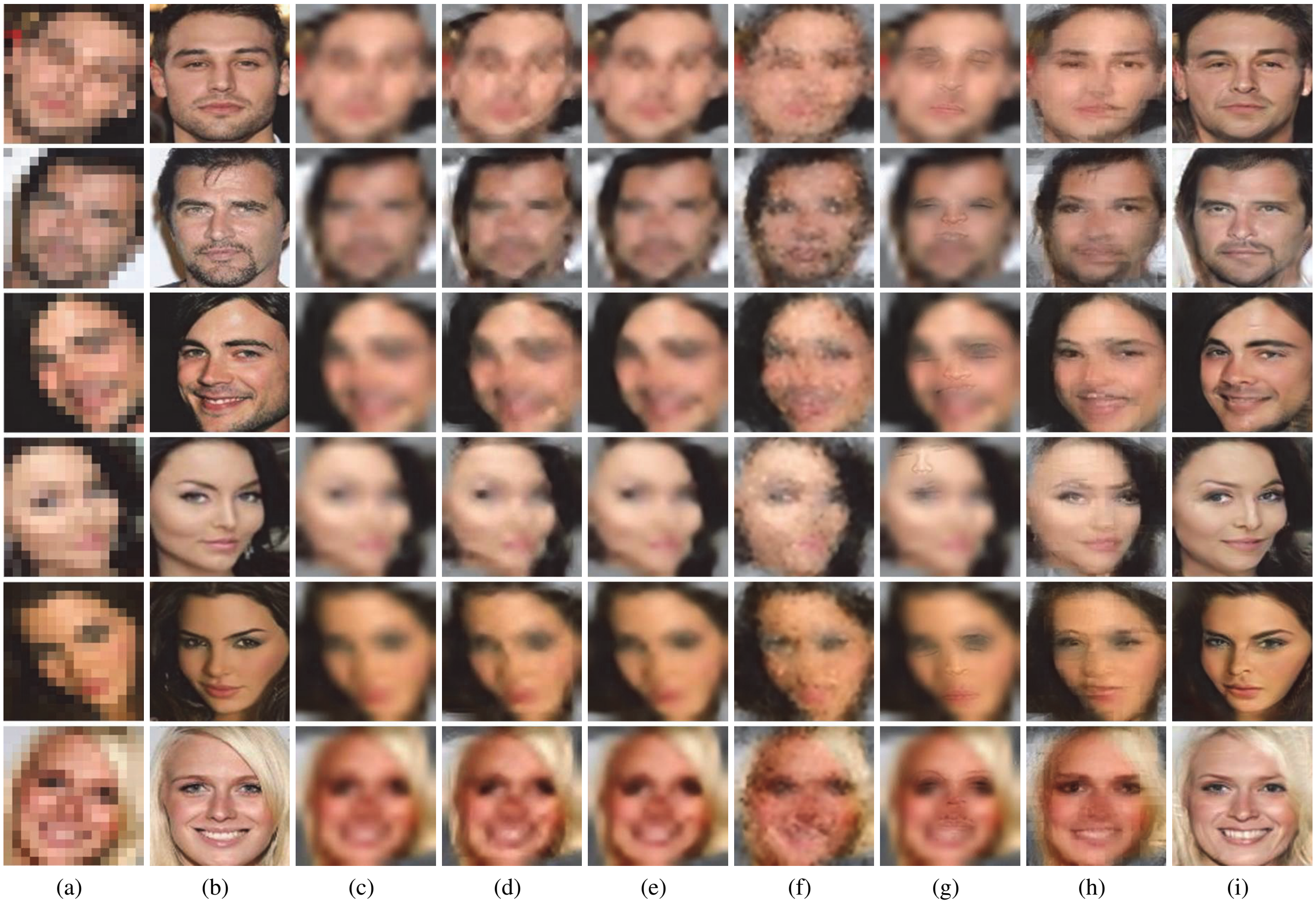

Figure 3: Comparison with the state-of-the-arts methods. (a) LR inputs. (b) Original HR images. (c) Bicubic interpolation. (d) Yang et al.'s method (2010). (e) Dong, Loy, and He's method (2016) (SRCNN). (f) Liu, Shum, and Freeman's method (2007). (g) Yang, Liu, and Yang's method (2013). (h) Ma, Zhang, and Qi's method (2010). (i) Our method.

very LR face images when the faces are precisely aligned (*i.e.*, face positions and head poses). Because the alignment errors of LR faces by STN_0 exist, the aligned LR faces have shifts with the appearance model. Besides, we use all the faces in the training dataset to train an appearance model, and there are different facial expressions and poses in the training dataset, which make the appearance model noisy. Hence, as shown in Fig. 3(f), their results suffer severe artifacts without hallucinating authentic facial details.

The structured face hallucination method (Yang, Liu, and Yang 2013) looks for the most similar facial components in the dataset and then transfer those HR components to the LR input ones. However, when the resolution of the input images is very small, localizing facial landmarks in LR inputs is difficult. Thus their method cannot accurately find the most similar facial components in the dataset and fails to output HR transferred components, as illustrated in Fig. 3(g). Therefore, this method is unsuitable to hallucinate very LR face images.

Ma, Zhang, and Qi's method (2010) exploits position patches to hallucinate HR faces. Thus this method requires the LR inputs to be precisely aligned with the reference images in the training dataset. As seen in Fig. 3(h), when there are obvious alignment errors in the aligned LR faces, their method will output mixed faces in their results. Furthermore, as the upscaling factor increases, the correspondences between LR and HR patches become more inconsistent. Hence, this method suffers from obvious block artifacts around the boundaries of different patches.

As shown in Fig. 3(i), our method reconstructs authentic facial details. Note that, the reconstructed faces have different poses and facial expressions. Since our method applies multiple STNs on feature maps to align face images, we can achieve better alignment results without damaging input LR images. Furthermore, our method does not need to warp input images directly, so there are no blank regions in our results. It implies that our method can exploit information better than the other methods.

Conclusions

We presented a transformative discriminative network to super-resolve unaligned very low-resolution face images in an end-to-end manner. Our network learns how to align faces and how to upsample them by making use of the class-specific information. It attains a significant upsampling factor of 8× while hallucinating rich and authentic facial details. Since our method does not require any feedback of face poses and facial expressions, it is very practical.

Acknowledgments

This work was supported under the Australian Research Council‘s Discovery Projects funding scheme (project DP150104645).

References

Arandjelović, O. 2014. Hallucinating optimal high-dimensional subspaces. *Pattern Recognition* 47(8):2662–2672.

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

Bruna, J.; Sprechmann, P.; and LeCun, Y. 2016. Super-resolution with deep convolutional sufficient statistics. In *ICLR*.

Denton, E.; Chintala, S.; Szlam, A.; and Fergus, R. 2015. Deep Generative Image Models using a Laplacian Pyramid of Adversarial Networks. In *Advances In Neural Information Processing Systems (NIPS)*, 1486–1494.

Dong, C.; Loy, C. C.; and He, K. 2016. Image Super-Resolution Using Deep Convolutional Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38(2):295–307.

Fischer, P.; Dosovitskiy, A.; Ilg, E.; Häusser, P.; Hazırbaş, C.; Golkov, V.; van der Smagt, P.; Cremers, D.; and Brox, T. 2015. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2758–2766.

Goodfellow, I.; Pouget-Abadie, J.; and Mirza, M. 2014. Generative Adversarial Networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2672—-2680.

Hennings-Yeomans, P. H.; Baker, S.; and Kumar, B. V. 2008. Simultaneous super-resolution and feature extraction for recognition of low-resolution faces. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 1–8. IEEE.

Hinton, G. Neural Networks for Machine Learning Lecture 6a: Overview of mini-batch gradient descent Reminder: The error surface for a linear neuron.

Jaderberg, M.; Simonyan, K.; Zisserman, A.; et al. 2015. Spatial transformer networks. In *Advances in Neural Information Processing Systems*, 2017–2025.

Kim, J.; Lee, J. K.; and Lee, K. M. 2015. Accurate Image Super-Resolution Using Very Deep Convolutional Networks. *arXiv:1511.04587*.

Kolouri, S., and Rohde, G. K. 2015. Transport-based single frame super resolution of very low resolution face images. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 4876–4884.

Li, Y.; Cai, C.; Qiu, G.; and Lam, K. M. 2014. Face hallucination based on sparse local-pixel structure. *Pattern Recognition* 47(3):1261–1270.

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

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

Liu, C.; Shum, H.; and Zhang, C. 2001. A two-step approach to hallucinating faces: global parametric model and local nonparametric model. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, 192–198.

Liu, C.; Yuen, J.; and Torralba, A. 2011. Sift flow: Dense correspondence across scenes and its applications. *IEEE transactions on pattern analysis and machine intelligence* 33(5):978–994.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ma, X.; Zhang, J.; and Qi, C. 2010. Hallucinating face by position-patch. *Pattern Recognition* 43(6):2224–2236.

Radford, A.; Metz, L.; and Chintala, S. 2015. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *arXiv:1511.06434* 1–15.

Tappen, M. F., and Liu, C. 2012. A Bayesian Approach to Alignment-Based Image Hallucination. In *Proceedings of European Conference on Computer Vision (ECCV)*, volume 7578, 236–249.

Wang, X., and Tang, X. 2005. Hallucinating face by eigen transformation. *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews* 35(3):425–434.

Wang, Z.; Yang, Y.; Wang, Z.; Chang, S.; Han, W.; Yang, J.; and Huang, T. 2015. Self-tuned deep super resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 1–8.

Yang, J.; Wright, J.; Huang, T. S.; and Ma, Y. 2010. Image super-resolution via sparse representation. *IEEE transactions on image processing* 19(11):2861–73.

Yang, C. Y.; Liu, S.; and Yang, M. H. 2013. Structured face hallucination. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 1099–1106.

Zhou, E., and Fan, H. 2015. Learning Face Hallucination in the Wild. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 3871–3877.