Exemplar Guided Face Image Super-Resolution without Facial Landmarks

Berk Dogan, Shuhang Gu, Radu Timofte
Computer Vision Lab, D-ITET, ETH Zurich
doganb@student.ethz.ch, shgu@ee.ethz.ch, radu.timofte@vision.ee.ethz.ch

Abstract

Nowadays, due to the ubiquitous visual media there are vast amounts of already available high-resolution (HR) face images. Therefore, for super-resolving a given very low-resolution (LR) face image of a person it is very likely to find another HR face image of the same person which can be used to guide the process. In this paper, we propose a convolutional neural network (CNN)-based solution, namely GWAInet, which applies super-resolution (SR) by a factor $8\times$ on face images guided by another unconstrained HR face image of the same person with possible differences in age, expression, pose or size. GWAInet is trained in an adversarial generative manner to produce the desired high quality perceptual image results. The utilization of the HR guiding image is realized via the use of a warper subnetwork that aligns its contents to the input image and the use of a feature fusion chain for the extracted features from the warped guiding image and the input image. In training, the identity loss further helps in preserving the identity related features by minimizing the distance between the embedding vectors of SR and HR ground truth images. Contrary to the current state-of-the-art in face super-resolution, our method does not require facial landmark points for its training, which helps its robustness and allows it to produce fine details also for the surrounding face region in a uniform manner. Our method GWAInet produces photo-realistic images in upscaling factor $8\times$ and outperforms state-of-the-art in quantitative terms and perceptual quality.

1. Introduction

Face image super-resolution or face hallucination aims at reconstructing details / high-frequencies in low-resolution (LR) face images. This is an important problem due to the increasing need for high-resolution (HR) face images for different applications such as security, surveillance or other application that involves face recognition.

Due to the increasing interest in visual media and the development of the social media, it is very likely that given a LR face image of a person, we can find another HR face image of the same person possibly taken at a different time in different conditions. This guiding face could be used in the super-resolution (SR) process to guide the hallucination of high frequencies/details, which might increase the quality of the HR result and help to preserve the identity related features. Fig. 1 shows such a case and our result.

Figure 1: Exemplar guided face image super-resolution result ($8\times$) of our proposed GWAInet approach.

The current state-of-the-art face image super-resolution approach [27] proposed the use of a guiding image together with a facial landmark detector, where an additional loss term is optimized such that the facial landmarks of the warped guiding image and those of the ground truth image are close to each other. However, this approach seems to produce fine details for the face region in a non-uniform and unpredictable manner, resulting in SR images that look only partially sharp.

Although the recently proposed CNN-based SR solutions [37, 40] provide state-of-the-art quantitative results in terms of peak signal-to-noise ratio (PSNR) when they optimize for reconstruction losses such as L1 or L2 in image space, the results are smooth without the fine details required for a good perceptual quality. This problem is more visible with the increase of the upscaling factor [22, 5]. On top of that, the PSNR measure is unable to capture perceptually important differences between two images as it relies on the differences between pixel-level values at the same position [41, 42, 15]. One way to introduce perceptually important features into the SR image is to use generative adversarial networks (GANs) [13, 26, 5]. These networks help to create realistic SR images that look like HR images, which are naturally sharper and contain fine details.

In this paper, we introduce a novel CNN architecture capable of generating high quality HR face images with an upscaling factor $8\times$. During the SR process, the network utilizes the LR face image and the extra information provided by another HR face image of the same person while making the necessary processing through a warping subnetwork on this guiding HR image. By addressing the possible differences in contents (*e.g*. expression, pose, size) between two images the warper facilitates the extraction and integration of information from the guiding image. Contrary to the current state-of-the-art approach [27], our method does not require facial landmarks during training. This makes the network to learn and process the whole face region in a uniform manner and adds robustness. We add also an identity loss to further help in preserving the identity related features by minimizing the distance between the embedding vectors of HR result and ground truth images. Utilization of the guiding image expresses itself qualitatively as an improvement in visual content quality by correcting the inaccurate facial details. Finally, the adversarial loss that is incorporated via a GAN setting, will introduce fine details to the SR image and produce face images that are hardly distinguishable from real HR face images.

2. Related Work

Convolutional Neural Networks (CNNs) and Image Super-Resolution (SR). CNNs have emerged as a successful method in many computer vision applications [4, 25, 36]. Deep learning with CNNs has also become very widely used in image SR [23, 11, 26, 29]. CNNs have outperformed previous works [44, 49, 38, 39, 1] both quantitatively and qualitatively. These networks, on the other hand, are mainly used for SR of single or multiple LR images and do not utilize a guiding HR image for the given LR image. In our work, we use the network given in [29] as the main structural element of our subnetworks and specifically work on face images while utilizing the additional information provided by another HR face image of the same person.

Spatial Transformer Networks. Spatial transformer networks are modules that can be incorporated into an existing network and trained in an end-to-end fashion without any modification to the learning scheme or the loss function [21]. They increase the spatial invariance of the network and provide invariance for large transformations [21]. They are used to spatially transform input feature maps and consist of a localisation network and a sampler. In our work, we use the ideas from spatial transformer networks to create a flow field, which is then used in combination with a bilinear sampler to warp the guiding image, thus making the guiding image aligned with the contents of the input image.

Face Hallucination. CNNs have also shown great success in the field of face hallucination, where we apply super-resolution on face images [47, 46, 51, 27, 52, 43, 7, 50, 48]. [47] applies face hallucination on tiny 16×16 faces. In [46, 48], the authors again work on tiny 16×16 images but use spatial transformer networks [21] in their generator architecture to alleviate the effects of misalignment of input images. Zhou *et al*. [51] fuse two channels of information, namely extracted facial features and the LR input image, in order to overcome problems related with appearance variations and misalignment. However, they use resource intensive fully-connected layers in upscaling process and follow a simple fusion operation by just summing the upscaled LR input image via bicubic interpolation with the HR image created from the facial features. In our approach, on the other hand, we cope with the effects of appearance variations through the use of spatial transformer networks. However, contrary to [46, 48], we introduce the spatial transformer network as a subnetwork that is only applied to the input image rather than intermediate feature maps and contrary to [51], we use resource efficient convolutional layers in upscaling process and follow a complex feature fusion for the information coming from two channels. Most importantly, these works do not incorporate the use of an additional HR image of the same person. In a very recent work [27], Li *et al*. use a guiding image and a warper subnetwork to cope with appearance variations between the LR input and the HR guiding image. However, they apply direct concatenation of warped guiding image and upscaled LR image at the input of the generator, which is different than our feature extraction and fusion based approach through the use of secondary feature extractor subnetwork, which is called GFEnet, for the warped guiding image. They also use landmark loss and total variation loss for their warping subnetwork in the joint training phase, whereas we do not incorporate these losses in our overall objective, thus our network does not require facial landmarks during training. Another difference is that they use conditional adversarial networks [20] for generating the adversarial loss, whereas we use a Wasserstein generative adversarial network with gradient penalty (WGAN-GP) [14]. [27] is the only recent paper known to us that uses an additional guiding image in face hallucination. As in our approach, Zhang *et al*. [50] also use an identity loss in face hallucination problem, which is calculated between the SR image and the ground truth image.

Generative Adversarial Networks (GANs). Although the SR methods using CNN architectures provide state-of-the-art quantitative results in PSNR terms when optimized for reconstruction losses such as L1 or L2 in image space, they produce overly-smooth visuals and lack the ability to produce images with fine details. The PSNR metric does not correlate well with the human perception of image quality [16]. This is due to the fact that the reconstruction loss is calculated in image space and the optimum solution is the average of all possible solutions [12, 26, 6]. GANs [13]

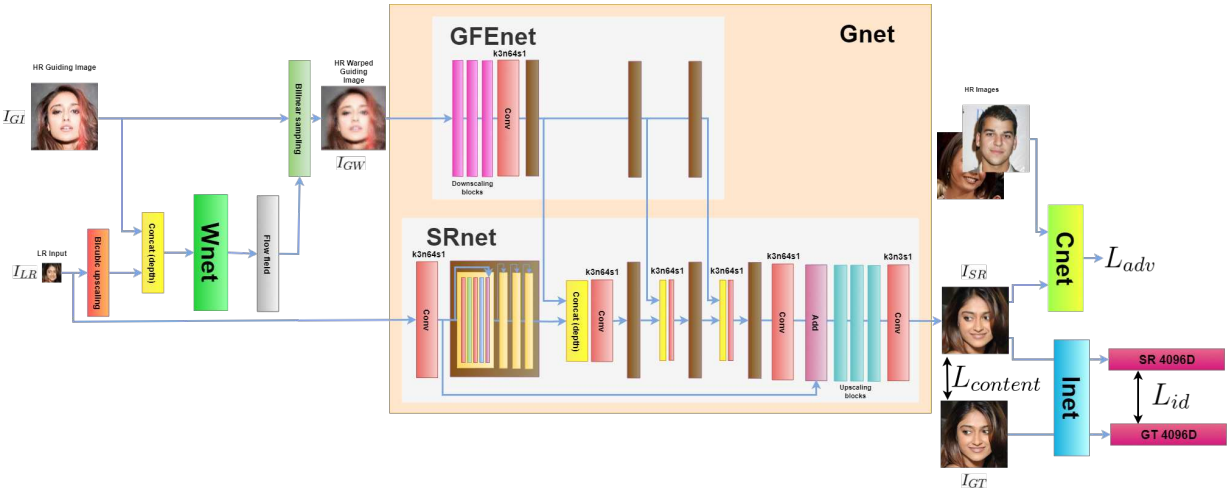

Figure 2: Proposed GWAInet and its Warper (Wnet), Generator (Gnet), Critic (Cnet) and Identity Encoder (Inet) subnetworks.

have become successful in creating realistically looking images thanks to their adversarial loss. As a result of this, many methods make use of GANs. [12] uses both a loss in feature space and an adversarial loss, in addition to the reconstruction loss in image space to generate sharp and natural looking images. Besides adversarial loss and reconstruction loss in image space, [31] uses an additional image gradient difference loss between the input and the output that sharpens the image prediction to predict future images from a video sequence. In [47, 46], they use adversarial loss and pixel loss to super-resolve tiny face images such that the resulting SR images have high frequency components. In [26], they use adversarial loss and feature loss for VGG-19 [34] network to produce sharp and photo-realistic SR images. In [27], they also include adversarial loss in their total loss to improve the output visual quality of face restoration tasks from degraded observations. In our work, we use adversarial loss together with the L1 reconstruction loss in image space. Reconstruction loss drives the networks to match the contents of the output SR image with the contents of the input LR image. The adversarial loss, on the other hand, tries to ensure that the SR image contains high frequency features that make it photo-realistic. As a result of this loss combination, we get an SR image that agrees with the input LR image in terms of facial feature location and the coarse specifications for these facial features but also agrees with the specifications imposed by the distribution of the HR face images. We specifically use WGAN-GP, which optimizes for a different metric than the traditional GANs and is found to be more stable and easier to train [14].

3. Proposed Method

Our proposed GWAInet solution (Guidance, Warper, Adversarial loss, Identity loss network) produces a SR image I_{SR} from a LR input image I_{LR} and a HR guiding image I_{GI}. I_{LR} is obtained from a ground truth high-resolution image I_{GT} by downscaling with bicubic interpolation in scale $8\times$. I_{GI} is another HR face image of the person, to whom the tuple (I_{LR}, I_{GT}) belongs to. We also denote the image that is obtained by warping I_{GI} as I_{GW}.

In the following, we provide detailed information about our GWAInet method. First, we briefly describe the WGAN-GP and then we present the network architecture of our model. Finally, we describe the loss functions used in guiding the optimization process.

3.1. WGAN-GP [14]

We use a generative adversarial network (GAN) approach [13] to generate perceptually good and sharp images. Specifically, we use a Wasserstein GAN with gradient penalty (WGAN-GP) [14]. With the help of this new architecture, we are aiming to make the SR images from our proposed network GWAInet as indistinguishable as possible from the HR images in our dataset. This is possible due to the structure of WGAN-GP and its training objective.

GANs consist of a generator subnetwork G and a discriminator subnetwork D, where the aim of G is to create samples that are as close as possible to the real data samples and the aim of the D is to classify these fake samples from the real ones. If we denote the real data distribution by $p(x)$ and generated data distribution by $q(y)$, then objective can be formulated as [13]:

$$min_G\, max_D \mathbf{E}_{x' \sim p(x)} \left[logD(x') \right] + \mathbf{E}_{y' \sim q(y)} \left[1 - logD(y') \right] \quad (1)$$

In the traditional GAN setting, for an optimal discriminator, we are trying to minimize the JS-divergence between the real and generated data distributions [13]. With this approach, training the model is a difficult process. This is due to the fact that in many practical problems, the real and the generated data distribution are disjoint in some low dimensional manifold in a high dimensional space, which makes it easier to find a perfect discriminator [2]. When the discriminator becomes perfect, the gradient coming from the JS-divergence vanishes [13]. There is an alternative method to avoid vanishing gradients by maximizing $\mathbf{E}_{y' \sim q(y)} \left[-logD(y') \right]$ for G, however it is shown that this method causes unstable updates [2].

Wasserstein generative adversarial network (WGAN) introduces a new loss function for GAN training, which depends on the Earth-Mover distance [3] formulated as:

$$W(p,q) = inf_{\gamma \in \prod(p,q)} \mathbf{E}_{(x',y') \sim \gamma(x,y)} \left[\|x' - y'\| \right] \quad (2)$$

In equation 2, $\prod$ denotes the set of all joint distributions $\gamma(x,y)$, where $\sum_y \gamma(x,y) = p(x)$ and $\sum_x \gamma(x,y) = q(y)$. $\gamma(x,y)$ can be seen as the amount of earth that should be transported from x to y to transform p into q. WGAN-GP is the improved version of WGAN, with an addition of gradient penalty term in the cost function instead of weight clipping procedure [14].

3.2. Network Architecture

The complete network architecture of the proposed solution is illustrated in Figure 2. The complete model, called GWAInet, consists of four network components, namely Gnet, Wnet, Cnet and Inet.

Warper (Wnet). Wnet's aim is to produce the flow field required to warp the guiding image such that it is well aligned with the contents of the input LR image, removing any difference in pose or size of the faces in both images. This warping procedure allows the extra information provided by I_{GI} to be better utilized. Wnet is essentially the localisation network for a spatial transformer network [21]. It produces the transformation parameters that is fed into the bilinear sampler along with I_{GI}. Before the first layer of Wnet, upscaling via bicubic interpolation is applied to I_{LR}, which scales the spatial dimensions by 8, producing the image I_{LRU}. After that, I_{GI} and I_{LRU} are concatenated along the depth axis. The resulting tensor forms the input of Wnet. Wnet outputs a 3D flow field with 2 depth channels. At each pixel location, the first value determines the sampling motion horizontally and the second value determines the sampling motion vertically. It should be noted that flow field values are not scaled into a specific range, meaning that no constraints are applied at the output. The flow field and I_{GI} are used in the bilinear sampling module to produce the warped guiding image I_{GW}. Let us denote the flow field as $\Omega \in \mathbb{R}^{h_{HR} \times w_{HR} \times 2}$ and denote the pixel location grid for I_{GW} as $\delta \in \mathbb{R}^{h_{HR} \times w_{HR} \times 2}$, where $\delta(i,j,0) = \frac{2 \times i}{h_{HR}-1} - 1$ and $\delta(i,j,1) = \frac{2 \times j}{w_{HR}-1} - 1 \; \forall i \in \{0,1,...,h_{HR}-1\}$ and $\forall j \in \{0,1,...,w_{HR}-1\}$. Note that the grid values are in the range $[-1,+1]$ instead of $[0, h_{HR}-1]$ for $\delta(i,j,0)$ and $[0, w_{HR}-1]$ for $\delta(i,j,1)$. In other words, in our setting, we assume that the top left corner of the image has coordinates (-1,-1) and the bottom right corner of the image has coordinates (+1,+1). Using Ω and δ, the sampling grid $\rho \in \mathbb{R}^{h_{HR} \times w_{HR} \times 2}$ can be calculated as [21]:

$$\rho = \Omega + \delta \quad (3)$$

This sampling grid dictates where to sample from the original input image, I_{GI}, for an output pixel in the output image, I_{GW}, which is the warped guiding image. After the calculation of ρ, the values are scaled back to $[0, h_{HR}-1]$ for $\delta(i,j,0)$ and $[0, w_{HR}-1]$ for $\delta(i,j,1)$ using the inverses of the previously given transforms. If we let $I(i,j,c)$ represents the pixel intensity value at the $(height = i, width = j, channel = c)$ location of the some image I, then the pixel intensity values at the output of the bilinear sampling module can be calculated using the following formula [21]:

$$\begin{aligned} I_{GW}(i,j,c) = \sum_{(a,b) \in H} & I_{GI}(a,b,c) \\ & max\left\{0, 1 - |\rho(i,j,0) - a|\right\} \\ & max\left\{0, 1 - |\rho(i,j,1) - b|\right\} \end{aligned} \quad (4)$$

In equation 4, H refers to the 4 closest pixel indices with respect to the coordinate given by $height = \rho(i,j,0)$ and $width = \rho(i,j,1)$. Wnet can be trained end-to-end with a loss function using gradient based methods due to the fact that I_{GW} is sub-differentiable with respect to the parameters of the Wnet [21].

Generator (Gnet). Gnet is the network that applies SR on the I_{LR} while using the additional information provided by the warped guiding image I_{GW}. It consists of two smaller subnetworks, which are called SRnet and GFEnet. These two subnetworks represent two channels of information. SRnet takes only I_{LR} as input, whereas GFEnet takes only I_{GW} as input.

SRnet is the same baseline architecture used in [29]. The architecture is given in Figure 2. It consists of 16 residual blocks [17], whose architecture can be found in the supplementary material. In our setting, scale parameter is set to $\alpha_{res} = 1$, which is recommended in [29]. Throughout SRnet, spatial dimensions of I_{LR} is preserved via zero padding. After the merging point of the global skip connection, upscaling blocks come, whose main responsibility is to gradually upscale the feature maps such that their spatial dimensions match with the spatial dimensions of I_{GT}. The upscaling is done via efficient sub-pixel convolutional

layers [33], that is in each upscaling block, $2\times$ upscaling is performed by cascading a convolutional layer and a pixel shuffler layer. These convolutional layers apply a 3×3 filter with stride 1 and they have 256 feature maps.

GFEnet, which is used as a feature extractor for the I_{GW}, consists of 3 downscaling blocks and 12 residual blocks. As can be seen in Figure 2, each downscaling block is used to downscale the spatial dimensions of its input by 2. In each downscaling block, first, a convolutional layer with 3×3 kernel, 64 feature maps and stride 1 is applied, which is followed by a ReLU. Then another convolutional layer with 3×3 kernel, 64 feature maps and stride 2 is applied, which is again followed by ReLU. Downscaling of the I_{GW} is done through series of stride 2 convolutions instead of max-pooling operation. The motivation is to let the model learn the downscaling procedure instead of fixing it [35]. After every 4th residual block, the current features that come from GFEnet and features that come from SRnet are fused. This feature fusion is done via concatenation along the depth axis. Since only convolutions with 64 output feature maps are used in both subnetworks, after the concatenation, a feature map of depth 128 is obtained. A convolution operation follows this concatenation before the signal resulting from the fusion operation enters to the next residual block of SRnet. After 12th residual block, which also means that after the 3rd feature fusion, GFEnet reaches to an end.

Critic (Cnet). The critic network is the same discriminator network that is used in DCGAN architecture [32] except we do not use batch normalization layers [19]. The exact specifications of the architecture is given in the supplementary material.

Outputs of Gnet, I_{SR}, form the generated samples, which should be criticized as fake images by the critic. The real images, which are samples from the real data distribution, are the same images that are used as the ground truth HR images for the LR input images. These should be criticized as real images by the critic.

Identity Encoder (Inet). We use a Siamese network [9] for generating embedding vectors related with the identity of the person. We have selected VGG-16 network [34] as the architecture of our siamese Inet, whose details can be found in the supplementary material. Given a SR face image I_{SR} and its corresponding HR ground truth image I_{GT}, Inet is used to evaluate their similarity. This similarity information is then used to penalize I_{SR} that has characteristics that differ from the characteristics of its corresponding I_{GT}. To learn the parameters θ_{Inet}, we cast the problem as a binary classification problem, in which Inet tries to predict whether the two input images belong to the same person. This procedure can be guided by cross-entropy loss function, where the output is equal to

$$y = sigmoid(w^T|Inet(x_1;\theta_{Inet}) - Inet(x_2;\theta_{Inet})| + b) \tag{5}$$

where $Inet(x;\theta_{Inet}), w \in \mathbb{R}^{4096}$ and $b \in \mathbb{R}$. Note that the parameters w and b are only used during pretraining of Inet. Moreover, during the optimization of Wnet, Gnet and Cnet, θ_{Inet} is frozen.

3.3. Loss Functions

Content loss $L_{content}$. Content loss is equal to L1 loss in our setting and can be calculated as:

$$L_{content} = \frac{1}{3h_{HR}w_{HR}} \sum_{j=1}^{h_{HR}} \sum_{k=1}^{w_{HR}} \sum_{c=1}^{3} |I_{SR}(j,k,c) - I_{GT}(j,k,c)| \tag{6}$$

$L_{content}$ ensures that contents of super-resolved image I_{SR} match with those of I_{GT}. Although this loss is vital in keeping the connection between I_{SR} and I_{GT}, and optimizes for high PSNR values, it results in I_{SR} images that are formed by smooth regions and that lack high-frequency details [26].

Adversarial loss L_{adv}. The aim of adversarial loss is to to make SR images look perceptually good and photo-realistic, making generated SR data distribution and real HR data distribution as close as possible to each other. With the help of the adversarial loss, the SR image will have fine details and the network will combat the smoothing effect caused by the content loss.

The adversarial loss incurred by the WGAN-GP for the generator is equal to $-L_{fake}$. Note that $L_{fake} = D(I_{SR};\theta_{Cnet})$, where θ_{Cnet} represents the parameters of the critic and $D(I_{SR};\theta_{Cnet})$ represents the output of the critic for I_{SR} image.

Identity loss L_{id}. Identity loss is calculated as the squared Euclidean norm of the distance between the embedding vectors of I_{SR} and I_{GT}. Thus,

$$L_{id} = \frac{\|Inet(I_{SR};\theta_{Inet}) - Inet(I_{GT};\theta_{Inet})\|^2}{4096} \tag{7}$$

L_{id} is used to penalize I_{SR} that has characteristics that differ from the characteristics of its corresponding I_{GT}, thus increasing the perceptual quality of the SR image.

Critic loss L_c. We can calculate the loss incurred by the WGAN-GP for the critic as:

$$L_c = L_{fake} - L_{real} + \lambda_{gp}L_{gp} \tag{8}$$

where $L_{real} = D(I_{GT};\theta_{Cnet})$ and $D(I_{GT};\theta_{Cnet})$ represents the output of the critic for I_{GT} image. λ_{gp} is the coefficient for the gradient penalty and L_{gp}, which is the gradient penalty term, is a function of I_{SR}, I_{GT}, θ_{Cnet} and $\epsilon \sim Uniform\,[0,1]$. Exact details are given in [14].

Overall objective. The overall objective function for the optimization of θ_{Wnet} and θ_{Gnet} can be written as:

$$L_{total} = L_{content} + \lambda_{adv}L_{adv} + \lambda_{id}L_{id} \tag{9}$$

where λ_{adv} and λ_{id} are the weighting coefficients for L_{adv} and L_{id} respectively. The overall objective function for the optimization of θ_{Cnet} is directly equal to L_c. Optimization of these two objectives are done in an alternating fashion as also described in [14]. Note that during this training procedure θ_{Inet} is frozen.

4. Experiments

4.1. Datasets

CelebA [30]. We used this dataset in developing our network and we moved to the dataset of [27] for comparing our method with the state-of-the-art. We use the aligned and cropped version of the CelebA dataset. We select the same train-validation-test partitioning used by the creators of the dataset. We have removed all of the identities that has a single image from the dataset, resulting in 162,734 training, 19,862 validation and 19,959 test images. It should be noted that the identities in each set are disjoint. To remove as much background as possible and to focus on the faces, we further crop the images to dimensions 168×144. For a given LR input image, the guiding image is sampled uniformly from the remaining HR images of the same person.

Dataset of [27]. This dataset is a subset of VggFace2 [8] and CASIA-WebFace [45] datasets. All of the images are collected from the wild and therefore include different expressions, pose and illumination conditions. For each identity, pairs of HR guiding and ground truth images are available. All HR images have spatial dimensions 256×256. Different from [27], we randomly select 2,273 among 20,273 training images as validation images, which means that we are working with a smaller training set of size 18,000.

4.2. Experiment Settings

As a preprocessing step, we scale the input pixel intensity values from $[0, 255]$ to $[0, 1]$ and then subtract the mean of the training dataset. We also scale the range of the I_{GT} to $[-1, 1]$.

We always use Adam optimizer [24]. During training, whenever $\lambda_{adv} = 0$, we use the suggested parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$ [24]. Whenever $\lambda_{adv} \neq 0$, we use $\beta_1 = 0.5$ and $\beta_2 = 0.9$. During this adversarial training, we always apply 5 critic updates per each generator update and we set $\lambda_{gp} = 10$.

Training on CelebA. This dataset is used only during the development of the proposed method. The identity loss is not used during the training of GWAInet, thus λ_{id} in Equation 9 is always set to 0. The training consists of three steps. We first pretrain the Wnet using L1 reconstruction loss between the warped guiding image and the ground truth image with learning rate 0.0001 for 1.25 epochs with batch size 4. In the second step, we train the whole network by setting $\lambda_{adv} = 0$ in Equation 9 with learning rate 0.0001 and batch size 16 until the validation PSNR reaches its peak. Then we set $\lambda_{adv} = 0.001$, and continue training for 4 epochs using batch size 4 and then another 2 epochs with learning rate 0.00005.

Figure 3: CelebA results without (BAnet) and with (GWAnet) the use of the guiding face. [$8\times$ upscaling, LR input spatial dimensions 21×18]

Training on the dataset of [27]. We train the full model on this dataset. The training of GWAInet consists of two steps. We first train the network by setting $\lambda_{adv}, \lambda_{id} = 0$ in Equation 9 with learning rate 0.0001 and batch size 48 until the validation PSNR reaches its peak. Then we set $\lambda_{adv} = 0.001$ and $\lambda_{id} = 0.05$, and continue training for 8 epochs using batch size 16. During the training of GWAInet, parameters of Inet is fixed. The Inet is pretrained on the same training set for 12 epochs with learning rate 0.0001 and batch size 8. [1]

4.3. Results on CelebA

After training on the CelebA dataset, we obtain the model GWAnet, which is the same model as the proposed full model GWAInet except that it does not include the identity loss in its optimization objective. Note that the identity loss is not related with the utilization of the guiding image because it is calculated between the super-resolved image and the ground truth image.

Model without guiding image. To evaluate the importance of the guiding image and the subnetworks related with the guiding image, i.e. Wnet and GFEnet component of Gnet, we create a new model called BAnet, which only includes Cnet and SRnet component of Gnet. It is trained exactly in the same fashion as GWAnet. GWAnet, with the help of the guiding image, almost always improves the quality of the face image by adding some missing details about the facial features of the person over BAnet. GWAnet

[1] Our codes, models and results are publicly available on the project page: `https://github.com/berkdogan2/GWAInet`

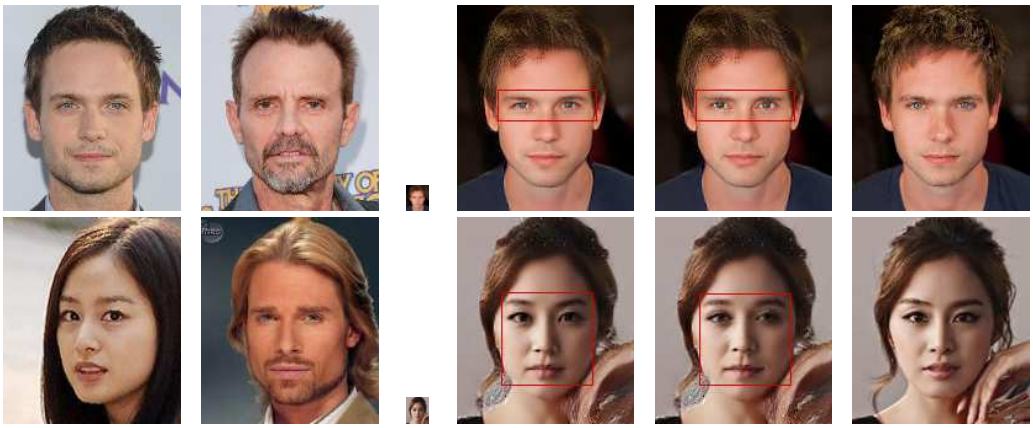

Figure 4: Comparison of I_{SR} face super resolved images for the cases when a guiding image with the same identity is used (GWAnet) and when a guiding image with a different identity is used (GWAnet-R) for CelebA dataset. [$8\times$ upscaling, LR input spatial dimensions 21×18]

outperforms BAnet in generating perceptually good looking face images due to the fact that GWAnet provides better visual content quality by correcting the inaccurate facial details through the use of I_{GI}. In this context, visual content quality refers to the extent that the characteristics of the facial contents of I_{SR} image match with those of I_{GT} image. As can be seen in Figure 3, BAnet is fully capable of generating photo-realistic images as well as GWAnet but the point that sets GWAnet apart from BAnet is its ability to complete the missing facial details in I_{SR} image by utilizing the guiding image I_{GI}. The improvements express themselves as location and shape improvements of facial features such as eyes, eyebrows, nose, mouth, hair and wrinkles.

Guiding image with a different identity In order to evaluate the magnitude of the effect of the guiding image in generating I_{SR}, we have carried over an experiment, where for a given identity, we feed a randomly selected guiding image with a different identity. We denote the resulting model as GWAnet-R. The comparison of I_{SR} images for GWAnet and GWAnet-R is shown in Figure 4. In general, when the guiding image has a different identity, the resulting differences from the standard model are noticeable. For some cases, as exemplified by the second row in Figure 4, the complete facial structure of the person changes. The mentioned differences mainly express themselves as location and shape differences of eyes and eyebrows as shown in the first row in Figure 4. There are also cases, where wrinkles appear or disappear according to the selected guiding image. The qualitative differences between GWAnet and GWAnet-R suggest that apart from being an additional information about high-resolution face images, the identity of the guiding image also plays an important role in generating high quality face images.

4.4. Comparison with state-of-the-art Methods

We compare our results quantitatively with the state-of-the-art face hallucination methods CBN [52], WaveletSR [18], TDAE [48], GFRNet [27] and super-resolution methods SRCNN [10], VDSR [23], SRGAN [26]. For all those methods, we directly use the results reported in [27]. Moreover, we compare our results qualitatively with GFRNet [27], which is the current state-of-the-art in face hallucination. Note that all of our experiments are performed for upscaling factor $8\times$.

Quantitative comparison. The quantitative results are shown in Table 1. As can be seen from Table 1, GWAInet outperforms the state-of-the-art in VggFace2 dataset by 1.47dB. It is the second best method in WebFace dataset and lags behind GFRNet [27] by 0.1dB. However, we should note that the training of GWAInet is not optimized for highest PSNR due to the adversarial loss and identity loss terms in the overall objective, which conflict with the objective of maximizing PSNR. PSNR is not well capable of capturing perceptual quality in an image [26, 41, 42, 15]. Moreover, it is possible to get highest PSNR values by training GWAInet shorter or longer with very small decrease in perceptual quality. We did not follow such a path because the focal point of this paper is presenting the capability of GWAInet in producing perceptually high quality SR images.

Qualitative comparison. As can be seen from Figure 5, our method GWAInet produces better looking and sharper face images than the state-of-the-art. GFRNet [27] only sharpens a small area in the face region, whereas our method GWAInet introduces high frequency details for all parts of the image, including the hair. Moreover, GFRNet [27] generally completely hallucinates the face of the per-

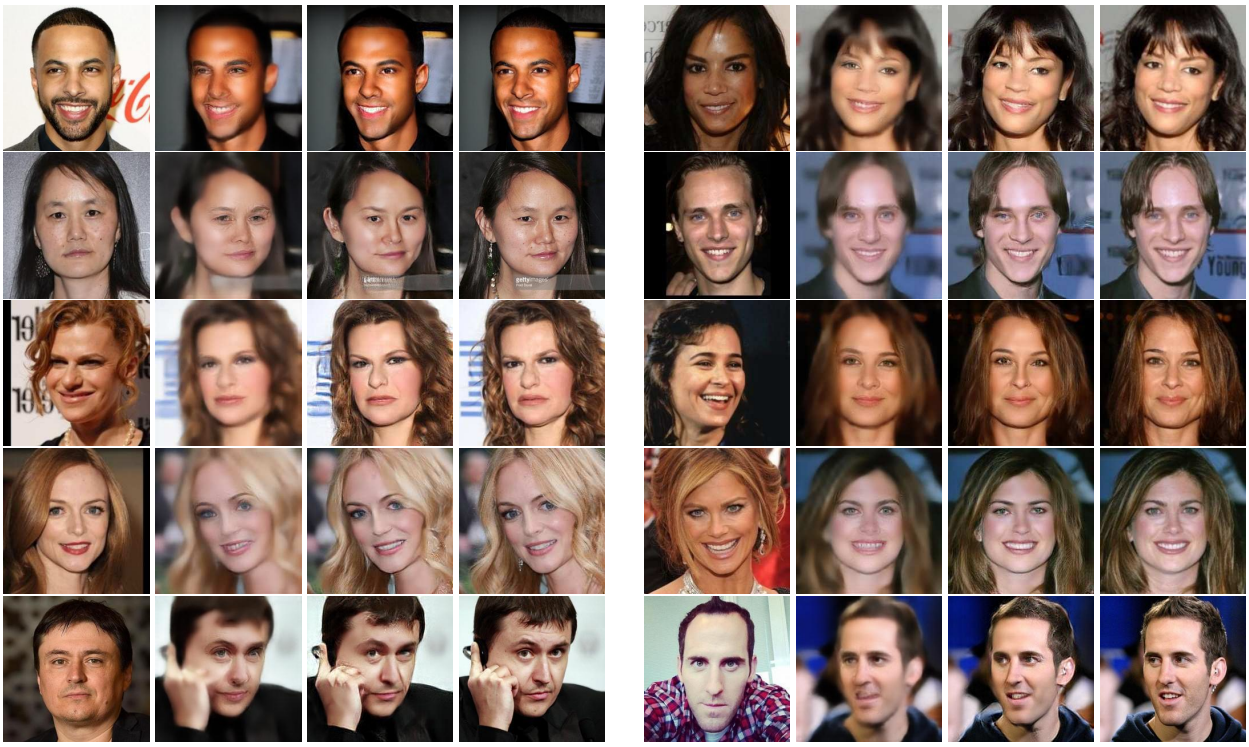

Figure 5: Comparison with state-of-the-art. Our full model GWAInet produces perceptually high quality images while retaining the facial features related with the identity. To obtain the results for GFRNet, their publicly available model is used [27, 28]. [8× upscaling, LR input spatial dimensions 32 × 32]

Method	VggFace2 [8]	WebFace [45]
SRCNN [10]	22.30	23.50
VDSR [23]	22.50	23.65
SRGAN [26]	23.01	24.49
CBN [52]	21.84	23.10
WaveletSR [18]	20.87	21.63
TDAE [48]	20.19	20.24
GFRNet [27]	24.10	27.21
Ours (Full-GWAInet)	25.57	27.11

Table 1: PSNR (dB) values for all models on the dataset of [27]. Upscaling factor is 8× for all experiments. All results apart from the results of our models are taken from [27]. Red and blue markers indicate the first and second highest value, respectively.

son such that the super-resolved face does not look like the same identity. Our method, on the other hand, is completely faithful to the identity of the person while super-resolving the face image. Furthermore, GFRNet [27] most of the time outputs a super-resolution image that is blurry and that contains artifacts, whereas our method GWAInet produces sharp, visually appealing and photo-realistic results.

5. Conclusion

We proposed a novel solution, namely GWAInet, for the task of face image super-resolution. Our GWAInet utilizes the additional information provided by a high-resolution guiding image of the same person. Our network does not use facial landmarks during training and is capable to produce fine details for the whole face region in a uniform manner. Moreover, in the training, the employed identity loss further helps in preserving the identity related features by minimizing the distance between the embedding vectors of the super-resolved and HR ground truth images. GWAInet produces photo-realistic images in upscaling factor 8× and outperforms state-of-the-art in PSNR terms and also for perceptual quality of super-resolved images.

Acknowledgments

This work was partly supported by ETH Zurich General Fund (OK), Huawei, and a GPU grant from Nvidia.

References

[1] E. Agustsson and R. Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[2] M. Arjovsky and L. Bottou. Towards Principled Methods for Training Generative Adversarial Networks. *ArXiv e-prints*, Jan. 2017.

[3] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein GAN. *ArXiv e-prints*, Jan. 2017.

[4] Y. Bengio. Learning deep architectures for ai. *Found. Trends Mach. Learn.*, 2(1):1–127, Jan. 2009.

[5] Y. Blau, R. Mechrez, R. Timofte, T. Michaeli, and L. Zelnik-Manor. 2018 pirm challenge on perceptual image super-resolution. *arXiv preprint arXiv:1809.07517*, 2018.

[6] J. Bruna, P. Sprechmann, and Y. LeCun. Super-Resolution with Deep Convolutional Sufficient Statistics. *ArXiv e-prints*, Nov. 2015.

[7] Q. Cao, L. Lin, Y. Shi, X. Liang, and G. Li. Attention-Aware Face Hallucination via Deep Reinforcement Learning. *ArXiv e-prints*, Aug. 2017.

[8] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. VGGFace2: A dataset for recognising faces across pose and age. *ArXiv e-prints*, Oct. 2017.

[9] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01*, CVPR '05, pages 539–546, Washington, DC, USA, 2005. IEEE Computer Society.

[10] C. Dong, C. C. Loy, K. He, and X. Tang. Learning a deep convolutional network for image super-resolution. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 184–199, Cham, 2014. Springer International Publishing.

[11] C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(2):295–307, Feb. 2016.

[12] A. Dosovitskiy and T. Brox. Generating Images with Perceptual Similarity Metrics based on Deep Networks. *ArXiv e-prints*, Feb. 2016.

[13] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Networks. *ArXiv e-prints*, June 2014.

[14] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville. Improved Training of Wasserstein GANs. *ArXiv e-prints*, Mar. 2017.

[15] P. Gupta, P. Srivastava, S. Bhardwaj, and V. Bhateja. A modified psnr metric based on hvs for quality assessment of color images. In *2011 International Conference on Communication and Industrial Application*, pages 1–4, Dec 2011.

[16] 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)*, pages 1–8, July 2013.

[17] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. *ArXiv e-prints*, Dec. 2015.

[18] H. Huang, R. He, Z. Sun, and T. Tan. Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1698–1706, 2017.

[19] S. Ioffe and C. Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *ArXiv e-prints*, Feb. 2015.

[20] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-Image Translation with Conditional Adversarial Networks. *ArXiv e-prints*, Nov. 2016.

[21] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu. Spatial Transformer Networks. *ArXiv e-prints*, June 2015.

[22] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. *ArXiv e-prints*, Mar. 2016.

[23] J. Kim, J. K. Lee, and K. M. Lee. Accurate image super-resolution using very deep convolutional networks. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1646–1654, 2016.

[24] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. *ArXiv e-prints*, Dec. 2014.

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pages 1097–1105, USA, 2012. Curran Associates Inc.

[26] C. Ledig, L. Theis, F. Huszr, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 105–114, July 2017.

[27] X. Li, M. Liu, Y. Ye, W. Zuo, L. Lin, and R. Yang. Learning warped guidance for blind face restoration. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[28] X. Li, M. Liu, Y. Ye, W. Zuo, L. Lin, and R. Yang. Learning Warped Guidance for Blind Face Restoration. *ArXiv e-prints*, Apr. 2018.

[29] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee. Enhanced Deep Residual Networks for Single Image Super-Resolution. *ArXiv e-prints*, July 2017.

[30] 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.

[31] M. Mathieu, C. Couprie, and Y. LeCun. Deep multi-scale video prediction beyond mean square error. *ArXiv e-prints*, Nov. 2015.

[32] A. Radford, L. Metz, and S. Chintala. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *ArXiv e-prints*, Nov. 2015.

[33] 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. *ArXiv e-prints*, Sept. 2016.

[34] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[35] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for Simplicity: The All Convolutional Net. *ArXiv e-prints*, Dec. 2014.

[36] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[37] R. Timofte, E. Agustsson, L. V. Gool, M. Yang, L. Zhang, B. Lim, S. Son, H. Kim, S. Nah, K. M. Lee, X. Wang, Y. Tian, K. Yu, Y. Zhang, S. Wu, C. Dong, L. Lin, Y. Qiao, C. C. Loy, W. Bae, J. Yoo, Y. Han, J. C. Ye, J. Choi, M. Kim, Y. Fan, J. Yu, W. Han, D. Liu, H. Yu, Z. Wang, H. Shi, X. Wang, T. S. Huang, Y. Chen, K. Zhang, W. Zuo, Z. Tang, L. Luo, S. Li, M. Fu, L. Cao, W. Heng, G. Bui, T. Le, Y. Duan, D. Tao, R. Wang, X. Lin, J. Pang, J. Xu, Y. Zhao, X. Xu, J. Pan, D. Sun, Y. Zhang, X. Song, Y. Dai, X. Qin, X. Huynh, T. Guo, H. S. Mousavi, T. H. Vu, V. Monga, C. Cruz, K. Egiazarian, V. Katkovnik, R. Mehta, A. K. Jain, A. Agarwalla, C. V. S. Praveen, R. Zhou, H. Wen, C. Zhu, Z. Xia, Z. Wang, and Q. Guo. Ntire 2017 challenge on single image super-resolution: Methods and results. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1110–1121, July 2017.

[38] R. Timofte, V. De Smet, and L. Van Gool. Anchored neighborhood regression for fast example-based super-resolution. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2013.

[39] R. Timofte, V. De Smet, and L. Van Gool. A+: Adjusted anchored neighborhood regression for fast super-resolution. In *Asian Conference on Computer Vision*, pages 111–126. Springer, 2014.

[40] R. Timofte, S. Gu, J. Wu, and L. Van Gool. Ntire 2018 challenge on single image super-resolution: Methods and results. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[41] 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.

[42] Z. Wang, E. P. Simoncelli, and A. C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003*, volume 2, pages 1398–1402 Vol.2, Nov 2003.

[43] X. Xu, D. Sun, J. Pan, Y. Zhang, H. Pfister, and M. Yang. Learning to super-resolve blurry face and text images. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 251–260, Oct 2017.

[44] J. Yang, J. Wright, T. S. Huang, and Y. Ma. Image super-resolution via sparse representation. *IEEE Transactions on Image Processing*, 19(11):2861–2873, Nov 2010.

[45] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning Face Representation from Scratch. *ArXiv e-prints*, Nov. 2014.

[46] X. Yu and F. Porikli. Face hallucination with tiny unaligned images by transformative discriminative neural networks, 2017.

[47] X. Yu and F. M. Porikli. Ultra-resolving face images by discriminative generative networks. In *ECCV*, 2016.

[48] X. Yu and F. M. Porikli. Hallucinating very low-resolution unaligned and noisy face images by transformative discriminative autoencoders. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5367–5375, 2017.

[49] R. Zeyde, M. Elad, and M. Protter. On single image scale-up using sparse-representations. In *International conference on curves and surfaces*, pages 711–730. Springer, 2010.

[50] 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 *The European Conference on Computer Vision (ECCV)*, September 2018.

[51] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin. Learning face hallucination in the wild. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pages 3871–3877. AAAI Press, 2015.

[52] S. Zhu, S. Liu, C. C. Loy, and X. Tang. Deep cascaded bi-network for face hallucination. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision – ECCV 2016*, pages 614–630, Cham, 2016. Springer International Publishing.