Obfuscated Human Faces Reconstruction

Cecilia Casarella
cecilia.casarella@studenti.unipd.it

Sara Repetto
sara.repetto@studenti.unipd.it

Abstract

This paper examines the usage of residual neural networks to reconstruct obfuscated images of human faces. A slightly modified residual network was implemented by eliminating some modules, unnecessary for the property of the task. A Gaussian blur was employed to obfuscate images. The difficulty of the task was then increased by downsampling the input image by a factor two, thus requiring the model to perform a super resolution task. Our analysis focuses on the benefits of the addition of human face specific loss function, the attention function. As a baseline we employed the pixel-wise Mean Squared Error as a loss function which has been the state-of-the-art loss in previous super-resolution tasks. In the early stage of the analysis we investigated whether the employment of the perceptual loss , which compares high level features of images , could increase the performance of the model. This loss was then replaced by the facial attention loss which achieves better results as it restores facial attributes in greater details. Experimental results validate our hypothesis by showing that training the model with the attention loss leads to more accurate results in terms of reconstruction of facial details.

1. Introduction

Human Face Images Reconstruction aims at the restoration of facial details from low resolution, obfuscated human face images. Obfuscated human face images are widely employed in order to protect or hide people identities, however they may still contain enough details such that the original facial features can be reconstructed. On the other hand, deception of an image can depend on its spatial resolution which is directly linked to the number of pixels. An up-sampling procedure can augment images resolution and consequently the sharpness of the image details, but it is important to take into account that increasing the resolution of a low-resolved image can affect the real perception of the facial attributes. The final goal is to generate a high resolution image that is as similar as possible to the original high resolution image, while also being visually plausible and preserving important details. Human face images reconstruction is important for a variety of applications, including facial recognition and biometrics. It is also useful in forensic science and criminal investigations. Additionally, it can be used in medical imaging to help reconstruct the faces of patients who have suffered injuries. Overall, the ability to accurately reconstruct human faces in images can have significant practical and societal benefits. Our research aims at exploiting the architectural benefits of the EDSR (Enhanced Deep Super-resolution Network)[7], a popular CNN-based model frequently used in the super resolution field, by leveraging an human faces ad hoc training objective. Experimental results show that the introduction of a loss based on important facial landmarks achieves a better result than pixel-wise and perceptual losses. Code is available here

2. Related Work

There have been many different approaches to super resolution over the years, but recent advances in deep learning have led to a significant improvement in the quality of the results. One of the most popular approaches is to use a deep convolutional neural network (CNN) to learn the mapping from low to high resolution. This is typically done by training the network on a dataset of low resolution images and their corresponding high resolution images. One of the early successful works on super resolution task is the work by Dong et al. in 2014, they proposed a method called SRCNN which is a three-layer CNN architecture that learns the mapping between low and high resolution images.[2] Another popular approach is to use a generative model, such as a generative adversarial network (GAN), to generate high resolution images. The GAN architecture consists of two networks: a generator network that generates high resolution images, and a discriminator network that evaluates the generated images and tries to distinguish them from real high resolution images. This approach was first proposed by Ledig et al. in 2017 [6], in their work called SRGAN. The results of their method are visually more pleasing and closer to the ground truth images. In this same work Ledig et al. presented the attention loss for human faces landmarks which is exploited in our analysis. Another recent approach

is to use a pre-trained network, such as a pre-trained CNN, to extract features from the low resolution image, and then use these features to guide the generation of the high resolution image. This approach was first proposed by Lim et al.[7] in 2017, in their work called EDSR, they proposed an enhanced deep residual network for super resolution which uses pre-trained weights to extract features and then uses these features to guide the generation of the high resolution image. It uses a deep ResNet architecture to better recover high-frequency information in the image, which is often lost when decreasing the resolution. The results of this method are competitive with state-of-the-art methods and are faster to train.

3. Dataset

The "Labeled Faces in the Wild" (LFW) dataset is a dataset of face photographs designed for studying the problem of face recognition in unconstrained environments. The dataset was introduced in 2008 by Huang et al. [4] It contains 13,233 target face images of 5749 different individuals. Of these, 1680 people have two or more images in the database. The remaining 4069 people have just a single image in the database. The images are available as 250 by 250 pixel JPEG images. Most of them are colored images, although a few are gray scale only. The images vary in terms of pose, lighting, expression, and background, making it a challenging dataset for face recognition algorithms and the other tasks. The LFW dataset is available for use in PyTorch through the torchvision library. The torchvision library is a PyTorch library that provides tools for loading and transforming datasets, including the LFW dataset. By using the torchvision library, researchers can easily load and preprocess the LFW dataset in PyTorch, making it easier to train and evaluate algorithms. The pre processing phase of our analysis consisted on down-sampling the training images by a factor of two thus lowering their resolution and adding a general blurring effect which caused the obfuscation of the principal facial attributes. Gaussian filtering was used for smoothing the images by convolving with a Gaussian kernel 5x5 and parameter sigma at 2.6. From the resulting set of images 2048 images were extracted for training purposes , 500 for the validation set and 500 for the test set. These quantities were chosen accordingly to our computational capacity. Figure 1 displays some examples of the presented input to the network and expected target high resolution images.

4. Method

The following presented work is based on deep learning technologies for obfuscated image restoration. The overall idea is carried out as an end-to-end mapping between low-resolved blurred input images and super-resolved restored output images through a Convolutional Neural Network. We targeted the reconstruction task by testing different loss functions between the reconstructed output and the corresponding ground truth high resolved image.

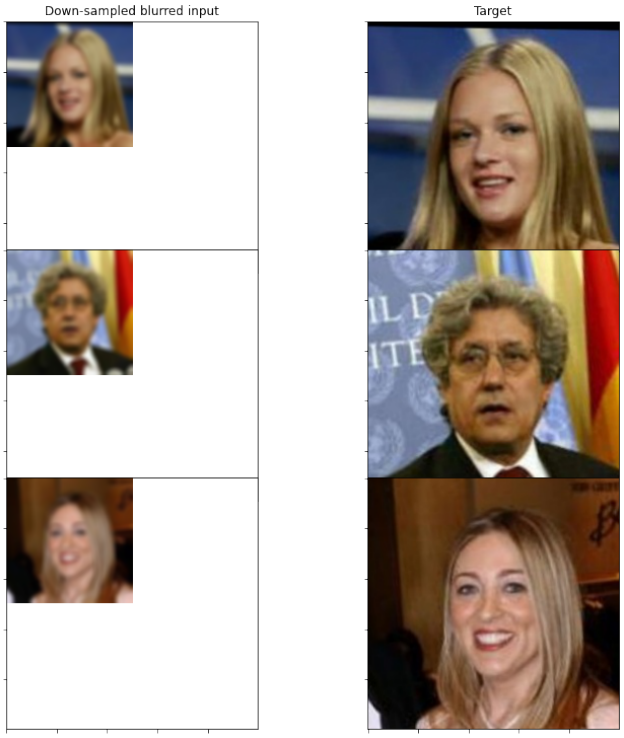

Figure 1. Example of low-resolved input images and high-resolved target images.

4.1. Architecture

A ResNet, short for Residual Network, is a type of convolutional neural network (CNN) that uses residual connections to allow for deeper networks without suffering from the vanishing gradient problem. The ResNet architecture from He et al. [3] was employed as a basic framework in order to reconstruct the images. As highlighted by Lim et al. [7] the original ResNet was proposed to solve higher-level computer vision problems such as image classification and detection. Applying ResNet architecture directly to low-level vision problems can lead to sub-optimal performance therefore some modules were removed. The network is trained to learn the mapping between the low-resolution and high-resolution images, and it uses the residual connections to better preserve the high-frequency information in the image during the up-scaling process. The architecture presented in Lim et al.[7] and showed in Figure 2 was adopted as an optimal solution in this research .

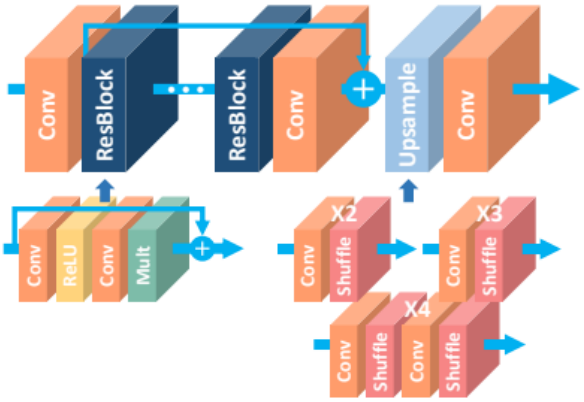

Figure 2. Archictecture of the proposed network (Lim et al).

4.2. Objective Function

Our project aims at assessing quantitative and quality results in training a leading model in the super resolution field with different loss functions. Results show how the addition of facial specific loss functions leads to better performance with respect to the baseline model.

4.2.1 Mean Squared Error

Pixel-wise Mean Square Error loss was used as baseline by following the experiments presented in [7]. It is computed as the sum of squared distances between each pixel value from the target image and each pixel value from the generated image. The pixel-wise MSE loss function can be expressed as follows:

$$l_{MSE} = \frac{1}{r^2WH}\sum_{x=1}^{rW}\sum_{y=1}^{rH}\left(I_{x,y}^{H\ R} - G_{\theta_G}\left(I^{LR}\right)x,y\right)^2$$

where G is the our network, I^{HR} and I^{LR} are target face images and input low-resolved images, respectively. W and H are width and height of the input image and r is the scale factor of the downsampling.

4.2.2 Perceptual Loss

Perceptual loss is used to measure the difference between an high-resolution image and the corresponding generated low-resolution image. The loss is calculated by comparing the high-resolution image to the generated image using a pre-trained deep neural network such as VGG-19 [8]. The neural network extracts features from both images and calculates the difference between the features, providing a measure of how similar the generated image is to the original high-resolution image. More specifically, it is computed as the Euclidean distance between the ReLU activation of the hidden layers of a pre-trained VGG-19 network applied to both the super resolved image and the target image. This loss function helps to improve the overall visual quality of the generated image by comparing high level features and make it more similar to the original high-resolution image.

Figure 3. Visualization of VGG-19 feature-space

4.2.3 Attention Loss

Facial attention loss was first introduced by Kim et al. [5]. It was employed to restore the attributes of the adjacent area to the facial landmarks. It was computed as a pixel-wise Mean of Absolute value of Errors (MAE) loss weighted by a facial component heatmap.

$$L_{\text{attention}} = \frac{1}{r^2WH}\sum_{x=1}^{rW}\sum_{y=1}^{rH}\left(M_{x,y}^{*}\cdot\left|I_{x,y}^{HR} - G\left(I^{LR}\right)_{x,y}\right|\right)$$

where M^* is heatmap of facial landmarks of target images. To construct the heatmaps, facial landmarks were initially identified using the Face Alignment Network by Bulat et al. [1]. An heatmap was then created as a matrix of the same size as the source image, taking value one in the indices associated to facial landmarks and zero elsewhere. A Gaussian filter was used to blur the heatmap to assure that the attention loss focuses on most elements that characterise the face even if they do not correspond to exact facial landmarks.

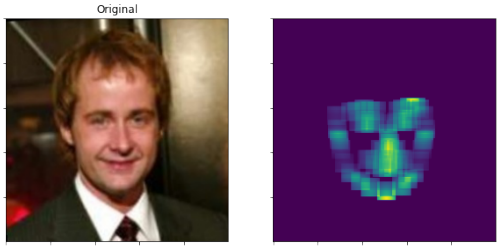

Figure 4. Visualization of the heatmap of facial landmarks

5. Experiments

Experiments were carried out on Google Colaboratory environment with default CPU hardware and GPU acceleration during training. Based on empirical results the optimal

batch sizes were found to be a multiple of 8, eventually 24 was chosen as a trade-off between computational efficiency and RAM storage capacity. In order to conduct comparable tests training with different loss functions is evaluated for 100 epochs each and Adam optimization algorithm. The learning curves as a result of the training process showed how Mean Squared Error loss easily reaches the minimum value in less epochs with respect to the others, while perceptual and attention loss require more computational time and effort given their intrinsic complexity. In order to verify our hypothesis evaluation is conducted by emulating the experiments in [7], thus using the same architecture and Mean Squared Error as loss function, and taking the results as baseline. Evaluation metrics and our quantitative and quality results are listed below.

5.1. Metrics

PSNR (Peak Signal to Noise Ratio) and SSIM (structural similarity index)[9] were used as quantitative metrics to measure the quality of the reconstruction. Both metrics evaluate similarity between images. Given $MAX\{I\}$ the maximum possible pixel value of the image, PSNR can be expressed as follows:

$$PSNR = 20 \cdot \log_{10} \left(\frac{MAX\{I\}}{\sqrt{MSE}} \right)$$

The higher the value of PSNR, the better will be the quality of the output image. By using SSIM the comparison between the two images is performed based on luminance, contrast and structure.

Performances were compared by using faces image patches. Table 1 shows final assessment of performance based on the previously mentioned metrics evaluated as the average on the test set.

Method	Average PSNR	Average SSIM
MSE	31.68	0.87
MSE + Perceptual Loss	31.46	0.88
MSE + Attention Loss	32.31	0.90

Table 1. Quantitative evaluation results in terms of average PSNR/SSIM on the test set.

Quantitative results show an higher average PSNR and SSIM when the attention loss function is implicated in training. Figure 5 presents qualitative results for a variety of human faces image patches as well as both colored and grey scale images. As a result of our training it was possible to generate high resolution face images from low resolved blurred ones. The first column in Figure 5 displays the original high resolution image used as target during training and final evaluation. The subsequent columns show the same image patches reconstructed by different models. The chosen baseline model, as already assessed in the Table 1, results in worse performances with respect to the other evaluated ones. The last column exhibits the best results both in terms of single-image and test set evaluation thanks to the employment of the attention function during training. Qualitative best results show an overall smoothness in the definition of the details as well as a better refinement.

6. Conclusion

Our research aims at exploiting the architectural benefits of the EDSR (Enhanced Deep Super-resolution Network) while introducing loss functions in such a way as to improve the performance of the model. Our experiments demonstrate that the introduction of an attention loss improves the performance of the model. Both attention and perceptual loss were introduced to allow the model to focus on more salient aspects of the image. However, while the perceptual loss focuses on general high-level features, the attention loss allows the model to focus specifically on elements of the face, thus guaranteeing a better reconstruction. Future extensions could concern the introduction of a new multi-scale model that efficiently reconstructs high-resolution images for various scales. The model would be more useful as it would allow good results for different types of input.

References

[1] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d &amp 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, oct 2017.

[2] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 184–199, Cham, 2014. Springer International Publishing.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[4] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[5] Deokyun Kim, Minseon Kim, Gihyun Kwon, and Dae-Shik Kim. Progressive face super-resolution via attention to facial landmark, 2019.

[6] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network, 2016.

[7] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution, 2017.

[8] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014.

[9] Zhou 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.

Figure 5. Qualitative results of patch comparison.