Two Studies on Convolutional Neural Network's Sensibility to Resolution

Antonio Augusto Abello

Thesis presented to the
Institute of Mathematics and Statistics
of the University of São Paulo
in partial fulfillment
of the requirements
for the degree of
Master of Science

Program: Computer Science
Advisor: Prof. Dr. Roberto Hirata Junior

During this work, the author was supported by CAPES

São Paulo
October 18, 2021

Two Studies on Convolutional Neural Network's Sensibility to Resolution

Antonio Augusto Abello

This version of the thesis includes the corrections and modifications suggested by the Examining Committee during the defense of the original version of the work, which took place on October 18, 2021.

A copy of the original version is available at the Institute of Mathematics and Statistics of the University of São Paulo.

Examining Committee:

- Prof. Dr. Roberto Hirata Junior (advisor) - IME-USP
- Prof. Dr. David Menotti - UFPR
- Prof. Dr. Zhangyang Wang - UTEXAS

Acknowledgements

First and foremost, I would like to acknowledge and thank the support and guidance provided by my supervisor, Professor Roberto Hirata Jr. Long and frequent conversations helped shape this work and steer it to the right directions.

Furthermore, I thank the other members of the Examining Committee of our qualifying exam, prof. Denis Deratani Mauá and prof. Zhangyang Wang, for the helpful comments and insight on our work. They were indispensable for making this thesis more mature and attuned to the state-of-the-art. I also take the liberty to thank in advance for the presence of the members of the Examining Committee of our Thesis Defence, prof. Zhangyang Wang and prof. David Menotti.

This work would not have been possible if not for the Vision Lab of IME-USP for several reasons. First for its infrastructure, which we used for most of the experiments. But the infrastructure alone would be useless if not for the continued maintenance work of IME's faculty and staff and the lab's volunteer admins, for whom I am profoundly thankful for. Last, but not least, for the friends and companions made on the physical lab, and the friendly and insightful conversations we were able to share.

Finally, I would like to thank the CAPES foundation for the financial support during my Masters, and my family for complementing this support when needed. I also could never have completed this journey without the emotional and psychological support from my family and friends.

Resumo

Antonio Augusto Abello. **Dois Estudos sobre a Sensibilidade de Redes Neurais Convolucionais à Resolução**. Dissertação (Mestrado). Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2021.

Redes Neurais Convolucionais (CNNs) recentemente se tornaram o estado-da-arte em várias áreas de Visão Computacional (CV). No entanto, por razões não completamente conhecidas, elas são bastante sensíveis à imagens de baixa resolução. Isso pode se tornar um problema para aplicações no mundo real, uma vez que para casos como o de vigilância ou direção automatizada nem sempre sensores de alta resolução podem ser utilizados. Neste trabalho conduzimos dois estudos sobre esse assunto: no primeiro estudamos empíricamente o efeito de perda de resolução e do uso de algoritmos de restauração de imagens em um modelo de Reconhecimento Facial (FR). No segundo, estudamos a hipótese do viés para altas frequências, uma das possíveis explicações para a sensibilidade de CNNs. No trabalho conseguimos desenvolver novas técnicas de restauração que ajudam melhor no problema de reconhecimento em baixa resolução e aprofundamos o entendimento atual sobre viés para altas frequências em CNNs.

Palavras-chave: Deep Learning. Super-Resolução. Reconhecimento Facial.

Abstract

Antonio Augusto Abello. **Two Studies on Convolutional Neural Network's Sensibility to Resolution:** . Thesis (Masters). Institute of Mathematics and Statistics, University of São Paulo, São Paulo, 2021.

Convolutional Neural Networks (CNNs) recently became the state-of-the-art for various Computer Vision tasks. However, for reasons not completely understood, they are very sensitive to low resolution images. This can be troublesome as real life applications such as automated driving or surveillance can not use high resolution sensors. In this work we perform two studies on this subject matter: on the first we empirically study the effect of resolution loss and image restoration algorithms on a Face Recognition model. On the second, we study the high frequency bias hypothesis, one of the current possible explanations for CNN's sensitivity. We are able to develop new techniques for image restoration that better deal with the low resolution recognition problem and advance the understanding of the high frequency bias in CNNs.

Keywords: Deep Learning. Super Resolution. Face Recognition

List of Abbreviations

DL	Deep Learning
NN	Neural Network
GAN	Generative Adversarial Network
SVM	Support Vector Machine
SISR	Single Image Super-Resolution
FR	Face Recognition
LFW	Labelled Faces in the Wild Dataset
YTFaces	YouTube Faces Dataset
IME	Instituto de Matemática e Estatística (*Mathematics and Statistics Institute*)
USP	Universidade de São Paulo (*University of São Paulo*)

List of Figures

2.1 Example of max pooling layer. The maximum value in each window prevails 5
2.2 Schematics of ResNet Layer. Reproduction He et al., 2016 9
2.3 5-layer dense block. Reproduction from G. Huang et al., 2017 9
2.4 Inception module. Reproduction from Szegedy, Wei Liu, et al., 2015 . . . 10

5.1 Testing Scheme for a SISR model 24
5.2 Scheme for training a SISR model on the FR loss 25
5.3 Comparison of results for an example downsampled in 8x 27
5.4 Comparison of results and specific inset 27

6.1 Energy Distribution Model Example 33
6.2 Example of distorted CIFAR10 images according to our model. Notice how color and edges are mixed in the first few bands, but the effect is barely noticeable in the last two 35
6.3 Example of distorted SVHN images according to our model. Notice how the less clear edges on rows 2 and 3 confuse even the human eye of the class of the digit 36
6.4 Example of distorted ImageNet images according to our model. Only five intervals amounting each 10% of total energy are shown for brevity. Notice how the effect is barely noticeable by the fifth interval, and how the fur texture is impoverished on the third and fourth intervals 37
6.5 Example of distorted VGGFaces2 images according to our model. Only five intervals amounting each 10% of total energy are shown for brevity . . . 38
6.6 Accuracy vs removed frequencies (frequencies not to scale) 40
6.7 Comparison of MSE of degraded images and decrease in performance . . 41
6.8 Effect of depth on different models trained on VGGFace2 42

List of Tables

4.1 Dataset Attributes 17

5.1 Results for intrinsic SISR evaluation on CelebA Test Split. The best results for each scale are bolded 28
5.2 kNN results using embeddings on CelebA Test Dataset. The best results for each scale are bolded 28
5.3 AUC and EER for embedding evaluation on face verification task on LFW. The best results for each scale are bolded 28

Contents

1 Introduction **1**

2 Fundamentals **3**

2.1 Digital Images . . . 3

2.2 Discrete Fourier Analysis . . . 4

2.3 Convolutional Neural Networks . . . 4

2.3.1 Image Classification . . . 6

2.3.2 Face Recognition . . . 7

2.3.3 Single Image Super Resolution . . . 7

2.4 CNN Architectures . . . 8

3 Literature Review **11**

3.1 Low Resolution Recognition . . . 11

3.2 Single Image Super Resolution . . . 13

3.2.1 Network Architectures . . . 13

3.2.2 Loss Functions . . . 13

3.2.3 Multi-task Learning and others . . . 14

3.2.4 Insufficiency of Current Metrics . . . 14

3.3 Deep Face Recognition . . . 15

3.4 High Frequency Bias in CNNs . . . 15

4 Experimental Basis **17**

4.1 Datasets . . . 17

4.2 Metrics . . . 18

4.2.1 FR Metrics . . . 18

4.2.2 SISR Metrics . . . 19

4.3 Implementation Details . . . 21

5 Optimizing Super Resolution for Face Recognition **23**

5.1 Task-Based Evaluation 24
5.2 FR Loss 25
5.3 Experimental Results 26
5.4 Discussion 28

6 Dissecting the High-Frequency Bias in Convolutional Neural Networks 31
6.1 Method 32
6.1.1 Frequency Importance 32
6.1.2 Energy Distribution Model 32
6.1.3 Implementation Details 33
6.1.4 Robust and Non-Robust Features 34
6.2 Experimental Results 39
6.3 Discussion 43

7 Conclusions 45
7.1 Optimizing Super Resolution For Face Recognition 45
7.2 Dissecting the High Frequency Bias in Convolutional Neural Networks . 46

Appendices

Annexes

A Article Published and Presented in SIBGRAPI 2019 **49**

B Article Published in Bridging the Gap Between Computational Photography And Visual Recognition 2021 **59**

Chapter 1

Introduction

Recently on the field of Computer Vision, Convolutional Neural Networks (CNNs) have achieved great results on various tasks, becoming the state-of-the-art for image recognition, restoration, generation, among others. However, low resolution recognition has increasingly become a problem. Low image resolution has a great effect on CNNs performance Dai et al., 2016 Dodge and Karam, 2016. At the same time industrial scaled supervised datasets are comprised mostly of human curated, high quality data while most applications, such as surveillance or automated driving, for example, use lower resolution cameras VidalMata et al., 2019 P. Li et al., 2019. In this work we perform two studies exploring respectively the practical and theoretical aspects of this problem.

On our first study, we evaluate how a CNN-based Face Recognition (FR) model reacts to resolution loss and restoration. It is known that image restoration techniques can lighten the issue of low-resolution recognition Dai et al., 2016. Nevertheless, a significant performance gap persists between using low resolution images, albeit restored, and high resolution ones Zhangyang Wang et al., 2016. We propose, then, to use the change of performance between low resolution and restored images as a proxy for the quality of restoration. Based on this idea, we also develop a new technique for training restoration models optimized for helping face recognition.

On the other hand, researchers have shown that CNNs are much more sensitive to resolution than humans Hendrycks and Dietterich, 2018, Dodge and Karam, 2017. This sensitivity, which expands to other degradations, even barely perceptible ones Szegedy, Zaremba, et al., 2013 is a characteristic of CNNs that is currently not well understood. A promising group of research tries to explain it through the existence of a bias of CNNs towards visual features which are highly informative, but also non-robust Ilyas et al., 2019. A common hypothesis is that these features consist of imperceptible patterns that lie on the high frequency part of the Fourier spectrum of the image Jo and Bengio, 2017 H. Wang et al., 2020 or are parts of the image that have a strong high frequency component, such as texture Geirhos et al., 2018.

This hypothesis is particularly interesting to the low-resolution recognition problem because lower resolution images have smaller sample rates and therefore, by the Shannon Sampling Theorem Broughton and Bryan, 2009, cannot represent precisely the high

frequency parts a higher resolution image can. We thus studied this hypothesis further hoping to understand more the causes of CNNs sensitivity to resolution. From our study of the literature, we found that the papers that had found evidence towards this hypothesis used a definition of "high" and "low" frequencies that was not rigorous, most of the time hand-adjusted. We developed then a way to divide the frequency spectrum in a fair, methodical way, and applied it to a variety of scenarios for a systematic study of the high-frequency bias.

Our works achieved interesting results, which were published in two papers: Antonio Augusto Abello and Hirata, 2019 and Antonio A Abello et al., 2021. We were able to reproduce the results of the literature in both cases. Our method for SISR training improved low-resolution recognition relative to standard SISR training. By expanding the tested scenarios and using our method for dividing the frequency spectrum, we were able to get a more nuanced view of the high-frequency bias. Overall we did not completely solve the low-resolution recognition problem, but we were able to alleviate and understand it further than before.

The rest of this work is divided as follows: we begin with a fundamentals Chapter (2), which defines notation and gives the basic background necessary for the following chapters. We then consolidate our literature review in 3. As both studies have much of their experimental basis in common, we describe it in Chapter 4 once to avoid redundancy. Each study is then presented in Chapter 5 and Chapter 6. Finally, we conclude with a summary of the work along with closing comments and possible future directions of research in Chapter 7. Both of our published papers are included on the annex part of this work for reference.

Chapter 2

Fundamentals

This chapter presents the theoretical basis necessary for understanding the rest of this work. It also introduces some notation that will be used throughout it. The chapter is based mostly on three books, each for one of it's sections: Klette and Rosenfeld, 2004 for the discussion of digital images, Broughton and Bryan, 2009 for Discrete Fourier Analysis and I. Goodfellow, Bengio, et al., 2016 for CNNs and Deep Learning

2.1 Digital Images

In this work when we refer to images we mean digital images, that is, a picture obtained through a process of digitization Klette and Rosenfeld, 2004 that represents an image as a 2 dimensional grid of pixels. We will represent an image with $n \times m$ pixels as a matrix $X \in R^{n \times m}$, in which each element $X[p, q], p \in \mathbb{N}, p \in [0, N-1], q \in \mathbb{N}, q \in [0, M-1]$ represents the intensity of the pixel located in p, q. For multi-channel images, such as RGB ones, each channel is represented by an individual matrix for the intensity of each color, and the image as a whole is the array of matrices.

The image resolution of a digital image is a complex concept, with various facets. Klette and Rosenfeld, 2004 (Section 1.1.2). It can refer to a parameter of the display medium, defined by the number of dots or pixels per inch of the display, or a parameter of the digitization process, referring to the amount of samples per inch. It can also mean the overall size of the image, measured by the amount of pixels, or the optical resolution, the ability to resolve the objects or details present on the image.

These two latter meanings are the ones we are interested in on this work, and the ones we will mean when we refer to resolution. They are closely related by the Shannon Sampling Theorem Broughton and Bryan, 2009 (Section 1.6.4). Considering that any signal can be expressed as a sum of basic waveforms with different frequencies, the theorem states that when an analog, one-dimensional signal is sampled at a rate of N, it can only be perfectly recovered if it contains no frequencies larger than $\frac{N}{2}$. An analogue result exists for two-dimensional signals, such as images Petersen and Middleton, 1962. This implies that images with smaller sizes (sample rates) are not able to represent higher frequencies well, thus impairing their optical resolution. The converse, however, is not true, as an

image with low optical resolution not necessarily will have a smaller image size.

2.2 Discrete Fourier Analysis

The Discrete Fourier Transform (DFT) analyses any discrete signal into a sum of basic waveforms Broughton and Bryan, 2009, and synthesize a sum of waveforms into a discrete signal through its inverse. This manner of expressing a signal is also called the *frequency domain* representation. This can be useful for performing specific computations and for conceptual analysis of the signal.

Thus, when we refer to the Fourier transform of an image and to its inverse, we are actually referring to the DFT. More specifically, the DFT is an operator $\mathcal{F} : R^{N \times M} \rightarrow C^{N \times M}$ such that, given a matrix X, results a complex valued matrix Y, given by Equation 2.1

$$Y[k,l] = \frac{1}{N * M} \sum_{p=0}^{N-1} \sum_{q=0}^{M-1} X[p,q] e^{-2\pi i (\frac{kp}{N} + \frac{lq}{M})} \tag{2.1}$$

For each pair (k, l) representing a frequency, we call the magnitude of the complex coefficient of that frequency the energy contributed by $Y[k, l]$. Notice that for the special case of $Y[0, 0]$ we have not a waveform, but instead the average intensity of the matrix X.

Given the Nyquist sampling rate Broughton and Bryan, 2009 (Section 1.6.4), frequencies above $\frac{N}{2}$ are aliased to negative frequency terms. Thus k and l can also be written in the range of $\frac{-N}{2} \leq k \leq \frac{N}{2}$, $\frac{-M}{2} \leq l \leq \frac{M}{2}$. This representation "centers" the zeroth frequency at the middle, and thus allows for a grouping of "low", "middle", or "high" frequencies, according to their distance to the center. This division can be conceptually useful as, for example higher frequencies tend to contain noise and can be filtered for a smoother representation of the signal Broughton and Bryan, 2009 (Section 2.3.5).

2.3 Convolutional Neural Networks

A Convolutional Neural Network (CNN) is a kind of Deep Learning model used for Computer Vision. Although most of its theory was already developed in the 90s LeCun, Bengio, et al., 1995 it gained attention and became the state-of-the-art only in the early 10s Krizhevsky et al., 2012, when the resources for training larger models became available. We present here a brief introduction of CNNs based off the contents of I. Goodfellow, Bengio, et al., 2016 and the course notes of F.-F. Li et al., 2015 [1].

As a Deep Learning model, a CNN is a model composed of multiple layers which process data individually and whose inputs and outputs are chained together in a graph I. Goodfellow, Bengio, et al., 2016 (Chapter 6). What distinguishes CNNs are the particular kinds of layers they are composed of, namely convolutional, pooling, flattening and

[1] Available at https://cs231n.github.io/

fully connected layers, among others F.-F. Li et al., 2015 (Lecture: Convolutional Neural Networks: Architectures, Convolution and Pooling Layers).

A convolutional layer receives as input a matrix, or array of matrices (which can represent a digital image as discussed above) and performs a linear operation called convolution. For a single-channel image $X \in R^{n \times m}$ and a kernel $K \in R^{i,j}$, $i << n, j << m$, the convolution output, $O \in R^{n \times m}$ is written on Equation 2.2

$$O(i,j) = (K * X)(i,j) = \sum_m \sum_n X(i+m, j+n) K(m,n) \tag{2.2}$$

When dealing with multi-channel images, the operation is done channel-wise and the result is summed. The kernel K is a changeable parameter of the network that can be adjusted to the data. The set of kernels of a network is also referred to as its *weights*. After the convolution operation, a non-linear activation is applied element-wise with the purpose of allowing the network to represent non- linear functions F.-F. Li et al., 2015 (Lecture: Neural Networks pt.1).

I. Goodfellow, Bengio, et al., 2016 gives three main motivations for the use of convolutions (Section 9,2). First, as the kernel is smaller than the input image, it leads to fewer parameters and fewer memory and processing footprints (what they call **sparse connectivity**). Secondly, the fact that the same kernel is applied to different sections of the image leads allows the network to learn one set of independent parameters instead of various parameters for each location on the image (**parameter sharing**). Finally, convolution is **equivariant** to translation, which means that if an image is shifted, the output of its convolution by a particular kernel will also be shifted. This means that once the network is able to detect an interesting feature, it can detect it wherever it is on an image.

A pooling layer is a transformation that lowers the dimension of its input, as one can see in Figure 2.1. Generally convolution and pooling layers will be interspersed together, the latter serving to progressively summarize and refine the network's output I. Goodfellow, Bengio, et al., 2016 (Section 9.3). One can use different functions, such as taking the maximum value of a neighborhood of pixels, the average or the L^2 norm.

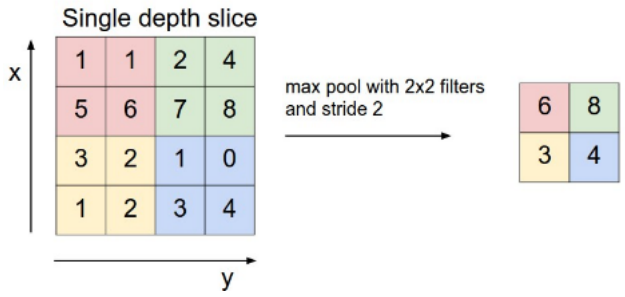

Figure 2.1: *Example of max pooling layer. The maximum value in each window prevails*

If the desired output of the function is one-dimensional, a flattening layer can convert the 2D data to 1D by concatenating rows and channels into a vector. Fully connected or "Dense" layers can further process one-dimensional data by performing a matrix multiplication along with a non-linear activation F.-F. Li et al., 2015 (Lecture: Neural Networks

Part 1). In this case, the matrix used for multiplication is the learnable parameter.

To determine the value of the learnable parameters, CNNs are optimized to minimize a certain cost-function. As we will describe below, this cost-function is tailored to each application, and it can be the direct final metric we are interested in (for example, the reconstruction loss for Single Image Super Resolution) or a cost-function that is related to the metric (for example, the negative log-likelihood for image classification, related to the classification accuracy) I. Goodfellow, Bengio, et al., 2016 (Section 8.1). Generally, the cost-function is defined for a single example and then averaged over a batch of examples or a whole dataset.

Due to CNN's non-linear nature, this optimization is done using gradient-based methods F.-F. Li et al., 2015 (Lecture: Optimization: Stochastic Gradient Descent). This is possible because an important common feature among all layers of neural networks is that they are all differentiable (or at least semi-differentiable). F.-F. Li et al., 2015 (Lecture: Backpropagation, Intuitions). The network format allows for the quick calculation of partial derivatives through iterative applications of the chain rule, an algorithm called backpropagation LeCun, Touresky, et al., 1988.

The specific order and combination of layers in a CNN is called it's architecture. F.-F. Li et al., 2015 (Lecture: Neural Networks Part 1: Setting up the Architecture). Different applications need different layers and architectures of CNNs, along with properly defined cost-functions. Next, we will describe how CNNs can be adapted to different Computer Vision tasks. The list of tasks is not exhaustive, but focus on tasks important to our work.

2.3.1 Image Classification

Classification is the most traditional ML application. Given a collection of images ($x \in X$) associated to a discrete collection of labels, also called classes ($y \in [0, ...n]$), we want to train a CNN, f, to be able to correctly associate previously unseen examples to their correct label.

To approach this problem, we use a CNN as an estimator of the conditional probability of an example belonging to each class given its input I. Goodfellow, Bengio, et al., 2016 (Section 8.1.2). The cost-function minimized to achieve this is the negative log-likelihood, also called the cross-entropy loss, depicted on Equation 2.3.

$$\mathcal{L}(y, f(x)) = -y \log f(x) \tag{2.3}$$

To ensure that the CNN's output is in fact a probability distribution, we use a flattening layer to make the network's output one-dimensional and in the final layer we use a special activation function for the output called softmax. For a given vector, $z \in R^n$, the softmax activation $\sigma(z)$ is defined by Equation 2.4

$$\sigma(z) = \frac{e^z}{\sum_{j=1}^{N} e^{z_j}}, z \in \mathcal{R}^N \tag{2.4}$$

2.3.2 Face Recognition

In this work, we will refer to Face Recognition (FR) as an umbrella term for a series of tasks related to face images, such as Face Verification and Face Identification. Face Verification is the simpler binary task of deciding whether two images belong to the same person Phillips, P. Grother, et al., 2003

On a Face Identification problem the model has a gallery of known identities, each containing one or more face images, and an unseen image, called a probe, with a face image. The task consists of successfully matching the probe to one of the identities of the model's gallery, or correctly rejecting the example when the face does not belong to any of them. Generally, the model will output a list of potential candidates, ordered by how likely the model considers each candidate to be the person on the probe P. J. Grother et al., 2018. We call an FR model a network that can be trained once and then adapted to perform both of these tasks.

Taigman et al., 2014 developed a framework for using CNNs as FR models that quickly became the state-of-the-art P. J. Grother et al., 2018. To do so, they trained a CNN as a identity classifier in large-scale datasets of human faces. Then, the output of an intermediate layer was used as an embedding in R^n for human facial images. Similar faces, even when not seen before by the models, tended to be closer together on the embedding space, and different faces tended to be more distance.

Diverse FR problems can then be solved by using these learned embeddings, as in Weiyang Liu et al., 2017. They solved Face Verification by using a distance metric and a threshold after which two images are considered of the same person. For Face Identification they used a k-Nearest Neighbors classifier on the embedding space.

2.3.3 Single Image Super Resolution

Single Image Super Resolution (SISR) is the task of inferring a high-resolution (HR) image from a single degraded, low-resolution (LR) one C.-Y. Yang et al., 2014. As noted by Shi et al., 2016, this degradation process destroys information, particularly high-frequency information. This, according to them, makes it so that a single LR image can be generated from various HR counterparts, making the whole problem of SISR highly ill-posed.

CNNs can easily be adapted to produce an image as output. In fact, the convolution operation already produces an image as an output, as we have seen on the discussion about convolutional layers. Thus, if we simply do not use pooling or flattening layers, the output of the CNN will already be an image. Initially, CNNs used as input a LR image resized through interpolationDong, Loy, He, et al., 2014, but later on special layers that increased the image size such as the deconvolution layer Dong, Loy, and Tang, 2016 or the sub-pixel convolution layer Shi et al., 2016 were preferred.

Using a degradation model $\phi : X \to X$, we can artificially degrade a high-resolution image into a lower-resolution one and have the CNN learn the inverse process, outputting a restored image, $f : X \to X \simeq \phi^{-1}$. Both the degradation model and the CNN can be made to preserve the image's dimensions, allowing us to express them as image transformations with same domain and codomain. In SISR the most common degradation model is to

simply downsample the image, preceded by some kind of antialiasing filter. To go back to the original image dimension, methods such as bicubic interpolation are used after the downsample operation.

The CNN's restored image can be compared with the original using a distance function, such as the Euclidean norm Dong, Loy, He, et al., 2014. In this context, it is commonly referred to as the pixel-by-pixel Mean Square Error (MSE). The CNN is then trained to minimize this distance as defined by Equation 2.6.

$$x' = \phi(x), x \in X \tag{2.5}$$

$$\mathcal{L}(x, x') = ||x - f(x')||_2 \tag{2.6}$$

Given different degradations and different degradation models, a similar approach might be used for different tasks of image processing and image restoration, such as de-noising Jain and Seung, 2008 or de-hazing B. Li et al., 2017, for example. For SISR the degradation model usually consists of a down-sampling operation preceded by a low-pass filter, that avoids aliasing Dong, Loy, He, et al., 2014.

2.4 CNN Architectures

Although CNN architectures can be adjusted for individual tasks, as we have shown above, there have been some improvements that found universal adoption. They usually involve changes in information and gradient flow, with the objective of training ever deeper networks. We present a tentative chronological history of the development of these improvements in architecture here, while focusing on the ones we explored on this work and not aiming for completeness.

The VGG network (Simonyan and Zisserman, 2014) improved upon the original AlexNet (Krizhevsky et al., 2012) by effectively building and training much deeper networks than the state-of-the-art. They found a connection between network depth and performance that did not seem to diminish in returns, being hindered mostly by hardware limitations at the time. To do so, they also removed convolutions with larger convolution filters, standardizing all of them to the smallest odd size (3x3).

He et al., 2016 observed that deeper VGG-like networks started to perform worse after a certain depth threshold. They proposed that this was not due to overfitting, though, but by optimization difficulties, such as gradients becoming too small before they get to the first layers. To solve this, they developed Deep Residual Networks, or ResNets He et al., 2016. In this architecture a layer, or group of layers, learn the residual to be added to it's input instead of the actual transformation from input to output. This can be seen in Figure 2.2

With this layer it was trivial for a deeper net to simulate a more shallow one: have some of it's layers map all inputs to zero. Adding more layers would then be, in the worst

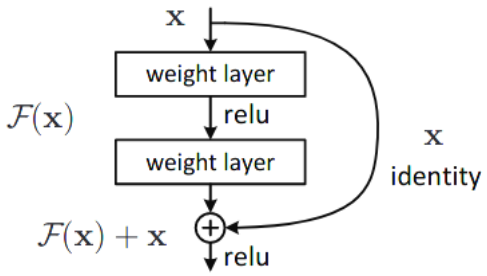

Figure 2.2: *Schematics of ResNet Layer. Reproduction He et al., 2016*

case scenario, irrelevant. They experimentally found that ResNets had better gradient flow and trained easier, becoming the then state-of-the-art quickly.

DenseNet (G. Huang et al., 2017) was developed as an improvement upon ResNets. Huang et al introduced dense blocks of layers, groups in which every residual layer receives as input all outputs from previous layers concatenated in the channel axis. An example can be seen in Figure 2.3. This would again facilitate gradient flow, but also allow for feature reuse and thus less redundancy. They found that this architecture not only facilitated the training of deeper, and thus better networks, but also that it was more efficient in parameter and computation time, as a smaller number of filters achieved an equivalent representation power of a ResNet.

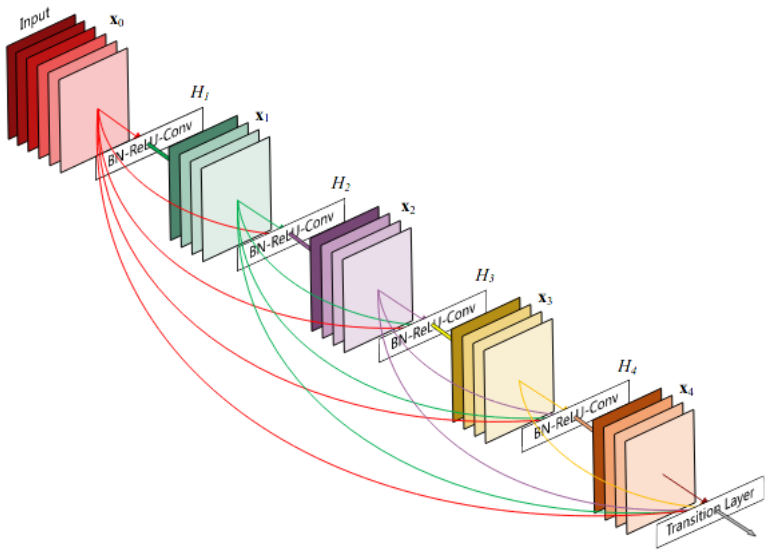

Figure 2.3: *5-layer dense block. Reproduction from G. Huang et al., 2017*

In parallel with the development of ResNets, an alternative to the simple (3x3) convolution layer, the Inception block, was developed Szegedy, Wei Liu, et al., 2015. In order to take advantage of different scaled details, this block branches into differently sized convolutions that are depth-wise concatenated at the end. This can be seen in Figure 2.4. The idea of an Inception block was iteratively improved, eventually leading to the merge of ResNets and Inception blocks, the InceptionResNet Szegedy, Ioffe, et al., 2017

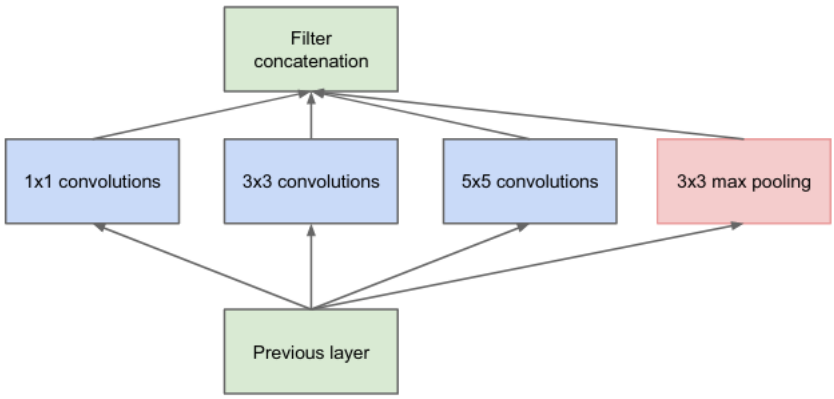

Figure 2.4: *Inception module. Reproduction from Szegedy, Wei Liu, et al., 2015*

Chapter 3

Literature Review

In this chapter we perform a literature review of the main areas of research relevant to our studies. We begin with a section about the problem of Low Resolution Recognition in general. After that we review the state-of-the-art of Single Image Super Resolution, one of the possible approaches to solve the Low Resolution Recognition problem. We end with a section on the High Frequency Bias hypothesis, a possible explanation for the magnitude of the impact of resolution on CNNs.

3.1 Low Resolution Recognition

It is well-known that low resolution images hinder performance of CV models. Studies from as early as 2009 show that low resolution is a problem for face recognition models. Lui et al., 2009 compares different models of the then state-of-the-art and finds that, while increasing resolution may not have much effect, when resolution drops below a threshold performance decreases sharply. Hu et al., 2012 produced a similar work with similar conclusions, focusing on surveillance applications. Dai et al., 2016 expands to a broader selection of CV tasks, such as border detection, digit classification and segmentation.

CNNs also seem to be affected similarly by resolution. Considering Gaussian blur as an appropriate proxy for resolution loss, Dodge and Karam, 2016 found that CNNs get progressively worse performances as blur increases, when testing traditional image classifiers. On a subsequent study (Dodge and Karam, 2017), they have also found that this degradation is much stronger than the one human subjects present. This was further experimentally verified by Hendrycks and Dietterich, 2018.

A natural idea for tackling this issue is to use Single Image Super Resolution (SISR) models as a pre-processing step, in order to first recover a higher quality image before performing some other CV task on it. This idea was tested a number of times in different contexts. Zou and Yuen, 2011 used then state-of-the-art face SR algorithms and found, at best, a slight increase in performance. Zhangyang Wang et al., 2016 furthers this study using deep neural networks to find a similar result for object classification. Finally, Dai et al., 2016 uses CNN-based SISR to achieve a consistent increase in performance across different tasks. In all of these studies, while there is a gain in using simple SR, the final

performances were closer to the ones on LR images than on the HR ones.

This performance gap motivated alternative methods. A first idea still involves two separate models, one for image restoring and other for recognition, but tries to associate the SISR models with the downstream task or a similar one. Haris, Greg Shakhnarovich, et al., 2018, for example, trains a SISR network to minimize the error of an object-detection neural network, leading to a result that is not visually pleasing for humans, but is more efficient for neural networks to process. This is also an idea we decided to explore, which we describe with more detail in Chapter 5.

K. Zhang et al., 2018 trains a SISR model collaboratively with a FR model. Although they do not share weights, the SISR model is trained to maximize recognition on the FR model and the FR model is trained to recognize images generated by the SISR model at the same time. This leads to a SISR model that allegedly resolves more identity-defining information and to an FR model robust to resolution differences.

D. Liu et al., 2019 used the weights of a SISR model as the initial point for the first layers of a larger classification model. This improved upon the idea of using SISR as a preprocessing step by letting the SISR model be adjusted along with the classification one during training

Another common approach is to train models that learn features which are invariant to resolution changes. Zhangyang Wang et al., 2016 used a domain adaptation approach with partially coupled networks, in which two networks trained on two different resolutions would share some weights during training. Lu et al., 2018 and Zangeneh et al., 2020 use CNNs to learn a coupled mapping in which similar images are close together but also the same image on different resolutions have approximately the same mapping.

When compared to the two-step (restoration, then recognition) approach, these more integrated methods achieved better results, and seem to represent the most promising direction for low-resolution recognition. Zhangyang Wang et al., 2016 and Zangeneh et al., 2020 explicitly compared their methods with SR ones, outperforming them. P. Li et al., 2019 surveys the field of low-resolution recognition and tests various methods on real-life LR images. They find that integrated methods outperformed SR ones on these situations as well.

Recently, some researchers have drawn attention to some limitations in most low resolution recognition studies which may hinder their real life applicability. Bulat, J. Yang, et al., 2018 argued that the degradation model commonly used for producing synthetic LR images may be overly simplistic, proposing a GAN-based method for learning a more accurate one. P. Li et al., 2019 notes that most studies test their methods on artificially degraded datasets. They perform a survey-like study of various techniques for low resolution classification such as SISR methods, domain adaptation, Siamese Networks and others on "in-the-wild", real LR datasets, getting far less optimistic results than their original papers.

It was also noted that the kind of degradation affecting the input images may not be known beforehand, may not be uniform among the data and may even be a combination of different kinds of degradation. In 2018, for example, the UG2 challengeVidalMata et al., 2019 was hosted at CVPR to aggregate and experimentally test various methods for enhancing both human and automated classification of images from different sources

(satellite, drone-based and ground-based) suffering from uneven and unknown distortions. While several teams from various top-level universities and laboratories participated, no proposed method was able to alleviate the image quality problems for all scenarios tested uniformly, even though some achieved success on specific cases.

3.2 Single Image Super Resolution

Different approaches for SISR were proposed, based on interpolation rules, image statistics, sparse-coding and others (C.-Y. Yang et al., 2014). However, since the introduction of the SRCNN (Super Resolution Convolutional Neural Network) by Dong, Loy, He, et al., 2014, Deep Learning based methods for SISR quickly surpassed and became the state-of-the-art.

The specific subset of super-resolution of human faces is called face super-resolution, or sometimes face hallucination when degradation is high enough. In general, the same super resolution methods apply normally to the domain of faces. But there have also been attempts to exploit specific properties of these kind of images such as in Y. Chen et al., 2018 or Bulat and Tzimiropoulos, 2018, better described below.

Further developments on Deep Learning based SISR are innumerable, and we refer to Zhihao Wang et al., 2019 for an in-depth survey of them. However, we would still like to single out a few crucial developments that will be of more importance to our work.

3.2.1 Network Architectures

Dong, Loy, and Tang, 2016 further developed their idea by adding a transposed convolution layer at the end of SRCNN that upsampled the input through a learnable operation, instead of using as input an interpolated version of the low-resolution image. Shi et al., 2016 developed a similar idea, using instead a sub-pixel convolution layer. Later on, Lai et al., 2017 would propose LapSRN (Laplacian Pyramid Super Resolution Network), that progressively upsampled images throughout the network, generating intermediate results for multiple scales of resolution. Finally, Haris, Gregory Shakhnarovich, et al., 2018 developed a network inspired on the backprojection procedure that iteratively upsamples and downsamples the input multiple times.

Since 2014 the SISR field also followed some of the general trends of DL network architecture in general. Kim et al., 2016a successfully trained Very Deep Super Resolution (VDSR) networks and Ledig et al., 2017 brought the Residual Network architecture to SISR with SRResNet. There have also been some architectural improvements specific to the Super Resolution problem too. Kim et al., 2016b proposed a Deeply Recursive Convolutional Network (DRCN), in which the same convolutional filter is applied repeatedly to its input. This can greatly increase the receptive field of each filter.

3.2.2 Loss Functions

Another source of improvements for Deep Learning-based SISR comes from improving and designing new loss functions that bring additional information during training. Lai

et al., 2017 argue that the L2-loss is sensitive to outliers, and prefer to use a robust differentiable version of the L1 loss called Charbonnier loss. Ledig et al., 2017, Johnson et al., 2016 and H. Huang et al., 2017 argue that minimizing exclusively a pixel-by-pixel loss function may miss more broad contextual information, and lead to results that are over-smooth and lacking in high-frequency detail. Each of them proposed different loss functions to address this issue.

Johnson et al., 2016 proposed a "perceptual loss", based on the activation from an intermediate layer of a VGG16 (Simonyan and Zisserman, 2014) pre-trained for object classification. The authors argue that in the process of learning how to classify objects, the network must have learned an intermediate representation that depends on a broader context than a single pixel, thus containing "perceptual" information. Ledig et al., 2017 applied the GAN architecture (I. Goodfellow, Pouget-Abadie, et al., 2014) to Super-Resolution, in which two networks are trained in an adversarial game. A discriminator network is trained to distinguish between super-resolved and original networks, while a generator network is trained to minimize the discriminator's certainty.

3.2.3 Multi-task Learning and others

A final approach for improving SISR models is to leverage information contained or used in other tasks to aid in resolving high-frequency details. A common way to do this is through multi-task learning. Bulat and Tzimiropoulos, 2018 train a network that performs super-resolution and facial landmark estimation at the same time, sharing an internal representation.

Another way to use different tasks to aid super-resolution is to have a sub-network that predicts relevant information pertaining to another task. Y. Chen et al., 2018's architecture includes a sub-network that performs face parsing. The inferred face maps are then fed to another part of the network that performs super-resolution. The network as a whole is trained then to minimize the reconstruction error and the face parsing error. This idea was most likely inspired by H. Huang et al., 2017, which proposes a similar model based on wavelet transformations. Simple, but not differentiable, wavelets transformations specifically chosen to represent facial high-frequency details are first predicted by the network and then used to reconstruct the image.

3.2.4 Insufficiency of Current Metrics

The intrinsic metrics commonly used in SISR will be properly defined in Chapter 4. We would like to point out, though, a common trend perceived in our literature review. The two most common metrics used, namely PSNR and SSIM, are generally deemed insufficient to properly evaluate image quality (Huynh-Thu and Ghanbari, 2008, Hanhart et al., 2013). It is hypothesized that this is due to the fact that both of these metrics are calculated pixel-by-pixel or patch-by-patch, ignoring then more global and contextual information Zhihao Wang et al., 2019. It is common for new methods to exceed the state-of-the-art qualitatively at the cost of a loss in these quantitative metrics, such as in Johnson et al., 2016 or Ledig et al., 2017. These works will then turn to opinion scores in order to justify the qualitative intuition of the superiority of their method. In the case of GANs, it is

common to present two versions of the final model: one trained solely on the MSE and the other trained with GANs, as the former will perform better quantitatively and the latter will perform better qualitatively or on opinion scores (Ledig et al., 2017, Y. Chen et al., 2018).

3.3 Deep Face Recognition

Face Recognition is a more marginal theme of this work. Nevertheless we felt the need to include a very brief literature review on the subject for proper contextualization

Before the advent of Deep Learning there was a significant gap between automated face recognition systems and human performance. Taigman et al., 2014 used Deep Face Recognition to close this gap and achieve competitive results. Their approach consisted of a CNN trained as a multi-class classifier on a closed set of face identities. The activation of an intermediate layer was then used as a representation of the face image. The authors trained different models using these embeddings as input, but even simple metrics such as the L_2 distance already surpassed by large previous results. Since then, CNNs quickly became the state-of-the-art for Face Recognition

As in Super-Resolution, there are innumerable improvements made upon Taigman et al., 2014's initial model. Parkhi et al., 2015 devised a training method based on collections of trios of an anchor, positive and negative images that could be used to leverage large-scale datasets whenever multi-class classification is turned unfeasible due to a high number of classes. Weiyang Liu et al., 2017 developed upon the "softmax loss" (the combination of a softmax activation and cross-entropy loss) used in the earlier works, constructing a variant of softmax that lead to a more efficient separation of different classes, that results in a better final embedding for faces.

3.4 High Frequency Bias in CNNs

High Frequency Bias is the name we give to a series of phenomena and hypothesis about CNNs predilection for higher parts of the image frequency spectrum. This would give them access to imperceptible patterns that are highly generalizable, but also very brittle, becoming a source of structural sensitivity.

Jo and Bengio, 2017 were, to the best of our knowledge, the first to study this phenomenon, calling it a preference for "surface statistical regularities". They showed that a model trained on a normal dataset was not able to generalize to a dataset with radial or random filtering of frequencies, but the inverse was not true, as a model trained on a filtered dataset would still perform fine on the original one.

The idea that CNNs brittleness could be explained by patterns on image data rather than on the model was first developed by Ilyas et al., 2019. In a first moment they did not theorize how these patterns might look like. H. Wang et al., 2020 expanded on their research proposing they could be patterns in the higher part of the spectrum. They build a qualitative case from a few examples on datasets such as CIFAR10 and ImageNet where

networks would classify an image correctly if and only if the high frequency information was present. They then perform experiments associating high frequency learning to adversarial vulnerability and memorization, and show that some recent heuristics such as BatchNorm Ioffe and Szegedy, 2015 increase the importance of high frequency components in training.

Tsuzuku and Sato, 2019 gives sound theoretical reasons as to why CNNs would be specially sensitive to noise consisting of Fourier basis directions, and might constitute what is known in the adversarial examples literature as Universal Adversarial Perturbations (UAP). Searching for suitable directions with high fool ratio, they found out that most networks do have specific sensitive spots on the Fourier spectrum, but also that it is not exactly in the higher frequency part, and would be more appropriately described as "medium" to "low". Analyzing the spectral content of other famous adversarial attacks and UAPs, they also found that they can vary a lot in terms of frequency content depending on network or dataset, also not lying strictly on the higher frequency part as it was once suggested.

Expanding the concept of surface regularities, Geirhos et al., 2018 proposed the "texture hypothesis". Contrary to the common understanding that CNNs construct shapes of ever-increasing complexity along their layers (the first ones representing edges, and so on), they propose that texture is the most decisive factor. They demonstrate this by creating situations where the shape and the texture of the object belong to contradictory classes, finding out that CNNs tend to err in favor of the latter.

There are other works which are not explicitly focused on the frequency bias, but nevertheless bring important information to consider in its research. Yin et al., 2019 showed that high frequency information is sufficient for achieving a reasonable classification accuracy on datasets such as ImageNet and CIFAR10. Similarly Brendel and Bethge, 2019 used CNNs with a very restricted receptive field size to achieve state-of-the-art performance on ImageNet, showing that shape information is not necessary, and texture information is sufficient for CNNs.

Chapter 4

Experimental Basis

In this chapter we present the basis for the experiments of this work. To avoid redundancy we preface our two chapters devoted to the two main interventions of this work with this one describing some of the common structure and experimental setup between them both. Namely, we detail the datasets, evaluation metrics and some of the implementation details.

4.1 Datasets

We list here the datasets utilized in this work. We used the CIFAR10 dataset as provided by the Keras Deep-Learning Library Chollet et al., 2015, the Street View House Numbers (SVHN) Netzer et al., 2011 as provided by the Tensorflow Deep Learning Library Martin Abadi et al., 2015, the Labeled Faces in the Wild (LFW) as provided by the UMass Lab of the University of Massachusetts Gary B. Huang et al., 2007, the CelebFaces Attributes Dataset (CelebA) as provided by the MMLab from the Chinese University of Hong Kong Z. Liu et al., 2015, and the VGGFaces2 as provided by the Visual Geometry Group of the University of Oxford Cao et al., 2018.

We also utilized a restricted version of the 2017 ImageNet Large Scale Visual Recognition Challenge (ILSVRC or ImageNet, for short) Berg et al., 2010. This restricted version, called RestrictedImageNet, was proposed in Ilyas et al., 2019, and groups several classes into nine superclasses of animals. A summary of the attributes of each dataset is given on Table 4.1. All datasets provide images in multicolor, on RGB format.

Dataset Name	Subject	Nº Examples	Classes/Identities	Image Size
CIFAR10	Object Classification	60000	10	32x32
RestrictedImageNet	Object Classification	112365	9	Varies
SVHN	Digit Classification	99289	10	32x32
LFW	Face Recognition	13233	5749	Varies
CelebA	Face Recognition	202599	10117	Varies
VGGFace2	Face Recognition	3310000	9131	Varies

Table 4.1: *Dataset Attributes*

As most of the models we use need a fixed input size, we standardize the images from datasets of varying image size to 160x160. In order to do this we use simple bi-cubic interpolation. For datasets that provide bounding boxes metadata (ILSVRC, VGGFaces2), we use them for cropping and centering the object of interest beforehand.

The SVHN dataset is available in different formats, from which we chose the cropped format, with each image resized to 32x32, intended for image classification. We use only the training and testing sets of the cropped version.

The LFW dataset also provides a list of face pairs containing both matching and mismatching face identity pairs that can be used as a benchmark for face verification algorithms. For benchmarking, the authors define some standard evaluation protocols in regards the use of LFW data and outside data.

The CelebFaces Attributes Dataset (CelebA) also comes with a proposed training/validation/testing split, which we use. It is also worth noting that the CelebA dataset has no overlapping identities with LFW.

4.2 Metrics

We detail here some of the metrics used for evaluating CV models in different tasks.

4.2.1 FR Metrics

As we described earlier we will use FR to refer to a series of different tasks. FR models essentially measure how close in similarity two human faces are, and this measure is then used to accomplish different tasks. Two of the most common are verification and identification.

Verification is a binary problem of telling whether two images belong to the same identity such as, for example, the image on a passport and a photo taken before allowing someone to board a plane. The model must then issue a simple yes or no decision, usually based on a threshold of similarity or distance. On an identification task the model has access to a collection of known identities associated with different images. The task is, then, to analyze a new face image or a scene containing possibly various images to identify whether some of the known identities are present. The model usually outputs an ordered list of candidates for each face according to their proximity. A threshold of distance or similarity is usually used to determine which identities are considered probable candidates or not.

On both cases, models can incur in two kinds of errors: Type I errors, or false positives are errors in which the model incorrectly associates two different people. Type II errors, or false negatives, are errors in which the model fails to recognize that two images belong to the same person. When the model correctly associates or deny an association between two images it is conversely called a true positive or a true negative, respectively.

The rate of Type I and Type II errors are usually inversely correlated, and their balance can be adjusted by changing the sensibility threshold of the model. Thus, a more "sensible"

model may be more cautious in declaring a match, reducing false positives, but may forgo opportunities of real matches in result, increasing false negatives. An inverse relationship may happen when reducing the threshold of a model. Performance metrics try to account for this balance between errors.

Face Verification Metrics

As Face Verification is a binary problem, it can be evaluated by all of the normal statistics for binary classification, such as sensitivity (the ratio of true positives by the amount of positive events in the dataset) or specificity (the ratio of true negatives by the amount of negative events in the dataset. Powers, 2020. A common practice to compare models is to adjust the sensitivity to a fixed quantity, such as 1 in 10.000 false positives and compare the resulting specificity P. J. Grother et al., 2018.

It is also common to illustrate the trade-offs involved graphically Fawcett, 2006. ROC (Receiver Operator Characteristic) curves display true positive rates as a function of false positive rates, and are obtained by varying the operating threshold and verifying these rates. Common metrics involving ROC curve analysis are the AUC (Area Under Curve) and the EER (Equal Error Rate), the rate at which false acceptances and false denials are equal. A good model usually maximizes the AUC while minimizing the EER. Fawcett, 2006

Face Identification Metrics

For identification the metrics need to be defined in a slightly different way, as the model outputs a list of possible candidates. The rate of false positives and rate of false negatives must be calculated in terms of this list, either by defining a specific place of the list or by defining a threshold. The identification rate of false positives would be the proportion of candidates above said threshold or place in the list, and similarly for the rate of false negatives P. J. Grother et al., 2018.

A simpler metric for identification is to check whether the correct candidate appears as the top suggestion on the list, referred as accuracy or identification score Phillips, Moon, et al., 2000. One can also calculate, for example, how many times the correct candidate appears as the fifth suggestion or before. This is called the rank-5 accuracy. For a more fine-grained analysis one can compare the accuracy at various ranks, on a curve called the Cumulative Match Scores, or Cumulative Match Characteristic Phillips, Moon, et al., 2000.

4.2.2 SISR Metrics

Metrics for assessment of SISR models can be divided into at least two groups: (1) "Intrinsic" and (2) "Extrinsic". The first compares the generated image to a ground-truth or other super-resolved versions of the same image. Generally, objective mathematical function are used to quantify this comparison. The second uses another downstream CV task to evaluate the quality of the generated or recovered image. The quality of the SISR method is then proxied by how well a method for another relevant task performs using the recovered images as input. This can be done either by degrading high quality images

and then comparing the results between downgraded, super-resolved and high quality images or by using the model on real-life low-quality images and assessing whether the performance improves or degrades after the application of the SISR model. We describe below the most common intrinsic metrics used for evaluating SISR models.

MSE and PSNR

The simplest quality metric given a ground-truth image and its reconstruction by a super-resolved algorithm is the pixel-by-pixel MSE (Mean Square Error). It is common in the literature to apply a logarithmic transformation to the MSE known as the PSNR (Peak Signal-to-Noise Ratio) and report it instead. The PSNR is defined by Equation 4.1.

$$\text{PSNR} = 10 * \log_{10}(\frac{\text{MAX}_I^2}{\text{MSE}}), \tag{4.1}$$

where MAX is the maximum pixel intensity possible for the image

The PSNR and MSE can be calculated over the 3 RGB channels wherever they are available. metric based on comparison with a ground-truth image. It is based on image moments

SSIM

The Structural Similarity Index (SSIM) Zhou Wang, Bovik, et al., 2004 is another quality calculated over different patches of the image. In specific, for two images $x_1, x_2 \in X$, three functions are defined with the intention of representing luminance $l(x_1, x_2)$, contrast $c(x_1, x_2)$ and structure $s(x_1, x_2)$, defined, respectively by Equations 4.2, 4.3 and 4.4.

$$l(x_1, x_2) = \frac{2\mu_{x1}\mu_{x2} + C_1}{\mu_{x1}^2 + \mu_{x2}^2 + C_1}, \tag{4.2}$$

$$c(x_1, x_2) = \frac{\sigma_{x1}\sigma_{x2} + C_2}{\sigma_{x1}^2 + \sigma_{x2}^2 + C_2}, \tag{4.3}$$

$$s(x_1, x_2) = \frac{\sigma_{x1x2} + C_3}{\sigma_{x1} + \sigma_{x2} + C_3}, \tag{4.4}$$

where μ_x1, μ_x2 are the mean pixel intensity value, σ_x1, σ_x2 the standard deviation and σ_{x1x2} the covariance. C_1 to C_3 are fixed coefficients for numerical stability

The SSIM is then defined on Equation 4.5 as an weighted geometric mean of these three statistics.

$$SSIM(x_1, x_2) = l(x_1, x_2)^\alpha * c(x_1, x_2)^\beta * s(x_1, x_2)^\gamma \tag{4.5}$$

With α, β, γ representing arbitrarily defined weights. On most implementations, such as the one on Tensorflow Martin Abadi et al., 2015 and on this work weights are kept with an equal value of one. When applied on colored images, we calculate the SSIM only

on the calculated luminance channel. The SSIM is also calculated over small patches of the image and then averaged.

Other Metrics

More complex metrics have been developed for SISR, and there have been extensions of the most common ones. Zhou Wang, Simoncelli, et al., 2003 extended the SSIM into MS-SSIM, which calculates some of the statistics on multiple scales of degradation. It is also common to extend PSNR into a weighted version, in which the weights are informed by some model of the Human Visual System (HVS) Hanhart et al., 2013. Finally, there are metrics based on natural scene statistics and other statistical models for both source images and degradation, such as the NQM (Noise Quality Measure) Damera-Venkata et al., 2000 and the IFC (Information Fidelity Criteria) Sheikh et al., 2005. Since these metrics rely on complex statistical modeling and assumptions that do not hold for images coming for multiple sources, they are not widely used. Lately, researchers have also found it useful to use subjective opinion scores for comparing different model outputs in a more holistic way. Researchers aggregate opinions from various human subjects using tests such as the Mean Opinion Score (MOS) as done in Ledig et al., 2017.

4.3 Implementation Details

We used the CNN implementations directly from the Keras Chollet et al., 2015 when available, or used it to program specific architectures we needed. For training and evaluating models we utilized Tensorflow Martin Abadi et al., 2015. This choice was made mostly out of the researchers' own previous familiarity with the libraries, as well as the fact that some of the pre-trained models used in the work were available only as Tensorflow models.

We ran our experiments on the machines and hardware of the Computer Vision Research Group at IME-USP, without which these experiments would not have been possible. We used different machines, all of which were equipped with two NVIDIA GeForce RTX 2070 GPUS, Intel Xeon Silver 4110 CPUs with 16 cores and 252 GB of RAM.

Chapter 5

Optimizing Super Resolution for Face Recognition

In this chapter we present a study on the effect of image resoulution loss on CNNs. As we described in Chapter 3, resolution seems to be a key factor for the performance of Computer Vision algorithms in general. Out-of-the-box SISR algorithms have been used as a pre-processing step, with marginal gains. At the same time, the SISR community has a problem in measuring the performance of their models objectively, as the most recent developments have produced more realistic images that nevertheless score lower on most traditional evaluation metrics.

In order to address these questions, we study ways to use other CV tasks as proxies for the quality of generated images, what we called *task-based evaluation*. Namely, we use the performance of Face Recognition algorithms on restored face images to evaluate how well the SISR algorithm was able to recover identity-defining, fine-grained information from the degraded images. This generates new, quantitative metrics that may help evaluate SISR methods in the future.

Furthermore, we develop a method for optimizing SISR networks to perform well in these evaluations. We take advantage of the fact that CNNs are the state-of-the-art of FR to design a regularizer similar to the Perceptual Loss Johnson et al., 2016, called the FR Loss. Our experiments show that using it helps close the gap of FR performance between HR images and super-resolved ones.

The main contributions of this study are:

- We develop a task-based evaluation protocol for assessing SISR models using FR tasks and apply it to state-of-the-art models
- We develop a new method that optimizes SISR models to aid in Face Recognition tasks

The rest of this chapter is divided as follows: we describe our evaluation protocol and the FR Loss each in a section. Then, we present our main experimental results. Finally, we discuss our results in relationship with the literature

5.1 Task-Based Evaluation

Our testing protocol for quantifying how much identity-defining information is lost with downgraded resolution and how much is recovered through SISR methods is described by Figure 5.1. In it, vertical cylinders represent collections of images (datasets), horizontal cylinders collections of embeddings (elements of R^n) and boxes represent CNNs.

From a given high-resolution dataset we produce a degraded, low resolution dataset using a degradation function $D(x, \sigma)$, where $x \in X$ is a single digital image and σ is a parameter representing the magnitude of degradation. From the low-resolution dataset we produce a restored one (super resolved dataset) using the SISR Network we would like to evaluate. A pre-trained FR model then produces embeddings for each version of the dataset, as described in Chapter 2.

We then use the different embeddings to perform FR tasks and compare the performance achieved on each version. We assume that the observed variations in performance are caused by variations in the amount of information contained on each version of the images. Thus, we expect that low resolution images will lead to lower scores on FR metrics, and super-resolved images will improve metrics relative to LR ones. Furthermore, we use the improvement achieved by different SISR CNN's as a quantitative metric of how much information the model was able to retrieve.

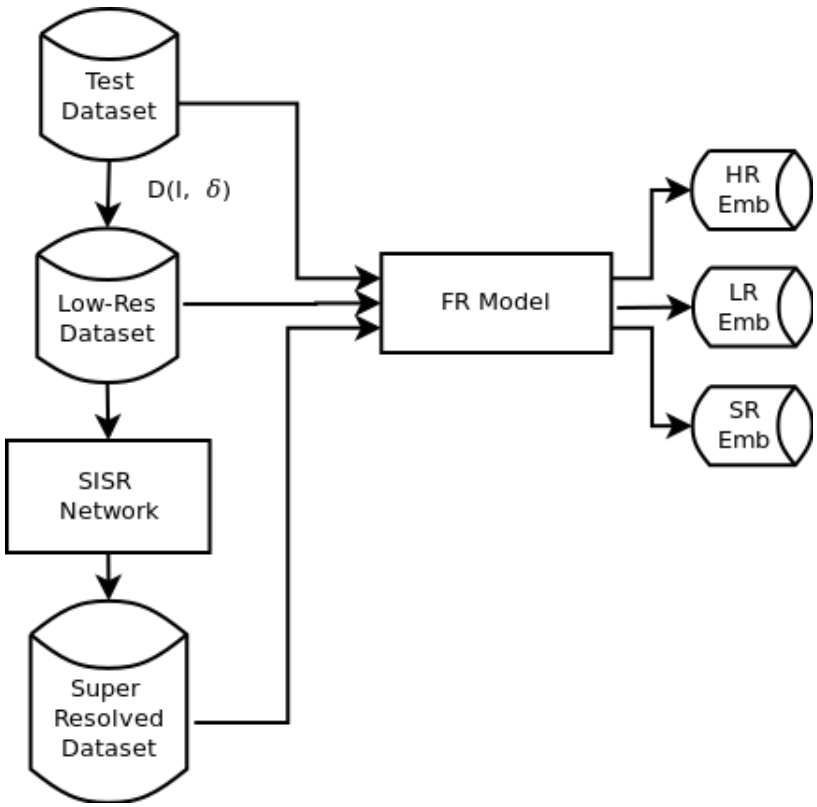

Figure 5.1: *Testing Scheme for a SISR model*

5.2 FR Loss

Recalling from Chapter 2, a Face Recognition CNN projects images onto a space in which proximity, as measured by simple distance metrics, means similarity. They are also currently the state-of-the-art of FR, as described in Chapter 3.

Considering a FR model: $f_{FR} : X \rightarrow R^n$ which transforms an image into a real-valued vector of size n. The distance between two vectors means how similar the model considers two face images to be. If a SISR model is able to minimize this distance, this would mean the network is recovering and reconstructing identity-defining information, producing faces that are similar to the original. To incentivize this behavior, we define and use the FR Loss as it is on Equation 5.1.

$$FR = ||\phi(Y) - \phi(\hat{Y})||_2 \tag{5.1}$$

We used the L_2 norm as a distance function inspired by the Perceptual Loss of Johnson et al., 2016, whose approach is very similar to ours. The FR loss is also similar to the task-based loss defined by Haris, Greg Shakhnarovich, et al., 2018. Our work distinguishes itself from them by being more specific than the perceptual loss, but nevertheless more abstract than the task-based loss. We use the end layer of an FR model differently from Johnson et al., 2016, which uses an intermediate output of a general-purpose classifier. We also do not optimize for the model's performance on a specific task, such as Haris, Greg Shakhnarovich, et al., 2018, which incentivize the SISR network to learn how to recover general identity-related information, instead of task-specific ones.

Effectively training SISR models with the FR loss is not straightforward. Directly optimizing the FR Loss lead to images with artifacts and poor quality in general, and training with the FR Loss from start led to difficulties in convergence and poor local optima. We have then devised a scheme for training that is similar to that by Ledig et al., 2017 for training with GANs, and is represented on Figure 5.2

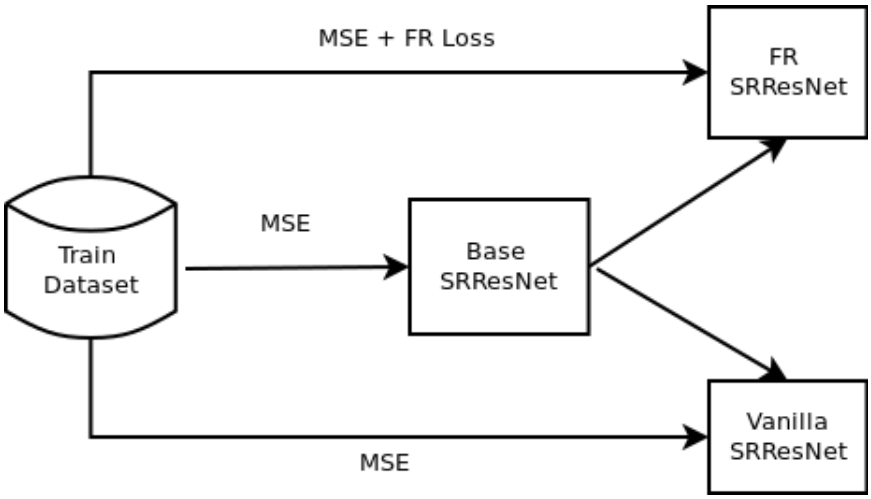

Figure 5.2: *Scheme for training a SISR model on the FR loss*

First, we trained a base network solely on MSE, and then we fine-tuned this network

to minimize the weighted average of the MSE loss and the FR loss. The weights were defined manually in such a way that at the beginning of training, each loss amounted to approximately half of the total loss. Finally, for fair comparison, we also fine-tuned the base network for the same number of epochs on MSE.

5.3 Experimental Results

We describe here the experiments performed to test our proposed protocols and a summary of our main results. We do not include specific implementation details but instead refer to the full paper included in the annex.

For calculating the FR Loss and for performing FR tasks we used a pre-trained FR model, trained on the VGGFace2 Dataset Cao et al., 2018 using simple softmax loss. [1]. We trained SISR Networks to perform 4x and 8x reconstruction by degrading images from the CelebA training partition. The networks were trained both solely on MSE ("VanillaSRResNet") and using our procedure ("FRSRResNet"). We then evaluated the trained networks first with traditional intrinsic metrics on the test set of the CelebA Dataset, and second with our task-based evaluation protocol using two different FR tasks on two different datasets, CelebA and Labelled Faces in the Wild.

For the evaluation on the CelebA Test Dataset we simulated an open-set identification problem on the test set of the CelebA Dataset. We used a simple k Nearest Neighbors classifier using the FR model's embedding as features. In other words, for each image we verified whether its nearest neighbor on the FR model's embedding (or a majority of its neighbors) belonged to the same identity as itself. We report the rank-1 accuracy of the kNN model as an assessment of the amount of information the FR model was able to retrieve.

To further evaluate if the SISR networks are retrieving information from faces in general and not overfitting to CelebA-style faces in particular, we also perform a face verification test following LFW's "unrestricted with labeled outside data" protocol Gary B Huang and Learned-Miller, 2014.

The key points we would like to bring out in the data are:

- Resolution loss affects heavily the performance of the FR model, in a non-linear fashion
- In traditional SISR metrics our method performed worse, but not by a large margin (less than a standard deviation typically)
- SISR methods were able to significantly improve performance of FR models
- SISR networks trained on the FR Loss further improved the performance
- This effect was better seen on 8x scale than 4x scale
- Qualitatively, our method produced some noise, but it seemed better at recovering characteristics such as the skin fold of the upper eyelid

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

Examples of super-resolved images are shown in Figures 5.3 and 5.4 for qualitative evaluation. Average MSE and SSIM between the super-resolved images and the original images can be seen on Table 5.1. An excerpt of the results for face identification are shown in Table 5.2. For the Face Verification task, we report two traditional metrics: AUC (Area Under the ROC Curve) and EER (Equal Error Rate). Results are shown in Table 5.3.

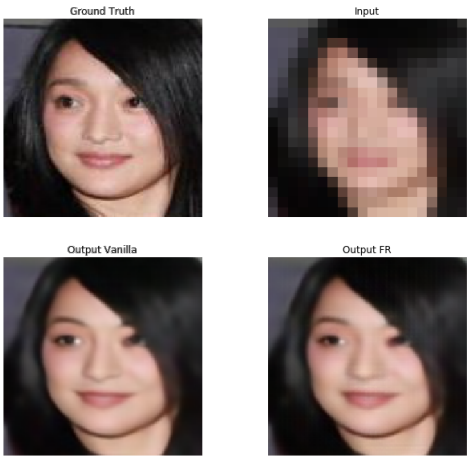

Figure 5.3: *Comparison of results for an example downsampled in 8x*

Figure 5.4: *Comparison of results and specific inset*

	PSNR (dB)	SSIM
VanillaSRResNet (8x)	**27.49 +- 2.07**	**0.875 +- 0.04**
FRSRResNet (8x)	27.30 +- 2.03	0.870 +- 0.04
VanillaSRResNet (4x)	**32.82 +- 2.56**	**0.956 +- 0.02**
FRSRResNet (4x)	32.57 +- 2.49	0.953 +- 0.02

Table 5.1: *Results for intrinsic SISR evaluation on CelebA Test Split. The best results for each scale are bolded*

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

Table 5.2: *kNN results using embeddings on CelebA Test Dataset. The best results for each scale are bolded*

Model	AUC	Equal Error Rate
High Resolution	0.999	0.012
FRSRResNet (4x)	**0.997**	**0.018**
VanillaSRResNet (4x)	**0.997**	**0.018**
Low Resolution (4x)	0.995	0.029
FRSRResNet (8x)	**0.981**	**0.065**
VanillaSRResNet (8x)	0.976	0.079
Low Resolution	0.906	0.174

Table 5.3: *AUC and EER for embedding evaluation on face verification task on LFW. The best results for each scale are bolded*

5.4 Discussion

In our work we found various results in agreement with the literature. Indeed, resolution loss degraded the performance of the downstream CV model, as in Lui et al., 2009 Hu et al., 2012. Specially similar to Lui et al., 2009, we have found that the decrease in performance happens sharply and suddenly after a certain threshold, and is not linear with degradation.

Our method was able to achieve better results in task-based evaluation and qualitative assessment. The improvements in task-based metrics were marginal on a 4x scale degradation, but more significant on an 8x scale degradation. However, these improvements were not reflected in traditional SISR metrics. On the contrary, our networks performed worse in this sense. This constitutes more evidence that these metrics are insufficient for evaluating SR quality, as in Johnson et al., 2016, Ledig et al., 2017, Y. Chen et al., 2018, and corroborates the use of task-based evaluation as a complement, as we did on this

study.

We have found that applying SISR methods in general as a previous step to recognition helped restore some of the performance lost by degradation, as Dai et al., 2016. But, even considering the improvements achieved by our method there is still a gap between the original and the restored images, even on the 4x scale, where the performance drop is less significant.

Chapter 6

Dissecting the High-Frequency Bias in Convolutional Neural Networks

From our previous investigations it remained a puzzling fact that resolution loss affected CNNs to such a great extent, even more than it does in humans Dodge and Karam, 2017. An answer to this question might in turn allow us to understand more precisely how SISR methods aid recognition, and how to design SISR methods for this end. To achieve this, we turned our work towards understanding the origin of this fragility.

The "high-frequency bias" hypothesis emerged from our study of the literature as a promising theory to understand these facts. As described in Chapter 3 this hypothesis holds that CNNs are prone to learn and over-rely on high frequency patterns, which are at the same time brittle and highly informative Jo and Bengio, 2017; Geirhos et al., 2018. This would explain the impact of resolution, but also other known facts and mysteries of CNNs, such as its overall brittleness I. J. Goodfellow et al., 2014 and its capability of generalization even when greatly overparametrized C. Zhang et al., 2016.

We have found on our review a reliable amount of evidence that corroborates this hypothesis. Nevertheless, we also found that different papers used different criteria for defining what constitutes a high or low frequency mode, most of them relying more on intuition and image recognizability than on a rigorous definition. Some of the papers also failed to consider how information might be unevenly distributed across the frequency spectrum.

We propose to study the high-frequency bias in a more systematic way. Using the concept of frequency energy defined in Chapter 2 as a proxy for information, we design a method for dividing the frequency spectrum in bands that have, on expectancy, the same amount of information. We then use a simple method inspired by feature importance procedures to quantify how much CNNs rely on each part of the frequency spectrum. To test these methods we replicate, and aggregate the diverse scenarios and conditions present on the literature.

To summarize, the main contributions of this study are:

- We develop a framework for assessing how sensible CNNs are to different parts of the frequency spectrum in a fair, methodical way
- We apply our framework to a variety of scenarios in a systematic study of the high-frequency bias

The rest of this chapter is divided as follows: we first describe our method and some important details on how we implemented it. Then, we detail the experiments we made and the results achieved. Finally, we discuss the results and their relationship with the literature.

6.1 Method

Our method consists of two interconnected parts: first we define how to assess the importance a certain frequency has on a given CV model. Then, we describe how to construct meaningful groups of frequencies. All images operations can be safely assumed to be applied channel-wise when dealing with multi-channel images, unless specified otherwise.

6.1.1 Frequency Importance

Recall from Chapter 2 that the DFT and its inverse can transform a digital image from image-space to frequency-space and vice-versa. We chose the DFT specifically to perform this transformation in order to be able to reproduce and build upon the literature, as it is the same transformation used in Jo and Bengio, 2017 and H. Wang et al., 2020, for example.

Using this, we develop a simple way to test if a given point in the frequency domain (k, l) is used by a model to classify an image X with a frequency representation of Y. We simply produce a version Y' that is equal to Y in every way but has a value of zero on the frequency (k, l). We then use the inverse DFT to produce a disturbed version of X without the specified frequency. We can see it formally defined on Equation 6.1.

$$X' = \mathcal{F}^{-1}(Y') \tag{6.1}$$

Thus, we compare the model's output on the original and on the image without the information coming from that frequency. If the prediction changes, then this constitutes evidence that the deleted point is important for the model's performance.

6.1.2 Energy Distribution Model

In order to observe greater effects we can repeat the procedure of the last section with groups of frequencies instead of individual ones. We are interested in groups of neighboring frequencies, so we can talk about properties of "low", "middle" and "high" frequencies. The format we chose to achieve this was, then, to divide the frequency spectrum in bands, or frequency discs defined by two radii, r_1 and r_2. For a given distance function, a disc

contains all frequencies which have a distance greater or equal to r_1 but strictly lesser than r_2.

As pointed out, previous studies of the high-frequency bias did not always consider that the distribution of energy throughout the spectrum is unequal. In fact, this distribution changes from image to image. Nevertheless the shape of the distribution is similar across natural images, and tends to follow a power law with relationship to the size of the frequency Burton and Moorhead, 1987. Considering the amount of energy as related to the amount of information, we propose to divide the frequency spectrum into bands with the same amount of energy.

We name the collection of integer-valued radii $r_1, r_2, ...r_n, r_n \in \mathcal{Z}$ an energy distribution model in n bands, where the frequency band $[r_i, r_{i+1}]$ represents $\frac{1}{n}$ of the total energy of the image. An example of an energy distribution model can be seen in Figure 6.1, with the x and y axis representing the frequency and each color representing a band. For visualization purposes, the frequencies are shifted as to make the zeroth frequency the center.

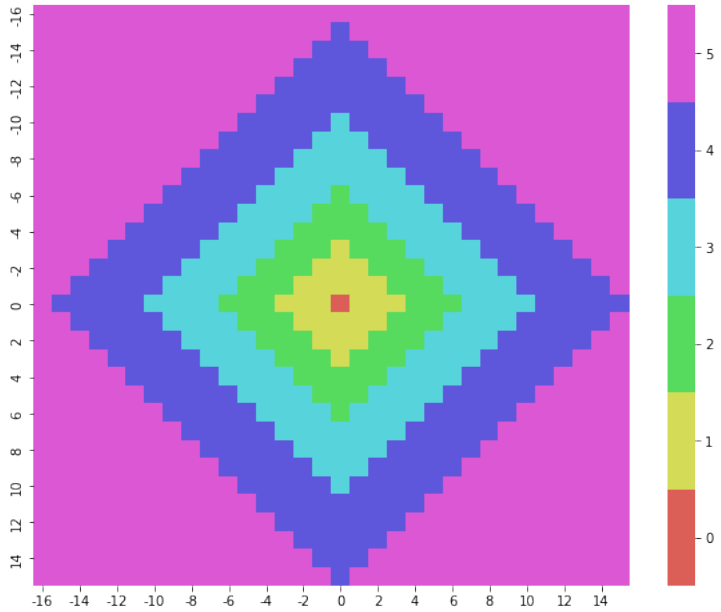

Figure 6.1: *Energy Distribution Model Example*

Besides dividing the frequency-space in a fair and methodical way, we also found that this method has another helpful property: it can be graphically shown in one -dimension in an immediate, straightforward way. As it can be seen in the experimental results section, this generates interesting and intuitive visualizations.

6.1.3 Implementation Details

To achieve statistical significance, we repeat the Frequency Importance experiment over an entire dataset of images not present on the model's training set. We aggregate the results by calculating the model's accuracy on the original images and on the distorted ones. We finally define the *estimated importance* of the frequency, or frequency importance, as the difference between these accuracies. This nomenclature is reminiscent of feature

importance procedures such as Mean Decrease Accuracy, used on tree-based classifiers and other traditional Machine Learning models Han et al., 2016.

To find the values for the energy distribution model we calculate the average energy distribution across the dataset instead of individually. This is done to avoid overly varying the energy distribution model between images while also allowing some change between different datasets of different subject matters.

For defining the radii of the frequency discs we chose the L_1 norm as a distance function, as it is more suitable for calculating distances in discrete spaces. As it assumes only integral values, we can safely restrict the radii to be integers as well. We can then define the coefficients greedily: we define $r_1 = 1$ and iteratively expand r_2 until the band formed contains $\frac{1}{n}$ of the energy distribution. Once we found r_2 we repeat this process for r_3 up to r_n.

We chose not to include the zeroth frequency because it has a qualitative different interpretation. It represents the average intensity of the pixels of the channel, and not a frequency per se. It's removal also causes a severe and disproportional distortion of the image. Finally, the calculation of the average energy distribution and the energy distribution module is done on gray-scale versions of the images. Subsequent filtering is done channel-wise. Figures 6.2, 6.3, 6.4 and 6.5 show the effect of removing some of the bands calculated acccording to our method on different datasets.

6.1.4 Robust and Non-Robust Features

We also study how importance is related to visual distortion, as measured by the l_2 distance (or mean-squared-error, MSE) of the distorted images to the original ones. This is inspired by the work of Ilyas et al., 2019, which argues that adversarial examples and overall lack of robustness in CNNs arises from the existence of non-robust features, which can be destroyed easily while not causing visually perceptible distortions.

From this perspective, we evaluate the ratio between the distortion introduced by removing a frequency band and its importance. The highest this ratio is, the easier it would be to exploit a given frequency for a small perturbation that can achieve a high fool score.

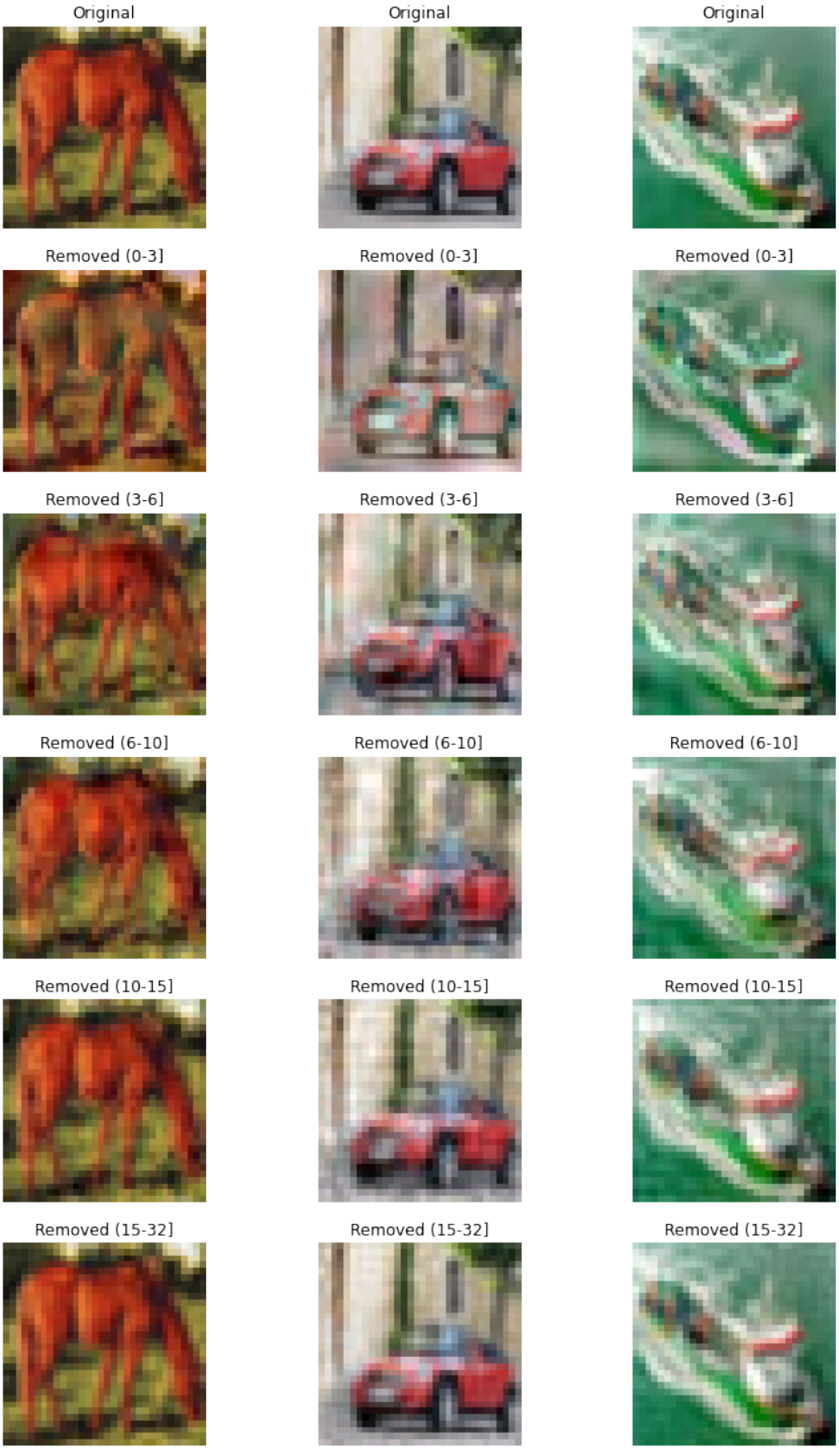

Figure 6.2: *Example of distorted CIFAR10 images according to our model. Notice how color and edges are mixed in the first few bands, but the effect is barely noticeable in the last two*

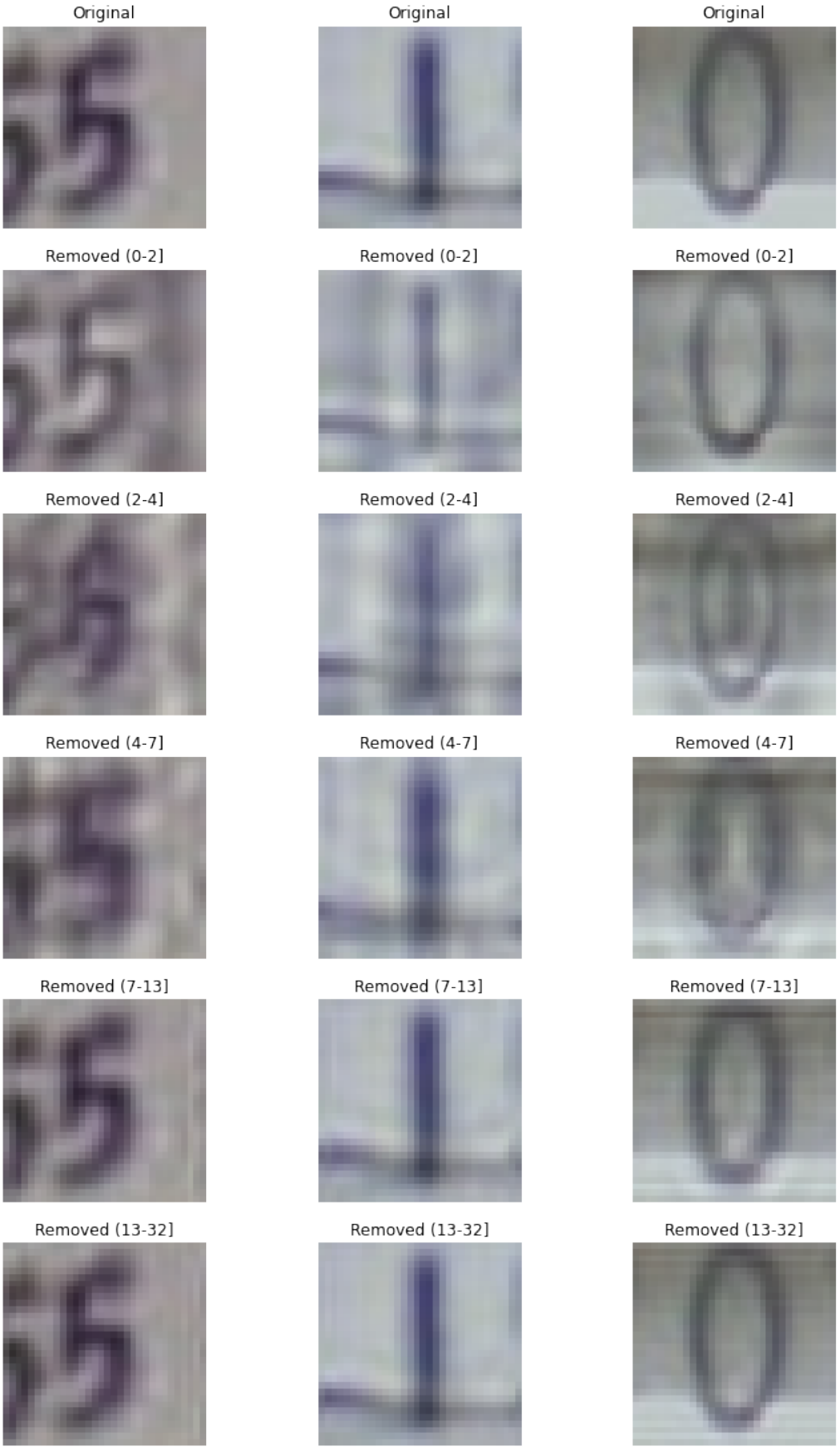

Figure 6.3: *Example of distorted SVHN images according to our model. Notice how the less clear edges on rows 2 and 3 confuse even the human eye of the class of the digit*

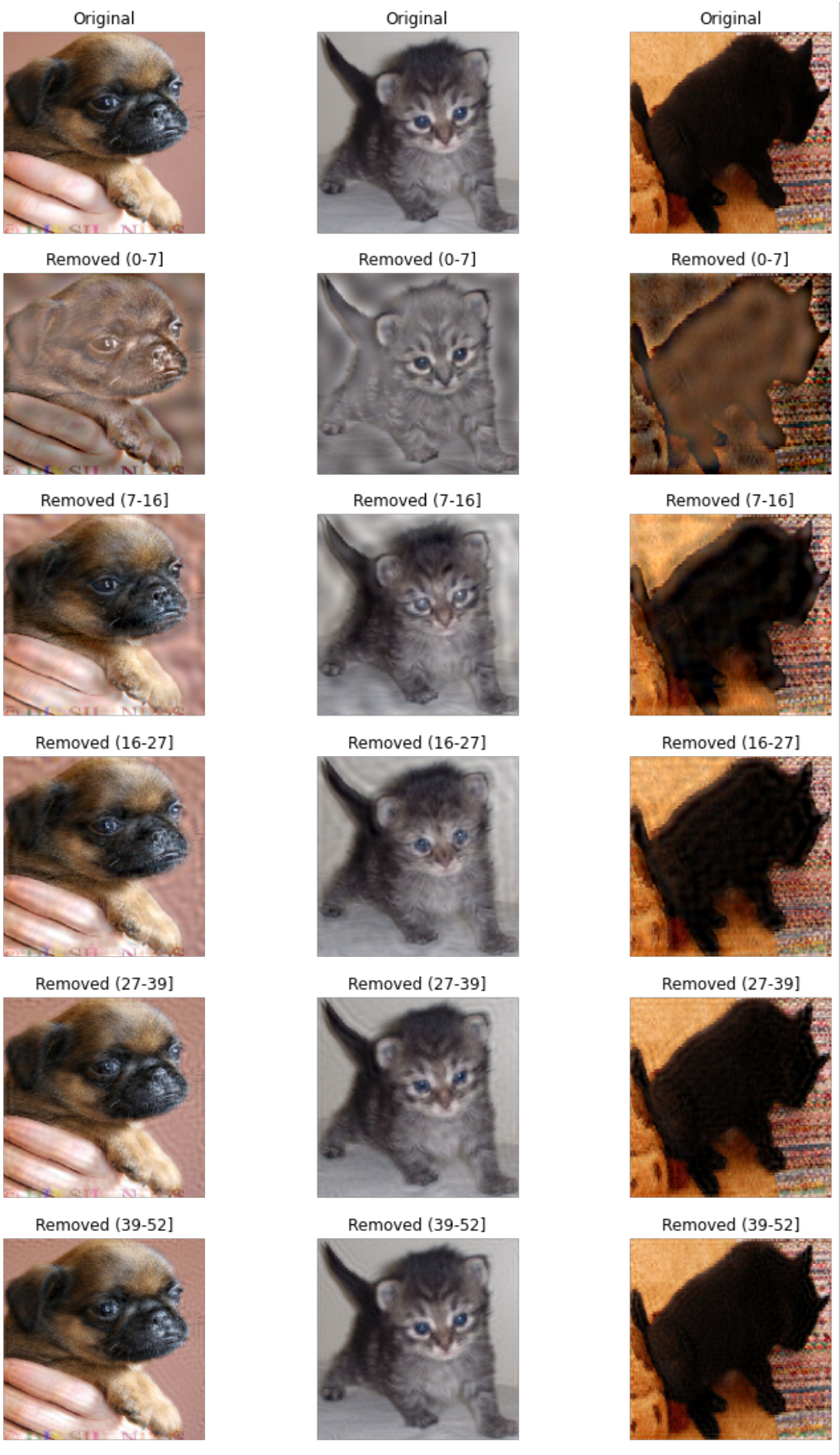

Figure 6.4: *Example of distorted ImageNet images according to our model. Only five intervals amounting each 10% of total energy are shown for brevity. Notice how the effect is barely noticeable by the fifth interval, and how the fur texture is impoverished on the third and fourth intervals*

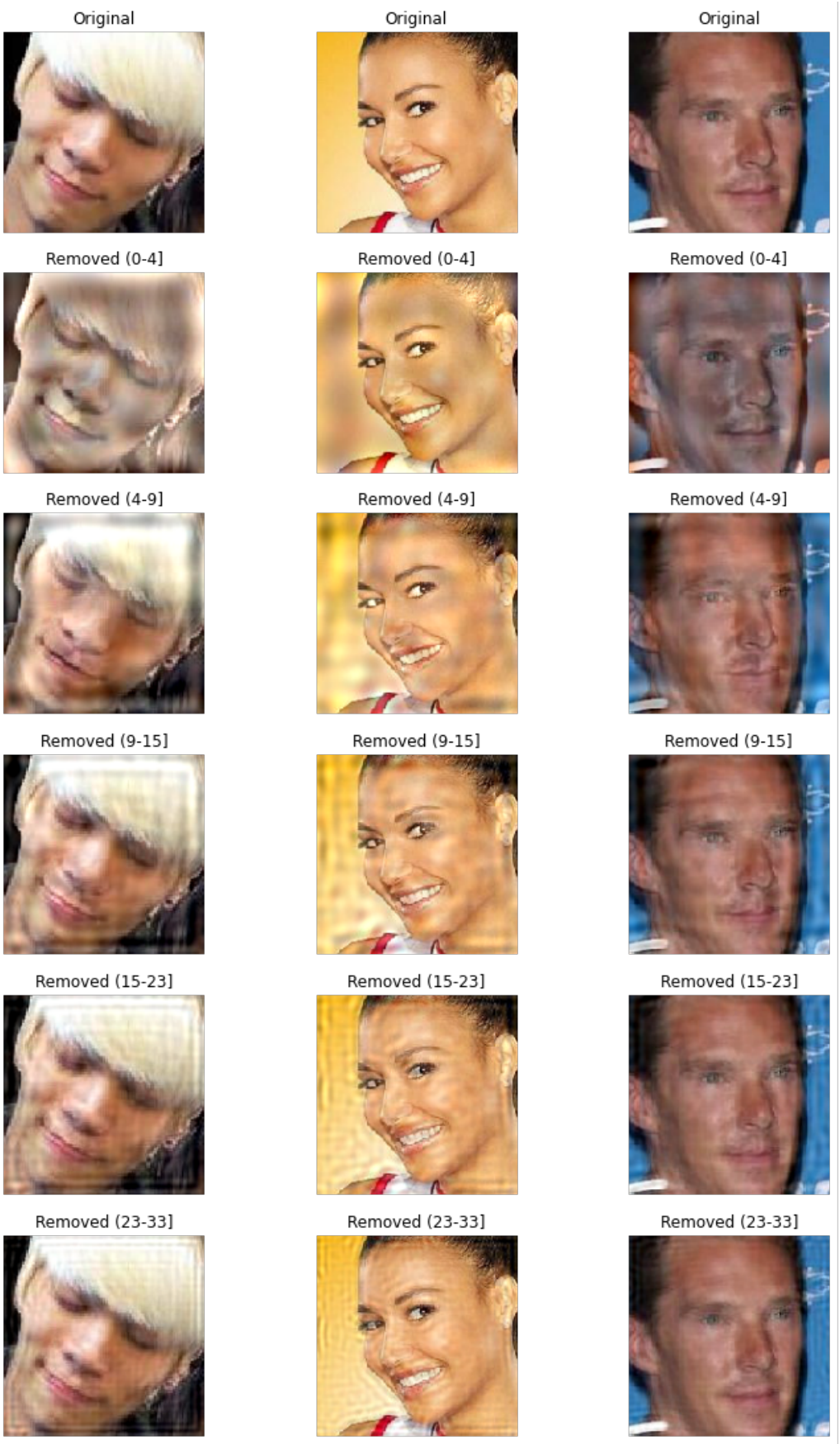

Figure 6.5: *Example of distorted VGGFaces2 images according to our model. Only five intervals amounting each 10% of total energy are shown for brevity*

6.2 Experimental Results

We conduct our studies on various scenarios to find out the relationship between frequency bias and training data. We performed experiments on two datasets of general object detection (CIFAR10, RestrictedImageNet), one of face recognition (VGGFace2) and one of in-the-wild digit recognition (SVHN). We also point out that CIFAR10 and SVHN are low resolution datasets, while VGGFace2 and ImageNet are high resolution ones. For this reason, we divided the frequency spectrum into five bands (each with 20% energy) for CIFAR10 and SVHN and into ten bands (each with 10% energy) for VGGFace2 and RestrictedImageNet.

Besides that, we also compared results between different typical confounding variables of CNNs: We trained three distinct network architectures families: VGG, ResNet and DenseNet. On the VGGFace2 dataset we trained two versions of each architecture with varying depth. Finally we compared the use of pre-processing normalization, the act of subtracting from each sample it's mean and dividing it by it's standard deviation, by training a DenseNet on all datasets with and without using it.

We present two visualizations for our results. Figure 6.6 plots for each dataset a series of line-graphs representing the test accuracy on different filtered versions of the dataset. For reference we include a line of the normal, unfiltered accuracy. Figure 6.7 plots for each image three bar graphs representing the average amount of distortion (measured by MSE) introduce while filtering the datasets, the decrease in performance caused by filtering and the ratio between these two quantities. Finally, Figure 6.8 uses both visualizations to present our comparison on network depth.

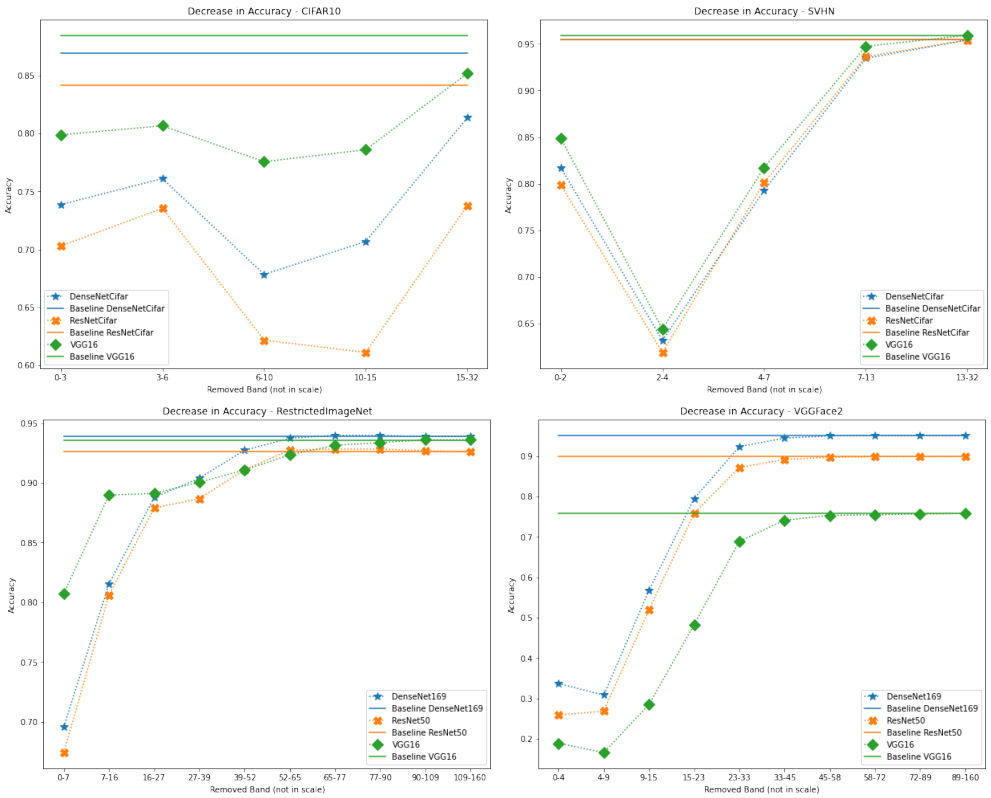

Figure 6.6: *Accuracy vs removed frequencies (frequencies not to scale)*

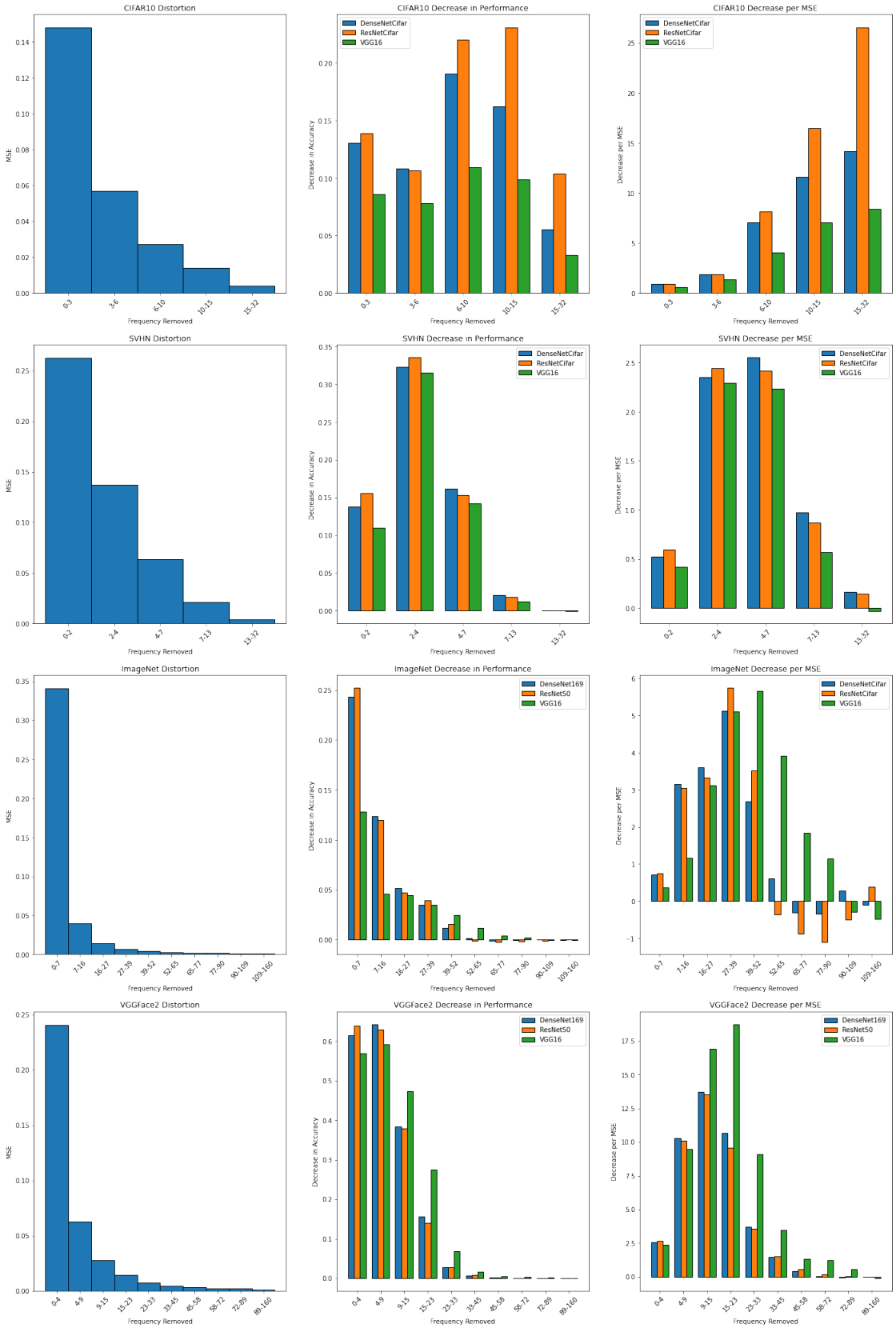

Figure 6.7: *Comparison of MSE of degraded images and decrease in performance*

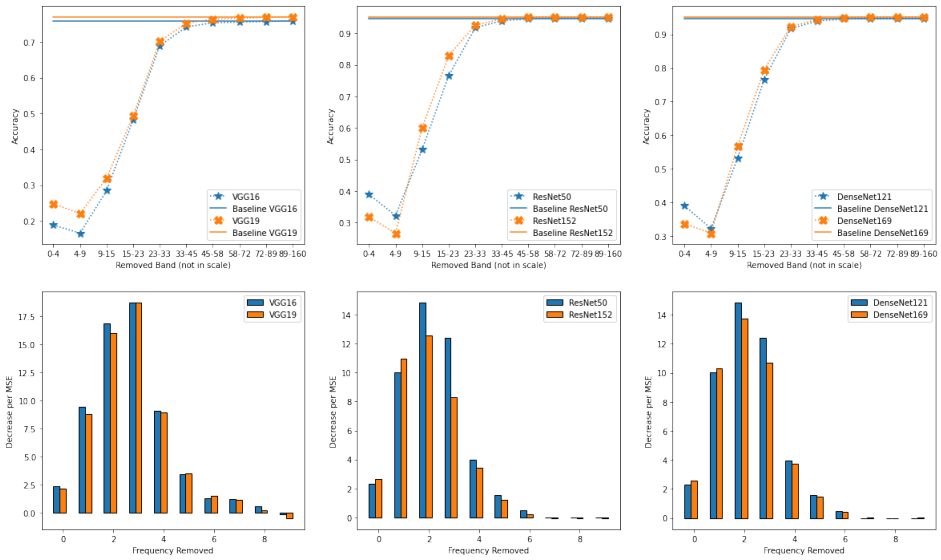

Figure 6.8: *Effect of depth on different models trained on VGGFace2*

6.3 Discussion

We were able to reproduce the results of Jo and Bengio, 2017 on the datasets they studied, with higher frequencies (the second and third on SVHN and third and fourth on CIFAR10) achieving greater importance than lower ones. This pattern, however, was not found to be universal on further testing. Overall results changed the most across datasets than on any other variable, which also made it difficult to develop a quantitative metric to analyze the results. It is also worth noting that the shape of the curve of importance is similar among datasets of same image size.

A clearer pattern arises when looking at the data through the lens of robust features. The ratio of MSE to decrease in accuracy was at its highest on mid to high frequencies either because the decrease was already high (on low-res datasets such as SVHN and CIFAR10) or because the MSE decreased at a much higher rate than the importance (hi-res datasets such as VGGFace2 and RestrictedImageNet). This is in accordance to H. Wang et al., 2020 findings, which associate higher frequencies with non-robust features.

There was no clear-cut evidence towards proving the existence or inexistence of CNNs frequency bias, but we found interesting evidence for higher frequencies being a source of adversarial vulnerability. There also wasn't significant trends when comparing results by model architecture, network depth or normalization. This could indicate that the frequency bias phenomenon is related either to universal properties in all CNNs or by particularities of the data.

Chapter 7

Conclusions

In this work, we studied some aspects of CNN's Low Resolution Recognition problem. On our first approach to the problem, in Chapter 5, we explored how a CNN-based FR model is affected by resolution loss and how to effectively counteract this effect. On our second study, in Chapter 6, we moved towards understanding the origin of CNN's sensitivity to resolution by exploring the high frequency bias hypothesis, and deepening our understanding of it with novel experiments. On this final chapter we summarize the results and conclusions we achieved with each study, and describe for each of them possible improvements and future developments.

7.1 Optimizing Super Resolution For Face Recognition

On our study in Chapter 5 we successfully reproduced some of the results of the literature. Indeed CNNs were very sensitive to resolution Dodge and Karam, 2017, but the use of SISR methods were able to boost the performance on low resolution images Dai et al., 2016. In particular, we noticed that the effect of resolution was very non-linear. Our method for training SISR networks achieved a bigger performance increase and some qualitative improvements as well while performing worse on traditional SISR metrics. In this sense, our evaluation method was successful in capturing this improvement when other metrics could not.

Even using our method for enhanced SISR training, there still was a significant gap between performance on high resolution and restored images. This could indicate that other methods of low resolution recognition are better suited for this problem, as found by Zhangyang Wang et al., 2016 and P. Li et al., 2019. Future work could bring comparisons with these other methods.

The FR loss could be improved in a number of ways as well. Concurrently and after the publication of our study, other researchers have used a similar approach to ours. Ataer-Cansizoglu et al., 2019 defines a "recognition loss" that is similar to the FR loss. Differently from us, they optimize a weighted sum of the MSE, SSIM and recognition loss from the beginning instead of using fine-tuning. Bayramli et al., 2019 uses the L1 distance

on the FR model's embedding instead, mixing it with the MSE, a GAN adversarial loss and the perceptual loss. Finally Grm et al., 2019 minimizes the cross-entropy loss of a face classification model directly. Future work should acknowledge these developments and compare their results.

Other future developments could be made using information from more than one image of the same identity. An idea could be to train the SISR network to minimize the distance between the reconstructed image and random pictures of the same identity, or the centroid of all pictures belonging to the same person. Dogan et al., 2019 trained a SISR network that received as input the FR model's representation of an image, so they could use a person's HR image to guide the model to reconstruct a LR image.

Our evaluation method could be further explored by applying it to a more varied selection of SISR networks and methods. Inclusion of different SISR models, perhaps even obsolete ones, in our experiments could also further clarify the relationship of SISR/PSNR and recovery of performance.

A question of more theoretical than practical interest would be whether our SISR models are learning how to reconstruct general identity clues, as some of our qualitative analysis shows, or just specific idiosyncrasies of a certain FR model. This could be investigated by using a "training FR model" for optimizing the SISR model and a "test FR model" for the task-based evaluation.

This study also lead to a research paper, Antonio Augusto Abello and Hirata, 2019, which was presented at the 32nd SIBGRAPI Conference on Graphics, Patterns and Images, and is included in the annex of this dissertation.

7.2 Dissecting the High Frequency Bias in Convolutional Neural Networks

On our high-frequency bias study in Chapter 6 we developed a methodical framework for testing CNNs sensitivity to the filtering of different parts of the spectrum, and applied it to a wide range of scenarios. We reproduced the results of Jo and Bengio, 2017, which was the first to propose the hypothesis, but when we expanded to other scenarios we found out that their results did not generalize. We have not found an universal phenomenon of high-frequency sensitivity, but instead the CNNs seemed to vary in their response to different frequencies. Most notably, the dataset used seemed to be the most important confounding variable.

We have not found other variables to be that important for the high-frequency bias but our experiments can be easily expanded to other scenarios. We hope that open sourcing our code can encourage this. In particular, we are interested on the effects of Batch Normalization Ioffe and Szegedy, 2015 and robust training techniques such as adversarial training Madry et al., 2017 or pre-training T. Chen et al., 2020; Jiang et al., 2020.

On our work we evaluate the influence of removed frequencies by differences in accuracy. However, there could be changes in model predictions from a wrong class to another wrong class or even from wrong class to the correct one after removing a

frequency. Initially, our hypothesis was that CNNs were drawn to informative patterns on the higher part of the spectrum, so comparing accuracy was the best experiment to test this hypothesis. On a future work, other metrics could be used to measure frequency influence in a wider way.

Further work could also refine the notion of "bias" or "preference" by trying to estimate how much useful information is actually contained on each frequency disc, perhaps by trying to fit models solely on that portion of the spectrum. Comparing the amount to information on a disc to the impact it has on a classifier may show a clearer view of the classifier's preference to it.

Our study of the high-frequency bias was also published as a research paper for CVPR's "Bridging the Gap Between Computational Photography and Visual Recognition" WorkshopAntonio A Abello et al., 2021. We also published the code used for performing our experiments on Github. It can be accessed through the link: https://github.com/Abello966/FrequencyBiasExperiments As in the other study, we include a copy of the paper in the annex of this work.

Annex A

Article Published and Presented in SIBGRAPI 2019

(This page left blank)

Optimizing Super Resolution for Face Recognition

Antonio Augusto Abello
Instituto de Matemtica e Estatistica
Universidade de Sao Paulo
Sao Paulo, Brasil, 05508-090
Email: abello.ime.usp.br

Roberto Hirata Junior
Instituto de Matemtica e Estatstica
Universidade de Sao Paulo
Sao Paulo, Brasil, 05508-090
Email: rhirata@ime.usp.br

***Abstract*—Face Super-Resolution or Face Hallucination is a subset of Super Resolution that aims to retrieve a high resolution image of a face from a lower resolution input.**

Recently, Deep Learning methods have improved drastically the quality of generated images. But these qualitative improvements are not always followed by quantitative improvements in the traditional metrics of the area, namely PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity Index). In some cases, models that perform better in opinion scores and qualitative evaluation have worse performance in these metrics, indicating they are not sufficiently informative.

To address this issue we propose a task-based evaluation procedure based on the comparative performance of face recognition algorithms on high-resolution and super-resolved images to evaluate how well the models retrieve high-frequency and identity defining information. Furthermore, as our face recognition model is differentiable, it leads to a novel loss function that can be optimized to improve performance in these tasks.

We successfully apply our evaluation method to validate this training method, yielding promising results

I. Introduction

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

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

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

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

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

The main contributions of this work are then:

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

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

II. Related Work

A. SISR and Face SR

Since Dong et al's introduced the SRCNN [12], Deep Learning methods became the state-of-the-art for single image super-resolution. Further developments on upsampling techniques [13], network architectures [3] and others have continued to improve results both in quantitative and qualitative ways. Wang et. al. [14] presents a more in-depth review of the development of the area.

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

Face super-resolution, sometimes called face hallucination is the subset of SISR that deals specifically with super-resolving images of human faces. Although methods of general super-resolution still work on face images, techniques exploiting unique properties of these images exist [16], [6].

B. Deep Face Recognition

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

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

Further developments in Deep Face Recognition were made in order to take advantage of large-scale face datasets. Parkhi et al [11] develop a new loss function that can be used to train the embedding on an open-set scenario, based on the distance between positive and negative examples of generic identities. Developments in the softmax loss [17] were also made for increasing discriminability and also facilitating training.

C. Neural Networks as Kernel Functions

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

The output of an intermediate layer of an Inception-like network [18] also pre-trained on ImageNet is commonly used as an evaluation metric for generative models. It is generally refined into the Inception Score(IS) [19], or Frchet Inception Distance (FID) [20]

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

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

D. Task-based evaluation and training

Dai et al [21] previously argued that SISR is mostly evaluated perceptually. They proceeded to do a review of the state-of-the-art methods and their effects on other CV tasks, with generally positive conclusions about the effect of super-resolution in other CV tasks, and asking for further integration between SISR and other subfields of CV.

Since Face Super-Resolution has a natural use-case in surveillance applications, task-based evaluation seems to be more common in this area. Before the emergence of Deep Learning, Hu et al [22] investigate the effects of super-resolution on surveillance applications. Rasti et al [7] train CNNs for super-resolution of faces and evaluate them using the performance of a HMM model for face recognition. These works focus more on face verification tasks, while our work extends also into face identification, as described in SectionIII

There are previous works using information from other CV tasks to aid super-resolution. This idea can also appear under the framework of multi-task learning. Bulat et al [23] train a network to perform both facial landmark estimation and super-resolution at once. Haris et al [24] develops an approach similar to ours but in regards of general object-detection instead of Face Recognition. They train a SR Network to minimize both the reconstruction loss and the error of a pre-trained neural network for object detection on the super-resolved images. We instead focus on face images and Face Recognition and define our loss function in a different way, presented in Section III.

Zhang et al. [25] propose to jointly optimize separate Face Recognition and Super Resolution models and develop techniques for this joint training, that would result in FR models robust to differences in resolution and SISR models that can recover identity information. The joint training leads to some confusion in the experimental design, though, which overlooks the generated images in favor of evaluating the jointly trained FR model.

They present three evaluation protocols: Visual Quality, quantitative and qualitative analysis of generated images, Identity Recovery, which measures the cosine similarity of super-resolved images and original images on the trained FR model's embedding, and Identity Recognizability, which trains a new FR Model on super-resolved images and test its performance on traditional FR benchmarks

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

III. Method

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

A. SISR Networks

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

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

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

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

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

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

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

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

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

B. Face Recognition Networks

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

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

where I is a set of images.

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

C. Face Recognition Loss

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

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

This FR Loss is different from the task-based loss of Haris et al [24] in the sense that it is not oriented towards an specific task, but to Face Recognition as a whole. We understand that by being more abstract our loss leads our models to recover facial characteristics in general, and not only characteristics relevant to a specific task.

D. Training for FR Loss

Experimentally, we have found that training SISR Networks exclusively on the FR Loss may lead to instability, poor local optima and overfitting that causes color aberrations, artifacts and other non-optimal behavior that seems to help minimize the FR Loss. In order to mitigate this we developed a training procedure that is illustrated in Fig. 1. First, we train a base network to minimize MSE exclusively. On a second phase, we fine-tune the base network to minimize a weighted sum of the FR Loss and MSE. We carefully define the weights so as to each one of the losses contribute approximately 50% of the total loss at the beginning of training. To provide fair comparison we also fine-tune the base network solely on MSE.

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

E. Evaluating Information Recovery of SISR Networks

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

For a high-resolution test dataset we produce a degraded version through our degradation model and a super-resolved

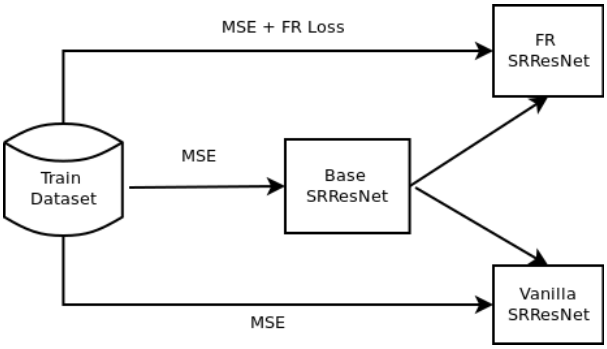

Fig. 1. Training Scheme for our Networks

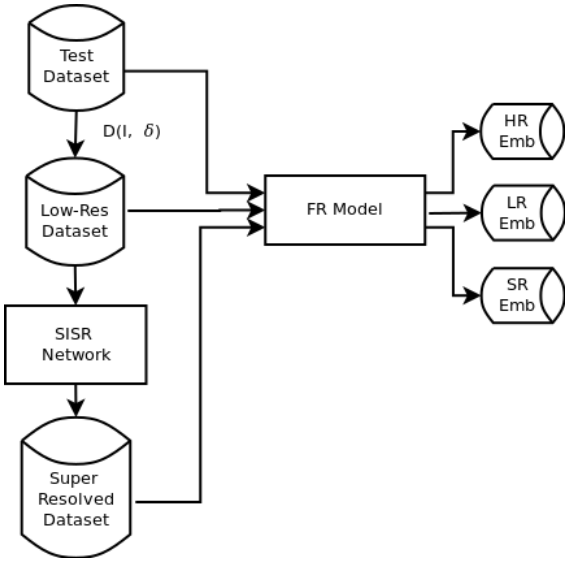

Fig. 2. Testing scheme for a single SISR Network

version through our SISR Networks. We then produce embeddings of these versions of the test dataset using an FR Model, and these embeddings are then evaluated on classical FR scenarios, which we describe more in depth on subsection IV-B. This procedure is illustrated in Fig. 2.

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

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

IV. Experiments

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

A. Datasets

1) CelebFaces Attributes Dataset: The CelebFaces Attributes Dataset (CelebA) [26] is the main dataset used in our work. It contains 202.599 face images from 10.177 distinct identities, the number of images per individual identity varies between one and thirty. The dataset is manually annotated to face landmarks and binary characteristics and there is a previous proposed partition into train, validation and test dataset containing strictly non-intersecting identities. We train our SISR Networks exclusively with the training partition of the dataset. For face identification testing and traditional SISR evaluation we select the identities of the test dataset which have exactly thirty image examples.

2) Labelled Faces in the Wild: The Labelled Faces in the Wild (LFW) Dataset [27] is a classical Face Recognition dataset that is comprised of 13233 images pertaining to 5749 different identities. It has become famous for providing a series of test protocols for diverse scenarios, some of which have become widely used benchmarks. In our work we follow the ”unrestricted with labeled outside data” protocol for testing face verification. Besides the faces themselves, this protocol offers a list of pairs of images of faces available in the dataset alongside with a classification of whether they belong to the same person or not.

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

B.

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

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

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

where MAX is the maximum pixel intensity possible for the image

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

The SSIM metric uses image moments to calculate statistical approximations for the difference in luminance, contrast and structure between two images [29]. The SSIM consists

then on a weighted geometric mean of these statistics, and ranges from 0 to 1, where 1 means a perfect match. The SSIM is generally calculated on a series of small windows of both images that is then averaged. As with the PSNR, SSIM is calculated exclusively on the luminance channel.

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

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

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

On most Deep FR applications [10] [17] the metric used for face verification is a distance metric between faces in the embedding space learned by the model. We use the simple Euclidean distance to produce a vector of distances for each pair and evaluate the embedding using the metrics described above. As discussed in Section III, the performance of the embedding can be used to gauge how much information the SISR Network could retrieve.

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

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

C. Pre-processing and Post-processing

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

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

D. Training

We trained SISR Networks to retrieve high-resolution images from the CelebA training partition after degrading them on a 4x and 8x scale. Our network's architecture was the SRResNet [3] with 10 residual blocks. We trained both the base and fine-tuned versions with Adam [33], and a learning rate of 10^{-4} and 10^{-5}, respectively. We compare the results between the network trained solely on MSE ("VanillaSRResNet") and our model ("FRSRResNet").

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

E. Testing

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

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

V. Results and Discussion

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

A. SISR Evaluation

Table I shows that, considering only classical SISR evaluation metrics and methods, our model performs slightly worse when it is optimized for the FR Loss. The super-resolved

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

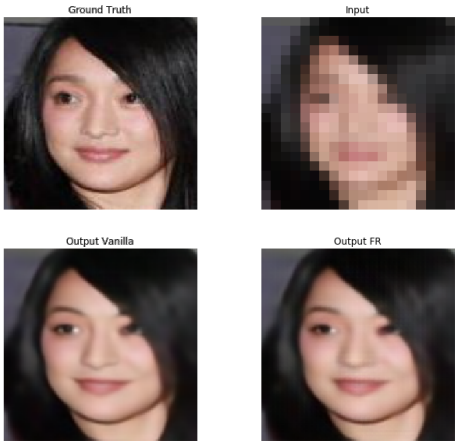

Fig. 3. Comparison of results for an example downsampled in 8x

Fig. 4. Comparison of results and specific inset

images for the FRSRResNet are generally closer than the ground-truth on our embedding space but this should come as no surprise, as this is what the model was trained to do. Otherwise, the images have less PSNR and SSIM. It is very common to models trained on different losses to behave like this while yielding seemingly better-looking images [4] [3] [6]. What usually follows is a qualitative argument, or the use of opinion scores to justify the model.

Figure 3 presents a high resolution image (top left) and its low resolution downsizing (top right). It also presents the output of the VanillaSRResNet (botton left) and our FRSRResNet model (botton right). As expected, the VanillaSRResNet

Fig. 5. Comparison of results and specific inset

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

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

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

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

B. Evaluation on Face Resolution Tasks

Beyond the qualitative argument, our evaluation procedure allows us to make quantitative arguments about the usefulness

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

TABLE II

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

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

TABLE III

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

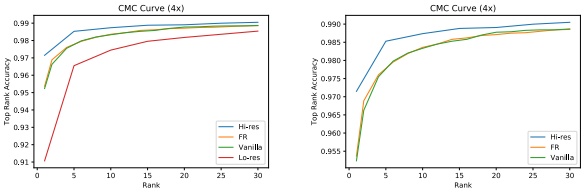

Fig. 6. Cumulative Match Characteristic for Face Identification Task at 4x Scale

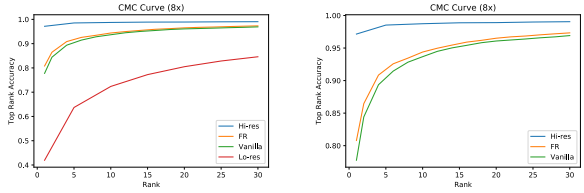

Fig. 7. Cumulative Match Characteristic for Face Identification Task at 8x Scale

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

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

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

C. Limitations and Future Work

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

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

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

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

strategy for greater improvements even in lesser degrees of degradation.

VI. Conclusions

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[16] Y. Chen, Y. Tai, X. Liu, C. Shen, and J. Yang, "Fsrnet: End-to-end learning face super-resolution with facial priors," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

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

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

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

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

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

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

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

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

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

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

[27] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," University of Massachusetts, Amherst, Tech. Rep. 07-49, October 2007.

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

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

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

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

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

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

Annex B

Article Published in Bridging the Gap Between Computational Photography And Visual Recognition 2021

(This page left blank)

Dissecting the High-Frequency Bias in Convolutional Neural Networks

Antonio A. Abello, Roberto Hirata Jr.
University of São Paulo
Butanta, São Paulo - State of São Paulo, Brazil
abello@ime.usp.br, hirata@ime.usp.br

Zhangyang Wang
The University of Texas at Austin
Austin, TX 78712, United States
atlaswang@utexas.edu

Abstract

For convolutional neural networks (CNNs), a common hypothesis that explains both their generalization capability and their characteristic brittleness is that these models are implicitly regularized to rely on imperceptible high-frequency patterns, more than humans would do. This hypothesis has seen some empirical validation, but most works do not rigorously divide the image frequency spectrum. We present a model to divide the spectrum in disjointed discs based on the distribution of energy and apply simple feature importance procedures to test whether high-frequencies are more important than lower ones. We find evidence that mid or high-level frequencies are disproportionately important for CNNs. The evidence is robust across different datasets and networks. Moreover, we find the diverse effects of the network's attributes, such as architecture and depth, on frequency bias and robustness in general. Code for reproducing our experiments is available at: `https://github.com/Abello966/FrequencyBiasExperiments`

1. Introduction

The machine learning community dedicates a considerable amount of research to understand Deep Learning's functioning in general and Convolutional Neural Networks (CNNs) in particular. Among various questions, two seem to be most intriguing: first, CNNs are capable of generalization even when they are greatly overparametrized [26]. Second, they seem to be excessively brittle and susceptible to adversarial examples [10]: small, imperceptible perturbations that make a model act in undesirable ways. A common hypothesis that would explain and unite both phenomena is that CNNs are somehow biased towards the higher frequency modes of images. Thus the network would be implicitly regularized to lean on imperceptible yet highly generalizable high-frequency patterns [17, 9]. In turn, this would make a network somehow fragile to noise and other common image corruptions that target especially this region of the frequency spectrum [25]. Additionally, this would make a CNN prone to adversarial examples that exploit how perceptually small changes in images could destroy these patterns [14, 24]. Confirming the existence of this high-frequency bias and understanding its nature would be an important step towards understanding how CNN's work and how to make them more robust.

The cited papers collectively present a reliable amount of evidence for the existence of some high-frequency bias in modern CNNs. However, most of the experiments are based on an intuitive but not rigorous definition of what constitutes a high or low image frequency mode, and some do not account for the fact that information may not be evenly distributed across the spectrum. The conditions and scenarios tested vary on each paper, leading to interesting discussions and conclusions that are also worth aggregating and consolidating on a systematic study. We propose to study the high-frequency bias by separating the image frequency spectrum in bands with the same amount of information. We then use a simple method reminiscent of feature importance procedures in traditional Machine Learning to quantify how much different models, under different circumstances, are biased towards each frequency band.

The rest of this paper is structured as follows: we perform a literature review on this topic in Section 2. Section 3 defines the notation used in this work, presents a quick recapitulation of feature importance metrics and describes our proposed method for separating the frequency spectrum. Next, in Section 4, we describe the experimental scenarios we would like to investigate in this work. After that, in Section 5, we present our results and a discussion of them and con-

clude with final remarks and future research directions on section 6.

2. Related Work

To the best of our knowledge, Jo and Bengio [17] were the first to show that the present generation of neural networks are biased towards higher frequencies, which they called surface statistical regularities. They showed that while a network trained on a low-pass filtered version of the dataset could generalize well to the unfiltered version, a network trained on the original dataset would perform much worse when the test set was low-pass filtered. This generalization gap showed that, while not indispensable, networks would latch on to these high-frequency patterns. They heuristically defined a threshold value for high and another for low frequency manually, adjusting the threshold for each dataset and maintaining the filtered images' human perceptual similarity.

This research led to exciting developments expanding the meaning of surface regularities to that of texture ones. Geirhos et al. [9] went further to propose the *texture hypothesis*, according to which the CNNs are biased more towards textural information than shape. They demonstrate the fact by creating experiments in which the shape and texture information are contradictory, finding out that CNNs tend to consider the texture information more than the shape one.

The high-frequency bias was also approached from the point of view of model robustness and adversarial perturbations. Tsuzuku et al. [23] presented sound theoretical reasons for CNN's sensibility to noise in the format of Fourier basis directions. Searching for directions that were effective in fooling classifiers, they found out that networks had increased sensitivity in some regions of the Fourier spectrum, more critically in what one could call a "middle" to "low" region. Wang et al. [24] proposed that highly generalizable but brittle high-frequency patterns in data may account both for CNN's capacity of generalization and sensitivity to adversarial attacks. They collect image examples where the absence of some higher frequencies, albeit unnoticeable by humans, would fool a CNN. They also performed experiments that associate the images' higher frequency components to memorization and overfitting.

Yin et al. [25] presents another related work that does not deal explicitly with high-frequency bias but shows that high-frequency information can be sufficient for reasonable classifying success if one trains a classifier exclusively on them. Similarly, Brendel et al. [3] achieved a competitive performance using CNNs with limited receptive field size, showing that shape information is not necessary, and texture information may be sufficient for image classification. [19] has provided a visualization study of CNN sensitivity to translations.

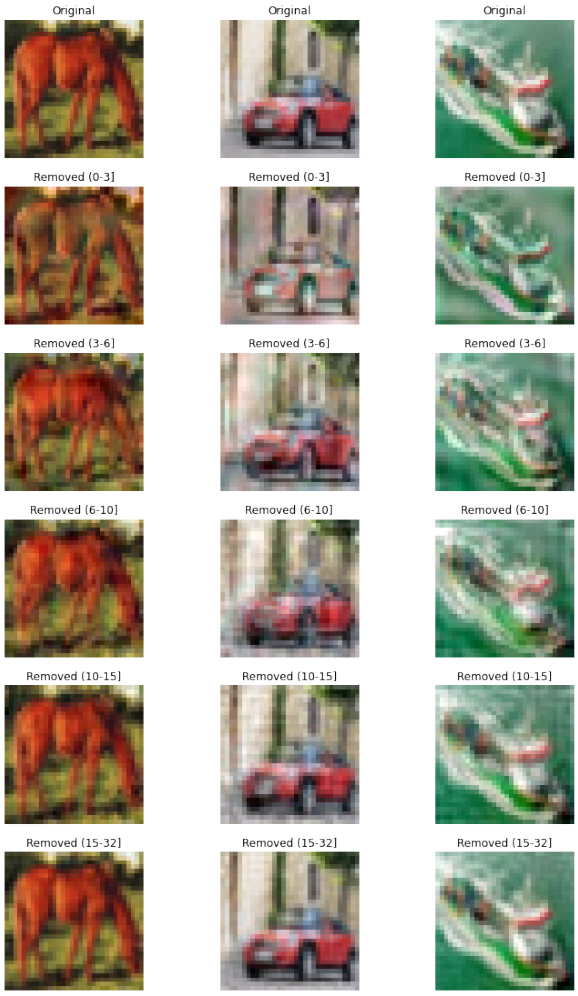

Figure 1. Example of distorted CIFAR10 images according to our model. Notice how color and edges are mixed in the first few bands, but the effect is barely noticeable in the last two

3. Method

We represent an image as a matrix, X, of pixel intensities, i.e., $X \in R^{N \times M}$, $X[p, q] \in R, p \in [0, N-1], q \in [0, M-1]$. We will omit channel information for simplicity, but all image operations are assumed to be applied channel-wise when relevant. When we refer to the Fourier transform of an image and its inverse, we are referring to the Discrete Fourier Transform (DFT) and its inverse [4]. More specifically, the Fourier transform is an operator $\mathcal{F} : R^{N \times M} \rightarrow C^{N \times M}$ such that, given a

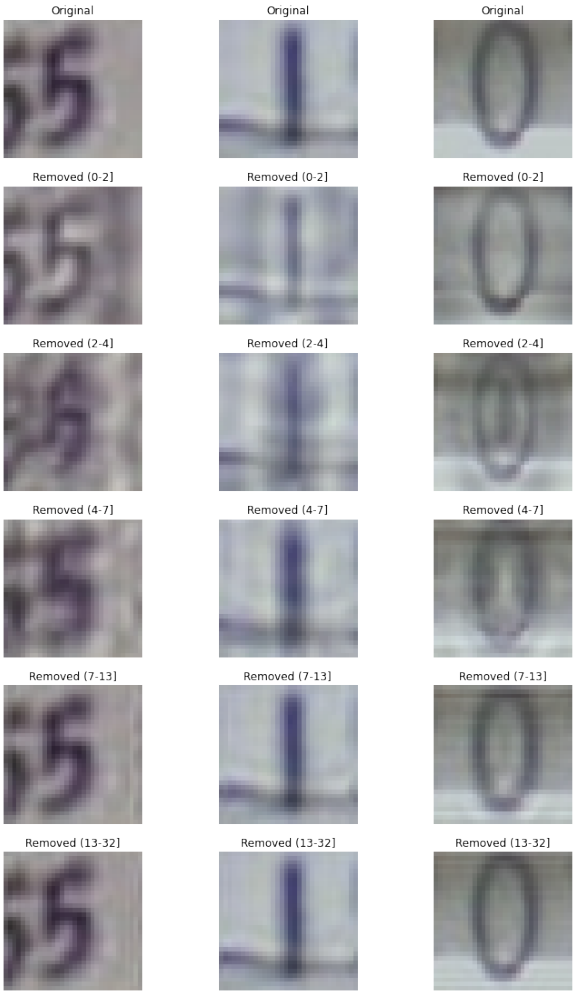

Figure 2. Example of distorted SVHN images according to our model. Notice how the less clear edges on rows 2 and 3 confuse even the human eye of the class of the digit

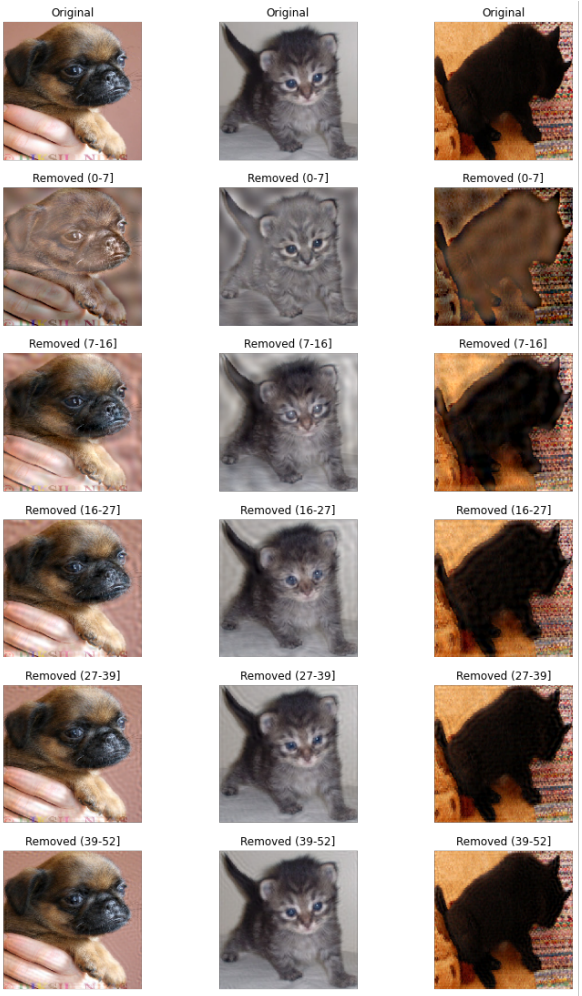

Figure 3. Example of distorted ImageNet images according to our model. Only five intervals amounting each 10% of total energy are shown for brevity. Notice how the effect is barely noticeable by the fifth interval, and how the fur texture is impoverished on the third and fourth intervals

matrix X, results in a complex-valued matrix Y:

$$Y[k,l] = \frac{1}{N*M} \sum_{p=0}^{N-1} \sum_{q=0}^{M-1} X[p,q] e^{-2\pi i (\frac{kp}{N} + \frac{lq}{M})} \quad (1)$$

For each (k,l) pair representing a frequency, the magnitude of the complex coefficient of that frequency is called the energy contributed by $Y[k,l]$ [4].

The DFT's resulting matrices are often shifted to leave the zeroth frequency ($Y[0,0]$) at the center. In this sense, the "distance," "distance from the center," or "size" of a frequency (k,l) is just the norm of the pair. The "height" of a frequency, in the sense of low and high frequencies, also refers to that.

3.1. Assessing Frequency Importance

The DFT provides a simple way to test whether a frequency is important for a model to classify an image. Given a frequency (k,l) and an image X we can construct a frequency mask M defined as:

$$M[p,q] = \begin{cases} 0, & \text{if } (p,q) = (k,l) \\ 1, & \text{otherwise} \end{cases} \quad (2)$$

This frequency mask can then be piece-wise multiplied by the Fourier transform of the image X, yielding the Fourier representation of an image without the frequency (k,l), X'. This representation can then be turned into a pixel representation of that image using the inverse DFT. To put shortly:

Figure 4. Example of distorted VGGFaces2 images according to our model. Only five intervals amounting each 10% of total energy are shown for brevity

$$X' = \mathcal{F}^{-1}C(\mathcal{F}(X) \times M) \tag{3}$$

By comparing a model's prediction on X to X', we can test if a specific frequency was important in classifying the image. If the prediction changes, then this constitutes evidence that the frequency (k, l) was important for the model's decision.

The test can be repeated throughout an entire dataset of images for a more statistically relevant test. We aggregate the information by calculating the difference in accuracy achieved by a model trained on original images tested on both natural and distorted versions of a test dataset. The *estimated importance* of a frequency is the deviance of the distorted version performance to the baseline performance. This nomenclature is reminiscent of feature importance procedures such as Mean Decrease Accuracy, used on tree-based classifiers and other traditional Machine Learning models[11].

3.2. Energy Distribution Model

The frequency importance test can be made with sets of frequencies rather than individual ones, as an individual frequency may cause an insignificant effect. For large images, on the other hand, it may be intractable to test each frequency. Since we are interested in studying the existence of a high or low-frequency bias in CNNs, we group neighboring frequencies in discs according to their distance to the zeroeth frequency. We chose to divide the frequency spectrum in bands, or frequency discs, with each disc represented by two radii r_1 and r_2, containing all the frequencies with L_1 distance greater or equal than r_1 but strictly lesser than r_2. We use the L1-norm rather than the L2-norm as it is more suit for calculating distances in discrete spaces.

To define the radii for the different discs, we refer to the aforementioned concept of energy carried by each frequency. In a sense, the amount of energy each frequency has is related to the amount of information it contains, so we consider it fair to divide the frequency spectrum into bands with the same amount of energy. We name the collection of integer-valued radii $r_1, r_2, ...r_n, r_n \in \mathcal{R}_n$ an energy distribution model, where the frequency band $[r_i, r_{i+1}), i \in [1, n-1]$ represents $\frac{1}{n}$ of the total energy.

To standardize calculations, we resize all images within a dataset to standard image size, so they always have the same amount of calculable frequencies. We also calculate the energy distribution models using the average energy distribution across all images within a test dataset instead of individually per image. This strategy allows the frequency bands to vary between different datasets of different subject matters and remain comparable within each experiment.

We find that this partition of the frequency space is appropriate for several reasons. It is of simple and straightforward interpretation, as it helps us divide the frequency spectrum from low to high frequencies in a one-dimensional fashion. This partitioning approach also allows us to define high and low-frequencies methodically.

The zeroth frequency represents the average intensity of the pixels of the channel. It has a disproportionate and qualitatively different meaning, and its removal causes severe distortion in the image. Therefore it is never included in the calculus of energy and partitioning of the frequency space. Figures 1, 2, 3, and 4 show examples of the images generated by the proposed method.

3.3. Robust and Non-Robust Features

Besides analyzing from our original point of view, in which energy should be compared with importance, another reasonable assumption would be that accuracy loss should be correlated with the amount of distortion introduced by eliminating each frequency disk. From this point of view, the *excess* of performance loss may constitute evidence for a frequency bias. This concept is related to Ilyas et al.[14]'s theoretical framework for studying robust and non-robust features. They develop a toy model in which non-robustness arises from a misalignment of the metric induced by the features with the metric used by adversarial perturbations. Applying this to our case, we study the ratio of the importance of each frequency band, as measured by our method, to the distortion introduced by removing it, as measured by the average L_2 metric (or mean-squared error, MSE) of distorted images with relation to the originals. The highest this ratio is for a frequency band, the more an adversary could exploit it to achieve a high fool ratio while maintaining a low perturbation score.

4. Experimental Setup

In order to study the effect on frequency preference produced by different data, in which discriminative features eventually lie on different parts of the spectrum, we experiment with various datasets, two of general object detection and classification (CIFAR10[18], ImageNet[2]), one of face recognition (VGGFace2[5]) and one of in-the-wild digit recognition (SVHN [21]). We train three distinct network architecture families on each dataset, VGG[22], ResNet [12] and DenseNet[13] to observe the effect of architecture on frequency bias.

Besides this general scenario, we are also interested in two other variables on frequency bias: depth of networks and pre-processing normalization. On the VGGFace2 dataset, we train and compare two versions of each architecture family of different depth. By normalization, we understand the act of subtracting from each sample its mean intensity and dividing it by its standard deviation before passing it to the neural network. To isolate architecture effects, we test the DenseNet architecture on all datasets trained with and without normalization. We divided all values between 0 and 255 by 255 whenever pixel intensities were in that range as an extra pre-processing step for the non-normalized scenarios.

4.1. Datasets

CIFAR10 is a traditional object classification dataset. It consists of 60.000 images with an already standardized 32x32 image size, divided into ten classes. It provides a train/test split of 50.000 images for training and 10.000 images for testing. We used the dataset precisely as provided by the Keras Deep-Learning library [7].

The Street View House Numbers (SVHN) [21] is an in-the-wild digit recognition dataset used for object recognition and object classification. Original images vary in image size and are provided along with bounding boxes for digits, intended for training and evaluation of object recognition. However, the dataset is also available in a cropped format, with each image resized to 32x32, intended for image classification. Collectively, the images have 73.257 digits for training, 26.032 digits for testing, and 531.131 additional examples, according to the official website [21]. We use only the cropped version training and testing sets, as they are provided in the Tensorflow Deep Learning Library [1].

The ImageNet Large Scale Visual Recognition Challenge (ILSVRC or ImageNet, for short)[2] is an annual challenge for object detection and classification with 1.000 classes. Training and validation images are provided with varying image sizes and bounding boxes for points of interest. We used the *Restricted* scenario suggested by Ilyas et al. [14], in which several classes are grouped into nine superclasses of animals. Using the ILSVRC 2017 version, this scenario includes 112.365 training images and 10.150 validation images. We crop the bounding boxes and resize all images to 160x160.

Finally, the VGGFace2 [5] is a Face Recognition dataset. It contains over 3 million images of 9.131 different identities, 8.631 included in the train set, and 500 in the test set. As a dataset of recognition and not classification, the sets of identities on the test and train set are disjoint. We artificially created a train/test classification split using the original training set. We used 5% of the data for the new test set in a stratified manner to ensure fair class representation.

We additionally pre-process VGGFace2 images by cropping the facial images using the bounding boxes provided with the dataset. We amplify the bounding boxes by 20 percent and scale all images to 160x160, interpolating and cropping when necessary but always retaining the original aspect ratio.

The energy distribution model was calculated exclusively on the test set of each dataset. For VGGFace2 and ImageNet, we divided the frequency spectrum into ten discs with 10% of the energy each. For CIFAR10 and SVHN, the small image size made it difficult to divide the calculated frequencies fairly, so we chose to use five discs with 20% of the energy each instead.

4.2. Networks and Training

We chose three different network architecture families that represent the recent evolution of CNNs for Computer Vision. We refer to their respective papers for a more in-depth explanation of their differences but highlight their essential aspects. The VGG network[22] achieved success by effectively building and training much deeper networks than the state-of-the-art. The ResNet[12] introduced residual connections, in which each layer would learn a residual to be added to the input rather than transforming it freely. This strategy proved to be a much more effective way of training CNNs. Finally, the DenseNet[13] built on ResNets by adding Dense blocks, in which every layer received as input the feature maps from each layer preceding it, improving gradient flow along with the network.

For the two datasets with larger image sizes (ImageNet and VGGFace2), the network implementations we used were the ones provided by the Keras Deep-Learning Library[7]. For the CIFAR10 and SVHN cases, we implemented the specific changes described in the ResNet and DenseNet papers to tailor these networks to datasets of smaller sizes. The VGG paper had no experiments in CIFAR10 or similar datasets, but we found that the network was able to perform well on them nevertheless.

All models are trained using standard SGD with $10^{-}2$ learning rate and 0.9 momentum, with different training duration and learning rate schedules depending on each dataset but standardized across networks. Regular data augmentation procedures are applied (random shifts, rotation, zoom, and horizontal flip, except on the SVHN case, where we do not flip images). All image data is normalized unless specified.

5. Results and Discussion

We present two visualizations for the main results. Figure 5 shows the test accuracy for each model on all degradations, along with the baseline accuracy of each model. Figure 6 shows the decrease in accuracy with relation to each model alongside the amount of distortion introduced by each filtering step and the proportion of accuracy decrease to distortion. Figure 7 uses both visualizations for our comparison on different network depth. Our main takeaways from the data are:

- Results change radically in shape and scale across datasets.
- In low-res datasets, higher frequencies tend to be disproportionately important, but the effect is less prominent on high-res ones.

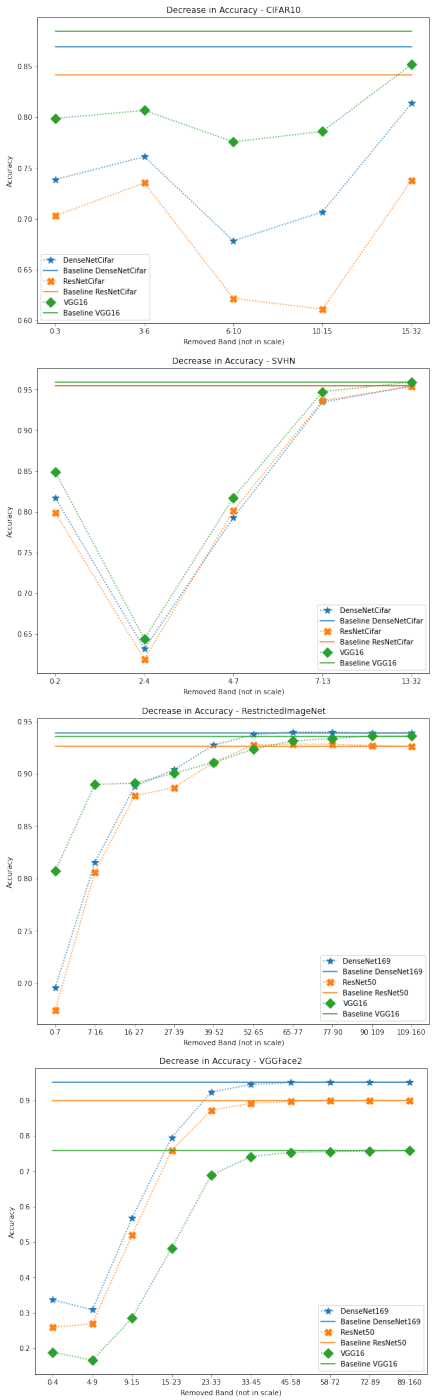

Figure 5. Accuracy vs removed frequencies (frequencies not to scale)

- Comparing accuracy decrease with MSE, mid to higher frequencies universally had a higher ratio.
- Network architectures may produce a small difference in scale, but not in the shape of the effect.

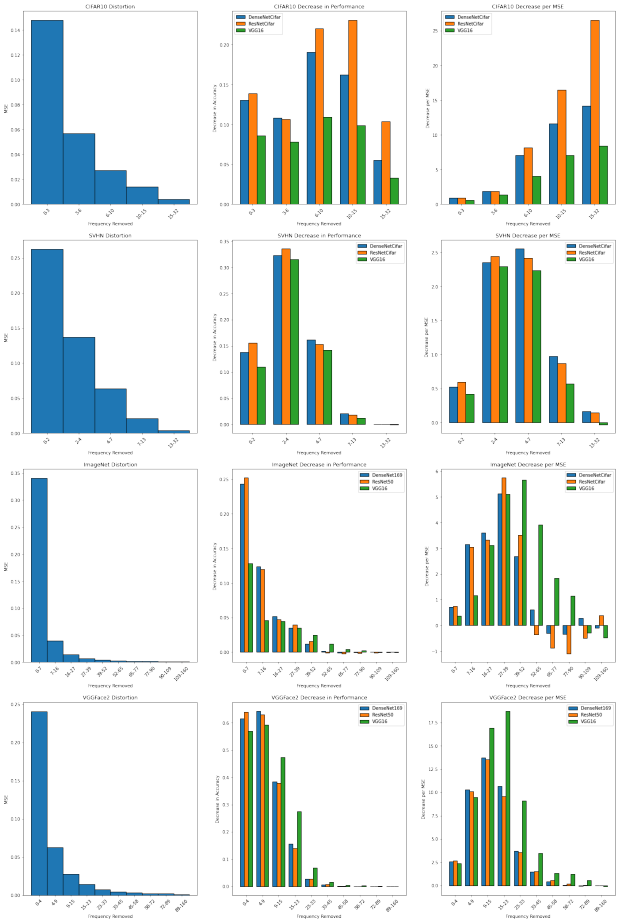

Figure 6. Comparison of MSE of degraded images and decrease in performance

- Network depth specifically does not seem to play a role in frequency bias.

Our results seem to reproduce the ones found by Jo and Bengio[17]. On the datasets they studied, we can see that higher frequencies (the second and third on SVHN and the third and fourth on CIFAR10) affected the classifiers more than the lower ones, suggesting some high-frequency bias. However, when we expanded our research to more datasets, we found that this phenomenon is not universal, as lower frequencies tend to be more critical in RestrictedImageNet and on VGGFace2, with the second frequency disc being slightly more critical on VGGFace2 in some cases. We can also see that the effect on different models is more of scale than of curve's shape. This fact suggests that the frequency bias is not related to the peculiarities of CNN architectures, either related to the universal properties of CNNs or data patterns. Image size plays a significant role in the frequency bias, as the curve varies most between the lower resolution (CIFAR10 and SVHN, with 32x32 images) and higher resolution (ImageNet and VGGFace2, which were scaled to 160x160). The difference between ImageNet and VGGFace2 can also be attributed to differences in the datasets' objectives, as discriminative facial features would lie on a higher frequency mode. This hypothesis will be the subject of further work.

From the point of view of robust features, Fig. 6 shows a more precise pattern, in which the discs with more decrease in accuracy per MSE are always in higher frequency modes. The reason is either that these discs were the ones with higher importance (CIFAR10, SVHN) or because the MSE distortion decreased way faster than the accuracy loss (ImageNet, VGGFace2). That points out that networks may be learning non-robust features in higher frequency modes, which can, in turn, be exploited in an adversarial setting, as also suggested by Wang et al. [24]. From either point of view, there is some evidence for a frequency bias. Interestingly enough, in both cases, it seems it would be more appropriate to name it a "mid-frequency" bias than a "high-frequency" one, a result similar to the one of Tsuzuku et al. [23].

We observe almost no significant patterns when comparing across network architectures. Our experiment on ImageNet seems to corroborate Geirhos et al. [9], which found that VGG-like networks were more prone to classifying ImageNet based on texture rather than shape, and Wang et al. [24], which found them more prone to learning from high-frequency components. Figure 6 shows how the decrease per MSE ratio for the VGG network lingers on and is the slowest to recede not only on the ImageNet but on the VGGFace2 case. However, this pattern was not found on the two low-res experiments, so it may not be a universal attribute of VGG-like networks.

Figure 7 shows that on our more variable-specific experiments, network depth has minimal effect on the frequency bias. Deeper models seem to attain a better accuracy at the cost of some loss in robustness, especially in the ResNet and DenseNet cases. However, looking through the robust features lens, the deeper networks seem to be less prone to high-frequency bias, which is somewhat surprising. Our experiments with normalization were not conclusive, with the results varying more across datasets and little by our confusing variable, not yielding any significant pattern.

6. Conclusion and Future Developments

We studied the common hypothesis that CNNs are prone to over-rely on imperceptible high-frequency patterns. We developed a method that allowed us to study

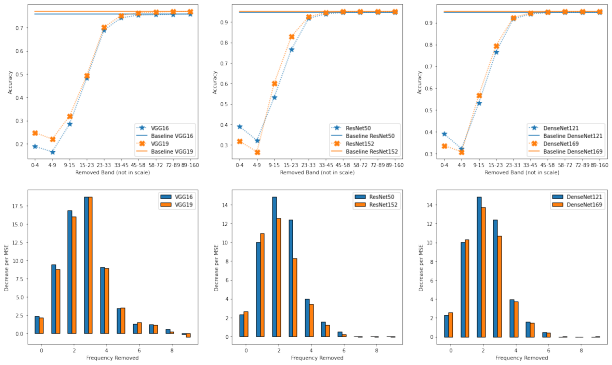

Figure 7. Effect of depth on different models trained on VGGFace2

how CNNs respond to different frequencies on different conditions. While our method has not yielded any quantitative metric, it is an improvement to the current state-of-the-art of research as it divides the frequency spectrum systematically using reasonable assumptions instead of relying on the researcher's discretion.

We found no clear-cut evidence for or against a high-frequency bias. However, we found some evidence that indicates CNNs tend to value more mid to high frequencies. This phenomenon also varied much more across datasets than by any other variable we studied, indicating that this may be more of a data phenomenon than a model phenomenon. We find it thus, improbable that the high-frequency bias hypothesis explains CNN's brittleness or capability of generalization. Our model could be applied as-is to research the effects of various other components of modern CNNs, such as Batch Normalization[15] or Adversarial Training[20]. For this reason, we provide open source code along with this paper.

There is also plenty of room for improvement in our model. Other strategies to divide the frequency spectrum could prove more informative, such as dividing by equal amounts of distortion introduced. Our analysis could also be complemented by estimating how much useful information is contained on each frequency disk, perhaps by training models exclusively on each disk. We are also interested in understanding how methods for training robust networks such as pre-training [6, 16] or architecture optimization [8] would affect those observations.

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] Alex Berg, Jia Deng, and L Fei-Fei. Large scale visual recognition challenge (ilsvrc), 2010. *URL http://www. image-net. org/challenges/LSVRC*, 3, 2010.

[3] Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. *arXiv preprint arXiv:1904.00760*, 2019.

[4] S Allen Broughton and Kurt Bryan. Discrete fourier analysis and wavelets. In *Applications to signal and image processing*. Wiley Online Library, 2009.

[5] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 67–74. IEEE, 2018.

[6] Tianlong Chen, Sijia Liu, Shiyu Chang, Yu Cheng, Lisa Amini, and Zhangyang Wang. Adversarial robustness: From self-supervised pre-training to fine-tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 699–708, 2020.

[7] François Chollet et al. Keras. `https://keras.io`, 2015.

[8] Minjing Dong, Yanxi Li, Yunhe Wang, and Chang Xu. Adversarially robust neural architectures. *arXiv preprint arXiv:2009.00902*, 2020.

[9] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

[10] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[11] Hong Han, Xiaoling Guo, and Hua Yu. Variable selection using mean decrease accuracy and mean

decrease gini based on random forest. In *2016 7th ieee international conference on software engineering and service science (icsess)*, pages 219–224. IEEE, 2016.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[13] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[14] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*, pages 125–136, 2019.

[15] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[16] Ziyu Jiang, Tianlong Chen, Ting Chen, and Zhangyang Wang. Robust pre-training by adversarial contrastive learning. *arXiv preprint arXiv:2010.13337*, 2020.

[17] Jason Jo and Yoshua Bengio. Measuring the tendency of cnns to learn surface statistical regularities. *arXiv preprint arXiv:1711.11561*, 2017.

[18] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[19] Jake Lee, Junfeng Yang, and Zhangyang Wang. What does cnn shift invariance look like? a visualization study. *arXiv preprint arXiv:2011.04127*, 2020.

[20] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[21] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

[22] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[23] Yusuke Tsuzuku and Issei Sato. On the structural sensitivity of deep convolutional networks to the directions of fourier basis functions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 51–60, 2019.

[24] Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8684–8694, 2020.

[25] Dong Yin, Raphael Gontijo Lopes, Jon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. In *Advances in Neural Information Processing Systems*, pages 13255–13265, 2019.

[26] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

References

[Antonio Augusto Abello and Hirata 2019] Antonio Augusto Abello and Roberto Hirata. "Optimizing super resolution for face recognition". In: *2019 32nd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. IEEE. 2019, pp. 194–201.

[Antonio A Abello et al. 2021] Antonio A Abello, Roberto Hirata, and Zhangyang Wang. "Dissecting the high-frequency bias in convolutional neural networks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 863–871.

[Ataer-Cansizoglu et al. 2019] Esra Ataer-Cansizoglu, Michael Jones, Ziming Zhang, and Alan Sullivan. "Verification of very low-resolution faces using an identity-preserving deep face super-resolution network". In: *arXiv preprint arXiv:1903.10974* (2019).

[Bayramli et al. 2019] Bayram Bayramli, Usman Ali, Te Qi, and Hongtao Lu. "Fh-gan: face hallucination and recognition using generative adversarial network". In: *International Conference on Neural Information Processing*. Springer. 2019, pp. 3–15.

[Broughton and Bryan 2009] S Allen Broughton and Kurt Bryan. "Discrete fourier analysis and wavelets". In: *Applications to signal and image processing*. Wiley Online Library, 2009.

[Brendel and Bethge 2019] Wieland Brendel and Matthias Bethge. "Approximating cnns with bag-of-local-features models works surprisingly well on imagenet". In: *arXiv preprint arXiv:1904.00760* (2019).

[Berg et al. 2010] Alex Berg, Jia Deng, and L Fei-Fei. "Large scale visual recognition challenge (ilsvrc), 2010". In: *URL http://www. image-net. org/challenges/LSVRC* 3 (2010).

[Burton and Moorhead 1987] Geoffrey J Burton and Ian R Moorhead. "Color and spatial structure in natural scenes". In: *Applied optics* 26.1 (1987), pp. 157–170.

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

[Bulat, J. Yang, et al. 2018] Adrian Bulat, Jing Yang, and Georgios Tzimiropoulos. "To learn image super-resolution, use a gan to learn how to do image degradation first". In: *The European Conference on Computer Vision (ECCV)*. Sept. 2018.

[Cao et al. 2018] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. "Vggface2: a dataset for recognising faces across pose and age". In: *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE. 2018, pp. 67–74.

[Y. Chen et al. 2018] Yu Chen, Ying Tai, Xiaoming Liu, Chunhua Shen, and Jian Yang. "Fsrnet: end-to-end learning face super-resolution with facial priors". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2018.

[T. Chen et al. 2020] Tianlong Chen et al. "Adversarial robustness: from self-supervised pre-training to fine-tuning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 699–708.

[Chollet et al. 2015] François Chollet et al. *Keras*. https://keras.io. 2015.

[Dai et al. 2016] Dengxin Dai, Yujian Wang, Yuhua Chen, and Luc Van Gool. "Is image super-resolution helpful for other vision tasks?" In: *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE. 2016, pp. 1–9.

[Damera-Venkata et al. 2000] Niranjan Damera-Venkata, Thomas D Kite, Wilson S Geisler, Brian L Evans, and Alan C Bovik. "Image quality assessment based on a degradation model". In: *IEEE transactions on image processing* 9.4 (2000), pp. 636–650.

[Dogan et al. 2019] Berk Dogan, Shuhang Gu, and Radu Timofte. "Exemplar guided face image super-resolution without facial landmarks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2019, pp. 0–0.

[Dodge and Karam 2016] Samuel Dodge and Lina Karam. "Understanding how image quality affects deep neural networks". In: *2016 eighth international conference on quality of multimedia experience (QoMEX)*. IEEE. 2016, pp. 1–6.

[Dodge and Karam 2017] Samuel Dodge and Lina Karam. "A study and comparison of human and deep learning recognition performance under visual distortions". In: *2017 26th international conference on computer communication and networks (ICCCN)*. IEEE. 2017, pp. 1–7.

[Dong, Loy, and Tang 2016] Chao Dong, Chen Change Loy, and Xiaoou Tang. "Accelerating the super-resolution convolutional neural network". In: *European conference on computer vision*. Springer. 2016, pp. 391–407.

[Dong, Loy, He, et al. 2014] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. "Learning a deep convolutional network for image super-resolution". In: *European conference on computer vision*. Springer. 2014, pp. 184–199.

[Fawcett 2006] Tom Fawcett. "An introduction to roc analysis". In: *Pattern recognition letters* 27.8 (2006), pp. 861–874.

[I. Goodfellow, Bengio, et al. 2016] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[Geirhos et al. 2018] Robert Geirhos et al. "Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness". In: *arXiv preprint arXiv:1811.12231* (2018).

[P. J. Grother et al. 2018] Patrick J Grother, Mei L Ngan, and Kayee K Hanaoka. "Ongoing face recognition vendor test (frvt) part 2: identification". In: (2018).

[I. Goodfellow, Pouget-Abadie, et al. 2014] Ian Goodfellow, Jean Pouget-Abadie, et al. "Generative adversarial nets". In: *Advances in neural information processing systems*. 2014, pp. 2672–2680.

[I. J. Goodfellow et al. 2014] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples". In: *arXiv preprint arXiv:1412.6572* (2014).

[Grm et al. 2019] Klemen Grm, Walter J Scheirer, and Vitomir Štruc. "Face hallucination using cascaded super-resolution and identity priors". In: *IEEE Transactions on Image Processing* 29 (2019), pp. 2150–2165.

[Hendrycks and Dietterich 2018] Dan Hendrycks and Thomas Dietterich. "Benchmarking neural network robustness to common corruptions and perturbations". In: *International Conference on Learning Representations*. 2018.

[He et al. 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[Huynh-Thu and Ghanbari 2008] Quan Huynh-Thu and Mohammed Ghanbari. "Scope of validity of psnr in image/video quality assessment". In: *Electronics letters* 44.13 (2008), pp. 800–801.

[Han et al. 2016] Hong Han, Xiaoling Guo, and Hua Yu. "Variable selection using mean decrease accuracy and mean decrease gini based on random forest". In: *2016 7th ieee international conference on software engineering and service science (icsess)*. IEEE. 2016, pp. 219–224.

[Hanhart et al. 2013] Philippe Hanhart, Pavel Korshunov, and Touradj Ebrahimi. "Benchmarking of quality metrics on ultra-high definition video sequences". In:

2013 18th International Conference on Digital Signal Processing (DSP). IEEE. 2013, pp. 1–8.

[Gary B Huang and Learned-Miller 2014] Gary B Huang and Erik Learned-Miller. "Labeled faces in the wild: updates and new reporting procedures". In: *Dept. Comput. Sci., Univ. Massachusetts Amherst, Amherst, MA, USA, Tech. Rep* (2014), pp. 14–003.

[Haris, Greg Shakhnarovich, et al. 2018] Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita. "Task-driven super resolution: object detection in low-resolution images". In: *arXiv preprint arXiv:1803.11316* (2018).

[Haris, Gregory Shakhnarovich, et al. 2018] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. "Deep back-projection networks for super-resolution". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 1664–1673.

[Hu et al. 2012] Shuowen Hu, Robert Maschal, S Susan Young, Tsai Hong Hong, and P Jonathon Phillips. "Face recognition performance with superresolution". In: *Applied optics* 51.18 (2012), pp. 4250–4259.

[Gary B. Huang et al. 2007] 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*. Tech. rep. 07-49. University of Massachusetts, Amherst, Oct. 2007.

[G. Huang et al. 2017] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. "Densely connected convolutional networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4700–4708.

[H. Huang et al. 2017] Huaibo Huang, Ran He, Zhenan Sun, and Tieniu Tan. "Wavelet-srnet: a wavelet-based cnn for multi-scale face super resolution". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 1689–1697.

[Ilyas et al. 2019] Andrew Ilyas et al. "Adversarial examples are not bugs, they are features". In: *Advances in Neural Information Processing Systems*. 2019, pp. 125–136.

[Ioffe and Szegedy 2015] Sergey Ioffe and Christian Szegedy. "Batch normalization: accelerating deep network training by reducing internal covariate shift". In: *arXiv preprint arXiv:1502.03167* (2015).

[Johnson et al. 2016] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. "Perceptual losses for real-time style transfer and super-resolution". In: *European conference on computer vision*. Springer. 2016, pp. 694–711.

[Jo and Bengio 2017] Jason Jo and Yoshua Bengio. "Measuring the tendency of cnns to learn surface statistical regularities". In: *arXiv preprint arXiv:1711.11561* (2017).

[Jiang et al. 2020] Ziyu Jiang, Tianlong Chen, Ting Chen, and Zhangyang Wang. "Robust pre-training by adversarial contrastive learning". In: *arXiv preprint arXiv:2010.13337* (2020).

[Jain and Seung 2008] Viren Jain and Sebastian Seung. "Natural image denoising with convolutional networks". In: *Advances in neural information processing systems* 21 (2008), pp. 769–776.

[Kim et al. 2016a] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. "Accurate image super-resolution using very deep convolutional networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 1646–1654.

[Kim et al. 2016b] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. "Deeply-recursive convolutional network for image super-resolution". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 1637–1645.

[Klette and Rosenfeld 2004] Reinhard Klette and Azriel Rosenfeld. *Digital geometry: Geometric methods for digital picture analysis*. Morgan Kaufmann, 2004.

[Krizhevsky et al. 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in Neural Information Processing Systems*. Ed. by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger. Vol. 25. Curran Associates, Inc., 2012, pp. 1097–1105. URL: https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[Lai et al. 2017] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. "Deep laplacian pyramid networks for fast and accurate super-resolution". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 624–632.

[LeCun, Bengio, et al. 1995] Yann LeCun, Yoshua Bengio, et al. "Convolutional networks for images, speech, and time series". In: *The handbook of brain theory and neural networks* 3361.10 (1995), p. 1995.

[LeCun, Touresky, et al. 1988] Yann LeCun, D Touresky, G Hinton, and T Sejnowski. "A theoretical framework for back-propagation". In: *Proceedings of the 1988 connectionist models summer school*. Vol. 1. 1988, pp. 21–28.

[Ledig et al. 2017] Christian Ledig et al. "Photo-realistic single image super-resolution using a generative adversarial network". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4681–4690.

[B. Li et al. 2017] Boyi Li, Xiulian Peng, Zhangyang Wang, Jizheng Xu, and Dan Feng. "Aod-net: all-in-one dehazing network". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 4770–4778.

[P. Li et al. 2019] Pei Li, Loreto Prieto, Domingo Mery, and Patrick J Flynn. "On low-resolution face recognition in the wild: comparisons and new techniques". In: *IEEE Transactions on Information Forensics and Security* (2019).

[Z. Liu et al. 2015] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. "Deep learning face attributes in the wild". In: *Proceedings of International Conference on Computer Vision (ICCV)*. 2015.

[Weiyang Liu et al. 2017] Weiyang Liu et al. "Sphereface: deep hypersphere embedding for face recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 212–220.

[D. Liu et al. 2019] Ding Liu, Bowen Cheng, Zhangyang Wang, Haichao Zhang, and Thomas S Huang. "Enhance visual recognition under adverse conditions via deep networks". In: *IEEE Transactions on Image Processing* 28.9 (2019), pp. 4401–4412.

[Lu et al. 2018] Ze Lu, Xudong Jiang, and Alex Kot. "Deep coupled resnet for low-resolution face recognition". In: *IEEE Signal Processing Letters* 25.4 (2018), pp. 526–530.

[F.-F. Li et al. 2015] Fei-Fei Li, Andrej Karpathy, and Justin Johnson. "Cs231n: convolutional neural networks for visual recognition". In: *University lecture* (2015).

[Lui et al. 2009] Yui Man Lui et al. "A meta-analysis of face recognition covariates". In: *2009 IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems*. IEEE. 2009, pp. 1–8.

[Madry et al. 2017] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. "Towards deep learning models resistant to adversarial attacks". In: *arXiv preprint arXiv:1706.06083* (2017).

[Martin Abadi et al. 2015] Martin Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015. URL: https://www.tensorflow.org/.

[Netzer et al. 2011] Yuval Netzer et al. "Reading digits in natural images with unsupervised feature learning". In: (2011).

[Phillips, Moon, et al. 2000] P Jonathon Phillips, Hyeonjoon Moon, Syed A Rizvi, and Patrick J Rauss. "The feret evaluation methodology for face-recognition algorithms". In: *IEEE Transactions on pattern analysis and machine intelligence* 22.10 (2000), pp. 1090–1104.

[Phillips, P. Grother, et al. 2003] P Jonathon Phillips, Patrick Grother, et al. "Face recognition vendor test 2002". In: *2003 IEEE International SOI Conference. Proceedings (Cat. No. 03CH37443)*. IEEE. 2003, p. 44.

[Petersen and Middleton 1962] Daniel P Petersen and David Middleton. "Sampling and reconstruction of wave-number-limited functions in n-dimensional euclidean spaces". In: *Information and control* 5.4 (1962), pp. 279–323.

[Powers 2020] David MW Powers. "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation". In: *arXiv preprint arXiv:2010.16061* (2020).

[Parkhi et al. 2015] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al. "Deep face recognition." In: *bmvc*. Vol. 1. 3. 2015, p. 6.

[Sheikh et al. 2005] Hamid R Sheikh, Alan C Bovik, and Gustavo De Veciana. "An information fidelity criterion for image quality assessment using natural scene statistics". In: *IEEE Transactions on image processing* 14.12 (2005), pp. 2117–2128.

[Shi et al. 2016] Wenzhe Shi et al. "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 1874–1883.

[Simonyan and Zisserman 2014] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[Szegedy, Zaremba, et al. 2013] Christian Szegedy, Wojciech Zaremba, et al. "Intriguing properties of neural networks". In: *arXiv preprint arXiv:1312.6199* (2013).

[Szegedy, Wei Liu, et al. 2015] Christian Szegedy, Wei Liu, et al. "Going deeper with convolutions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9.

[Szegedy, Ioffe, et al. 2017] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. "Inception-v4, inception-resnet and the impact of residual connections on learning". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 31. 1. 2017.

[Taigman et al. 2014] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. "Deepface: closing the gap to human-level performance in face verification". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 1701–1708.

[Tsuzuku and Sato 2019] Yusuke Tsuzuku and Issei Sato. "On the structural sensitivity of deep convolutional networks to the directions of fourier basis functions". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 51–60.

[VidalMata et al. 2019] Rosaura G VidalMata et al. "Bridging the gap between computational photography and visual recognition". In: *arXiv preprint arXiv:1901.09482* (2019).

[Zhou Wang, Bovik, et al. 2004] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. "Image quality assessment: from error visibility to structural similarity". In: *IEEE transactions on image processing* 13.4 (2004), pp. 600–612.

[Zhangyang Wang et al. 2016] Zhangyang Wang, Shiyu Chang, Yingzhen Yang, Ding Liu, and Thomas S Huang. "Studying very low resolution recognition using deep networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 4792–4800.

[H. Wang et al. 2020] Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P Xing. "High-frequency component helps explain the generalization of convolutional neural networks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 8684–8694.

[Zhihao Wang et al. 2019] Zhihao Wang, Jian Chen, and Steven C. H. Hoi. "Deep learning for image super-resolution: A survey". In: *CoRR* abs/1902.06068 (2019). arXiv: 1902.06068. URL: http://arxiv.org/abs/1902.06068.

[Zhou Wang, Simoncelli, et al. 2003] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. "Multiscale structural similarity for image quality assessment". In: *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*. Vol. 2. Ieee. 2003, pp. 1398–1402.

[Yin et al. 2019] Dong Yin, Raphael Gontijo Lopes, Jon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. "A fourier perspective on model robustness in computer vision". In: *Advances in Neural Information Processing Systems*. 2019, pp. 13255–13265.

[C.-Y. Yang et al. 2014] Chih-Yuan Yang, Chao Ma, and Ming-Hsuan Yang. "Single-image super-resolution: a benchmark". In: *European Conference on Computer Vision*. Springer. 2014, pp. 372–386.

[C. Zhang et al. 2016] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. "Understanding deep learning requires rethinking generalization". In: *arXiv preprint arXiv:1611.03530* (2016).

[K. Zhang et al. 2018] Kaipeng Zhang et al. "Super-identity convolutional neural network for face hallucination". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 183–198.

[Zangeneh et al. 2020] Erfan Zangeneh, Mohammad Rahmati, and Yalda Mohsenzadeh. "Low resolution face recognition using a two-branch deep convolutional neural network architecture". In: *Expert Systems with Applications* 139 (2020), p. 112854.

[Zou and Yuen 2011] Wilman WW Zou and Pong C Yuen. "Very low resolution face recognition problem". In: *IEEE Transactions on image processing* 21.1 (2011), pp. 327–340.