

Bachelor Project



**Czech
Technical
University
in Prague**

F3

**Faculty of Electrical Engineering
Department of Cybernetics**

Face Interpretation Problems on Low Quality Images

Adéla Šubrtová

**Supervisor: Ing. Jan Čech, Ph.D
May 2018**

I. Personal and study details

Student's name: **Šubrtová Adéla** Personal ID number: **457114**
Faculty / Institute: **Faculty of Electrical Engineering**
Department / Institute: **Department of Cybernetics**
Study program: **Open Informatics**
Branch of study: **Computer and Information Science**

II. Bachelor's thesis details

Bachelor's thesis title in English:

Face Interpretation Problems on Low Quality Images

Bachelor's thesis title in Czech:

Úlohy interpretace obličejů na obrázcích nízké kvality

Guidelines:

1. Get familiar with the literature on face image super-resolution, and face interpretation problems (predicting age, gender, or possibly identity), from a single image.
2. Train a cGAN network [1] to reconstruct a higher quality image from a low quality image. The primary degradation is assumed by a low resolution.
3. Train or adapt a convolutional neural network (CNN) to predict gender and age.
4. Connect both network together, i.e. the output of the super-resolution cGAN network goes into the input of the CNN predicting age and gender. Train the chain end-to-end or with fixed weights of either networks.
5. Evaluate the quality of the super-resolution and of the attribute prediction accuracy. Compare the accuracy of the chain with baselines: the original CNN and the CNN adapted to low-resolution images by data-augmentation.

Bibliography / sources:

- [1] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, Alexei A. Efros. Image-to-Image Translation with Conditional Adversarial Nets. In CVPR, 2017.
- [2] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, Wenzhe Shi. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In CVPR, 2017.
- [3] Dmitry Ulyanov, Andrea Vedaldi, Victor Lempitsky. Deep Image Prior. arXiv:1711.10925, 2017.
- [4] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In ECCV, 2016.
- [5] V. Franc, J. Čech. Face attribute learning from weakly annotated examples. In Proc. of International Conference on Automatic Face and Gesture Recognition Workshops, Biometrics in the Wild (BWILD), 2017.

Name and workplace of bachelor's thesis supervisor:

Ing. Jan Čech, Ph.D., Visual Recognition Group, FEE

Name and workplace of second bachelor's thesis supervisor or consultant:

Date of bachelor's thesis assignment: **09.01.2018** Deadline for bachelor thesis submission: **25.05.2018**

Assignment valid until: **30.09.2019**

Ing. Jan Čech, Ph.D.
Supervisor's signature

doc. Ing. Tomáš Svoboda, Ph.D.
Head of department's signature

prof. Ing. Pavel Ripka, CSc.
Dean's signature

III. Assignment receipt

The student acknowledges that the bachelor's thesis is an individual work. The student must produce her thesis without the assistance of others, with the exception of provided consultations. Within the bachelor's thesis, the author must state the names of consultants and include a list of references.

Date of assignment receipt

Student's signature

Acknowledgements

I would like to thank my supervisor Ing. Jan Čech, Ph.D for helpful guidance throughout the last year and my family for the support I received during my studies.

Declaration

I declare that the presented work was developed independently and that I have listed all sources of information used within it in accordance with the methodical instruction for observing the ethical principles in the preparation of university theses.

Prague, May 25, 2018

Prohlašuji, že jsem předloženou práci vypracovala samostatně a že jsem uvedla veškeré použité informační zdroje v souladu s Metodickým pokynem o dodržování etických principů při přípravě vysokoškolských závěrečných prací.

V Praze, 25. května 2018

.....
signature

Abstract

Automatic age and gender prediction is applicable in many real-world problems (e.g. surveillance, commercial profiling, etc.). Often, only low-resolution (LR) images are available. The goal of the thesis is to compare a baseline CNN age and gender predictor trained on high-quality images with two proposed strategies for improving prediction accuracy on low-resolution images: (1) Data-augmentation strategy trains a CNN classifier on synthetically generated LR images. (2) Super-resolution strategy enhances image resolution using conditional generative adversarial network (cGAN) and the age and gender prediction is subsequently made using the baseline CNN. The intermediate step provides human-readable interpretation, unlike in the case of data-augmentation. The experiments show that both methods outperformed the baseline method and indeed improve prediction accuracy on LR images. The super-resolution noticeably exceeding the data-augmentation given comparable amount of training data.

Keywords: facial images, age and gender classification, convolutional neural network, single-image super-resolution, conditional generative adversarial network, GAN

Supervisor: Ing. Jan Čech, Ph.D

Abstrakt

Automatický odhad věku a pohlaví má potenciální reálné aplikace (např. sledování osob, komerční profilování, apod.). Často jsou ale k dispozici pouze obrázky s nízkým rozlišením. Cílem této práce je porovnat základní klasifikátor věku a pohlaví, který byl trénován na obrázcích vysoké kvality, se dvěma navrhovanými strategiemi pro zlepšení přesnosti predikce na obrázcích nízkého rozlišení. (1) Rozšíření datové sady, které adaptuje základní CNN klasifikátor věku a pohlaví pomocí syntézy obrázků nízkého rozlišení. (2) Superrozlišení vylepšuje rozlišení za použití podmíněné generativní adversariální sítě a věk a pohlaví se následně odhadují za použití základního CNN klasifikátoru. Na rozdíl od metody rozšíření dat tento mezikrok poskytuje interpretaci výsledků srozumitelnou pro člověka. Experimenty ukazují, že obě zmíněné strategie překonaly základní metodu a opravdu zlepšují přesnost predikce na obrázcích nízkého rozlišení. Se srovnatelným počtem trénovacích dat poskytuje superrozlišení znatelně lepší výsledky.

Klíčová slova: obrázky obličeje, klasifikace věku a pohlaví, konvoluční neuronové sítě, superrozlišení, podmíněné generativní adversariální sítě, GAN

Překlad názvu: Úlohy interpretace obličeje na obrázcích nízké kvality

Contents

1 Introduction	1
2 Related Work	3
2.1 Super-Resolution	3
2.2 Generative Adversarial Networks	4
2.3 Age and Gender Prediction	4
3 The Background	5
3.1 Super-Resolution	5
3.1.1 Generative Adversarial Neural Network	5
3.2 Age and Gender prediction	7
4 Strategies for prediction accuracy improvement on LR images	9
4.1 Data-augmentation	10
4.2 Super-resolution	11
4.2.1 Chaining the networks: cGAN + CNN chain	13
4.3 Implementation details	15
5 Experiments	17
5.1 Dataset	17
5.2 Error Statistics	17
5.2.1 Age and Gender prediction accuracy	17
5.2.2 Super-resolution	18
5.3 Baseline	18
5.4 Data-augmentation	18
5.5 Super-resolution cGAN	21
5.6 Chained networks: cGAN + CNN	25
5.7 Final comparison	29
6 Conclusion	33
Bibliography	35
A CD contents	39

Figures

Tables

4.1 The Output of the CNN.	9
4.2 CNN architecture.	10
4.3 Generator architecture.	12
4.4 Discriminator architecture.	13
4.5 cGAN + CNN chain	13
4.6 Chained networks	14
4.7 cGAN context failure.	15
5.1 Baseline/Data-augmentation comparison - Gender error.	19
5.2 Baseline/Data-augmentation comparison - Mean Absolute Error.	20
5.3 Baseline/Data-augmentation comparison - Cumulative Score at 5.	20
5.4 Baseline/Data-augmentation comparison - Cumulative Score at 10.	21
5.5 Examples of super-resolved images with trained cGAN.	22
5.6 Examples of super-resolved images with trained cGAN.	23
5.7 cGAN / bilinear interpolation comparison - PSNR.	24
5.8 Chain comparison - Average PSNR.	26
5.9 Chain comparison - Average MSE.	26
5.10 Chain comparison - Gender error.	27
5.11 Chain comparison - Mean Absolute Error.	27
5.12 Chain comparison - Cumulative Score at 5.	28
5.13 Chain comparison - Cumulative Score at 10.	28
5.14 Final comparison - Gender error.	29
5.15 Final comparison - Mean Absolute Error.	30
5.16 Final comparison - Cumulative Score at 5.	30
5.17 Final comparison - Cumulative Score at 10.	31

Chapter 1

Introduction

Automatic age and gender prediction from facial images is applicable in many different fields (such as demographic data collection, commercial profiling, surveillance, etc.). Nowadays, due to the tremendous progress of convolutional neural networks, the prediction accuracy achieves and outperforms human estimates [27].

Generally in real-world, high-quality images are rarely available (e.g. in surveillance scenario, using wide angle cameras, poor quality due to the subject distance). The images are often corrupted by the low resolution, motion blur, compression artifacts, image noise or poor lightning. The thesis studies the impact of low resolution to the accuracy of age and gender prediction. The resolution is probably a prominent degradation factor and can be easily simulated.

Two strategies to improve the prediction accuracy are proposed:

- (i) **Data-augmentation.** Training or adapting a CNN by synthesizing low-resolution images.
- (ii) **Super-resolution.** Enhancing image resolution using conditional generative adversarial network [21] and feeding it into the CNN trained to predict from high-resolution images.

CNNs probably have enough capacity and can, to some extent, train or adapt to low-resolution images. A disadvantage of that approach is that large labeled dataset is needed to sufficiently train the network. Another disadvantage is that the interpretation is unclear. Extremely low-resolution images are difficult to understand for humans. A trained CNN performs the prediction with improved accuracy, however, the "black-box" nature of the approach is undesirable and may limit the practical applicability.

Whilst the super-resolution approach enhances the quality of images first, and then predicts the age and gender is human-interpretable, meaning we have an outlook on predicted attributes and the actual face. As opposed to data-augmentation, the super-resolution strategy is more general and can be applied to improve results of already trained networks. Moreover, to train the cGAN for super-resolution, there is no need to have a labeled dataset. The cGAN can be potentially trained using a huge set of *unlabeled* face images that

are widely available. The disadvantage is that the cGAN produces artifacts that can confound the age and gender classifier. To this approach, several scenarios exist, such as training super-resolution cGAN and CNN classifier end-to-end together, training with fixed weights of either network or to train both networks independently.

The goal of the thesis is to analyse and quantitatively compare the two previously discussed approaches to this problem.

In Chapter 2 we review related literature on super-resolution and age and gender prediction methods. Chapter 3 focuses on the description of specific methods used to conduct the experiments. In particular, generative adversarial networks for super-resolution, convolutional neural network for age and gender estimation. Two strategies are presented in detail in Chapter 4. Proposed strategies are compared and analysed in Chapter 5, followed by the conclusion of the thesis in Chapter 6.

Chapter 2

Related Work

This section goes briefly over related work of super-resolution and age and gender prediction. Especially, the super-resolution is very challenging problem and there exist many different approaches. Here we will review only basic principles.

2.1 Super-Resolution

Super-resolution is a classical computer vision problem [19]. The goal is to estimate a high-resolution(HR) image from low-resolution(LR) one.

Super-resolution (SR) is an ill-posed problem. HR solution is intrinsically ambiguous i.e. many HR images correspond to given LR image. Approaches to solving SR can be separated into two groups based on the number of input images.

1. Multiple-frame super-resolution (MISR) - usually bypasses the ill-posedness by utilizing similar but non-identical information (e.g. local geometry) from multiple LR images (e.g. from video sequence).
2. Single-image super-resolution (SISR) - relies on learned prior of the image classes. Since images within a class typically have a structure (e.g. facial images).

From now on we will focus on SISR only.

Most straight-forward attempts to solve SISR problem are filter-based without the need of previous learning. The result is obtained by applying a mathematical formula. Interpolation methods (e.g. Lanczos, bilinear, bicubic) are fast and still widely-used, but the solution is far too simple to solve whole complexity of the problem and typically results in overly smooth edges.

Another method of super-resolution aims to preserve edge sharpness [1]. Edges play a crucial role in human vision, so it only makes sense, that it brings a better visual quality of high-resolution outputs than when using prediction-based approaches. However, even with limited artifacts and sharp edges, the technique fails to reconstruct high-frequencies, resulting in a poor texture impression.

More complex approaches to super-resolution problem are patch-based methods, which strive to find a mapping between LR and HR images. The basic principle of patch-based methods is that the input image is decomposed into patches and subsequently super-resolved by matching its local geometry with exemplary high-resolution patches. Often, learning of the mapping functions is conducted by learning sparse dictionaries [2], [3], kernel regression [4] or support vector regression [5, 6].

Current state-of-the-art approaches are neural networks that endeavour to learn end-to-end mapping from LR input image to HR images. Convolutional networks [7] are able to learn upscaling filters and outperform previously mentioned methods. Ulyanov et. al [8] shows that even the structure of generator network contains lot of low-level statistics prior to any learning. The most recent, and probably most accurate, contributions were made with GANs¹ that proved to be successful at super-resolving [9].

2.2 Generative Adversarial Networks

Recently, GANs have gained solid reputation in many computer vision problems. Usually used in conditional setting. It is a feasible solution to problems such as single-image super-resolution [9], text to image synthesis [10], face ageing [11] and domain transfer [12, 13, 14].

2.3 Age and Gender Prediction

Age and gender classification is immensely useful practice, usually applied in human-computer interaction, surveillance, commercial profiling, psychology or demographic data collection. We will only focus on facial image classification methods given the face detections in the image.

Particularly, age prediction is challenging because there is no straightforward feature that discriminate the genders or ages, but many cues together play an important role.

Amongst the most practiced methods to predict age and gender are: principal component analysis (possibly independent component analysis) to reduce the feature space and extract the features and linear discriminant analysis to classify the gender [15]. Another approach employs Adaboost with pixel comparisons [16] or nonlinear support vector machines (SVM) with radial basis function (RBF) kernel [17] and lately convolutional neural networks. CNNs are usually trained to predict both age and gender, but the disadvantage is the necessity to have a large labeled dataset. The CNN takes raw images and learns the representation unlike using hand-engineered features in previous approaches. Franc and Cech [18] have dealt with this complication by training the CNN with weakly annotated images using an instance of EM algorithm.

¹generative adversarial networks; will be discussed more thoroughly in following chapters

Chapter 3

The Background

3.1 Super-Resolution

Generative adversarial networks have proven that are suitable for solving single-image super-resolution accurately [9].

3.1.1 Generative Adversarial Neural Network

GANs were proposed by Goodfellow et. al [21]. The idea was to introduce a generative model that allows to generate samples from very complex distributions capturing a manifold of e.g. facial images. It is implemented as a pair of convolutional neural networks - generator and discriminator. The structure corresponds to two-player minimax game - Generator and Discriminator compete against each other. Generator tries to generate output that cannot be distinguished from the real data. Whereas the discriminator tries to correctly discern the real data and the synthesized data from the generator.

Main advantage of GANs is that they produce high-quality photorealistic images with sharp edges. Competing approaches, e.g. auto-encoders [20], tend to produce overly smooth results.

Generator. Is a generative unsupervised model represented by a convolutional neural network. The input of a generator (G) is a noise vector $z \in Z$, typically a Gaussian distribution $\sim \mathcal{N}(0, 1)$ we can generate samples from. The generator maps the input to the output image via CNN.

Simply, the input of the G is a noise vector and the output is generated image x' from data space X' . The goal of G is to produce output image indistinguishable from the real data distribution X . In other words, to approximate the real data distribution as close as possible ($X' \stackrel{d}{=} X$) and thus minimize the probability that the discriminator will correctly predict the origin of its input image

$$G(\mathbf{z}; \theta_g) : Z \rightarrow X',$$

where θ_g are the generator's parameters.

Discriminator. As the name suggests, the discriminator is a discriminative supervised model. Correspondingly with G , the discriminator (D) is represented by a convolutional neural network. D 's goal is to correctly determine the origin of its input. Given an image (either from G 's distribution X' or the real data distribution X), D with parameters θ_d returns the probability of its input being from X

$$D(\mathbf{x}; \theta_d) : \{X, X'\} \rightarrow [0, 1].$$

Let $(\mathbf{x}_i, y_i)_{i=1}^N$ be the training set, where \mathbf{x}_i is an image and y_i its corresponding label ($y_i = 1$ if the input image is from the real data distribution and $y_i = 0$ if the image is synthesized by the generator).

Discriminator's loss is the cross-entropy

$$H((\mathbf{x}_i, y_i)_{i=1}^N, D) = - \sum_{i=1}^N y_i \log(D(\mathbf{x}_i)) - \sum_{i=1}^N (1 - y_i) \log(1 - D(\mathbf{x}_i)). \quad (3.1)$$

$$\text{with label } y_i = \begin{cases} 1 & \text{for } \mathbf{x}_i \sim p_{data} \\ 0 & \text{for } \mathbf{x}_i = G(\mathbf{z}), \mathbf{z} \sim p_z \end{cases}$$

Which can be rewritten as

$$J^{(D)} = H((\mathbf{x}_i, y_i)_{i=1}^N, D) = -\mathbb{E}_{\mathbf{x} \sim p_{data}} [\log(D(\mathbf{x}))] - \mathbb{E}_{\mathbf{z} \sim p_z} [\log(1 - D(G(\mathbf{z})))].$$

The GAN framework corresponds to two-player minimax game. Thus, for the generator we get

$$J^{(G)} = -J^{(D)} = \mathbb{E}_{\mathbf{x} \sim p_{data}} [\log(D(\mathbf{x}))] + \mathbb{E}_{\mathbf{z} \sim p_z} [\log(1 - D(G(\mathbf{z})))].$$

Therefore, the objective is

$$\min_G \max_D -J^{(D)} = \mathbb{E}_{\mathbf{x} \sim p_{data}} [\log(D(\mathbf{x}))] + \mathbb{E}_{\mathbf{z} \sim p_z} [\log(1 - D(G(\mathbf{z})))] \quad (3.2)$$

Convergence. The GAN framework is relatively hard to train because the convergence is not always ensured. G and D are trained simultaneously and it is often rather difficult to harmonize the training process of both networks. As opposed to other neural network set-ups, the generator often has a problem with under-fitting rather than over-fitting. A frequent failure case when training the GAN is so-called mode collapse. The situation emerges when the generator is supposed to learn a multimodal distribution but fails to produce data with enough variety.

■ Conditional GANs

GAN [21] framework is easy to extend to a conditional setting [22]. Conditioning both generator and discriminator on an additional information \mathbf{y} . Frequently in practice, the conditioning is done on class labels (generating MNIST digits [22]) or on input images [14]. Now, the objective is

$$\min_G \max_D -J^{(D)} = \mathbb{E}_{\mathbf{x} \sim p_{data}} [\log(D(\mathbf{x}|\mathbf{y}))] + \mathbb{E}_{\mathbf{z} \sim p_z} [\log(1 - D(G(\mathbf{z}|\mathbf{y})|\mathbf{y}))] \quad (3.3)$$

■ 3.2 Age and Gender prediction

Age and gender prediction can be divided into two groups according to the method of solution. The problem can be formulated as a multi-class classification [27, 28] or as a problem of regression [29, 30].

When using regression, facial features are extracted by learning the age manifold reducing the dimensionality (common method is, for example, principal component analysis). Subsequently, parameters of the regression function are computed to fit the training data. The output value (estimated age) is continuous, unlike in case of multi-class prediction, thus, can be more precise.

State-of-the-art methods for multi-class prediction are CNNs [27] performing feature extraction and classification together. An advantage of the multi-class prediction is that the output is a distribution, which provides a confidence in the prediction. Other multi-class approaches reduce input space (similarly as regression methods) and employ a classifier e.g. structured SVM [28].

Chapter 4

Strategies for prediction accuracy improvement on LR images

To conduct the experiment on age and gender classification, we used state-of-the-art method - a convolutional neural network.

We used the net architecture [18] depicted in Figure 4.2. The net classifies into two genders and 60 age categories (minimal age is 16 and maximal 75). Therefore, the output is a 120-dimensional vector with separated male and female age categories. For clarity the vector is depicted in Figure 4.1. Each dimension represents a probability of the category given an image. We computed age and gender as median over marginalized softmax distribution.



Figure 4.1: The Output of the age and gender prediction CNN. The 120-dimensional vector constitutes of two gender categories, each consists of 60 dimensions representing ages from 16 to 75. Each dimension can be interpreted as a probability of given category.

Dataset. The dataset comprises of 87,485 fully annotated (with age and gender) high-quality facial grayscale images. The faces were found by face detector, cropped and resized to 100×100 pixels. The smallest faces detected were 50×50 pixels. The dataset was split into training 78%, validation set 4% and test set 18% with no overlaps. Training and validation sets consist of 70% of PubFig database [23] and 30% of Labeled faces in the wild (LFW) database [24]. The test set is composed from PubFig (55%), LFW(17%), ChaLearnAge [26] (22%) and FG-NET [25] (6%). Age categories range from 16 to 75.

Baseline. The CNN was trained on previously described dataset using the resolution of 100×100 pixels.

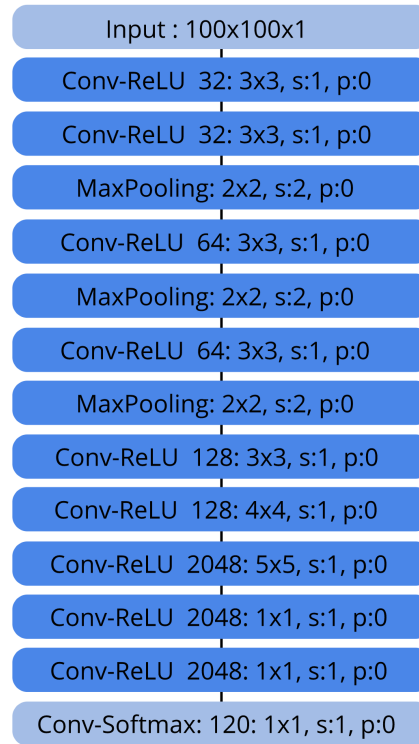


Figure 4.2: Age and gender CNN architecture. Each block carries additional information about its layers. Notation corresponds with following: number of convolutional filters: filter size, stride, padding. The output is a 120 dimensional vector depicted in Figure 4.1.

4.1 Data-augmentation

First strategy is to adapt a convolutional neural network to low-resolution images by data-augmentation. The same architecture as in case of Baseline method (Figure 4.2) was used. The augmentation was carried out by down-sampling images of the training set (100×100 pixels) randomly to scales: 1, 0.5, 0.25, 0.125 with uniform distribution, and subsequently upsampled to the former 100×100 size via bicubic interpolation. The augmentation was implemented within the SGD iteration.

In principle, it would be possible to adapt the CNN to single-scale downsampled images, but with additional information about the real-world problem environment (such as the camera resolution etc.). For each resolution a special

CNN needs to be trained.

Therefore, we decided to mix all resolutions and train a single CNN.

4.2 Super-resolution

To conduct the super-resolution part of experiments we used pix2pix image-to-image translation framework [14]. The paper proposes a general-purpose framework for image-to-image translation using generative adversarial network conditioned on an input image. The generator learns the mapping between the input image and the output image (mappings such as architectural labels \rightarrow photo, object edges \rightarrow photo, map \rightarrow aerial photo or day \rightarrow night). Whereas the discriminator serves as a structured loss, penalizing the structure at the scale of patches. Two losses are combined. The L_1 loss between generated image and the output image and the adversarial loss that proved to output photorealistic images. Experiments show that the framework provides impressive results, especially, with highly structured outputs.

Low-resolution input and high-resolution output have the same structure, the LR image is only lacking high frequencies. That makes our problem appropriate for previously mentioned framework. The generator’s architecture is depicted in Figure 4.3 and discriminator’s in Figure 4.4. Generator’s architecture with skip connection is convenient because of the same underlying structure in LR and HR images. The framework is conditioned on low-resolution images and instead of input noise vector, the noise is provided in the form of dropout (also during testing phase).

The final objective is

$$\begin{aligned}
 G^* = \arg \min_G \max_D \mathbb{E}_{\mathbf{I}_{LR}, \mathbf{I}_{HR} \sim p_{data}(\mathbf{I}_{LR}, \mathbf{I}_{HR})} [\log(D(\mathbf{I}_{LR}, \mathbf{I}_{HR}))] \\
 + \mathbb{E}_{\mathbf{I}_{LR} \sim p_{data}(\mathbf{I}_{LR}), \mathbf{z} \sim p_z(\mathbf{z})} [\log(1 - D(\mathbf{I}_{LR}, G(\mathbf{I}_{LR}, \mathbf{z})))] \\
 + \lambda \mathbb{E}_{\mathbf{I}_{LR}, \mathbf{I}_{HR} \sim p_{data}(\mathbf{I}_{LR}, \mathbf{I}_{HR}), \mathbf{z} \sim p_z(\mathbf{z})} [\| \mathbf{I}_{HR} - G(\mathbf{I}_{LR}, \mathbf{z}) \|_1].
 \end{aligned} \tag{4.1}$$

In all experiments λ is set to 100 following the provided code of [14].

Dataset. Dataset comprises of total 2650 (2419 for training and 231 for testing) high-quality facial images. Faces were detected¹ and cropped from images from IMDB database. High-quality was ensured by selecting only samples with detection score¹ over 130 and bounding-box size larger than 224×224 pixels. All images were resized to 224×224 px (a standardly used resolution of ImageNet CNNs - AlexNet, VGG). Resolution degradation was simulated by downsampling the input image and upsampling via bicubic interpolation before the training. The downsampling was implemented by Matlab *imresize* function. The antialiasing filter was on. Before upsampling to 224×224 , the intensities were quantized to integer levels 0 - 255. Downsampling scales $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$ were selected

¹Eyede Recognition, Ltd. www.eyede.cz

randomly with sampling probabilities² {0.2, 0.15, 0.1, 0.1, 0.1, 0.1, 0.1, 0.05, 0.05, 0.05}. Simply, the smallest resolution was 22×22 px ($0.1 \times 224 \doteq 22$) selected with probability 0.2 and the largest 224×224 with probability 0.05.

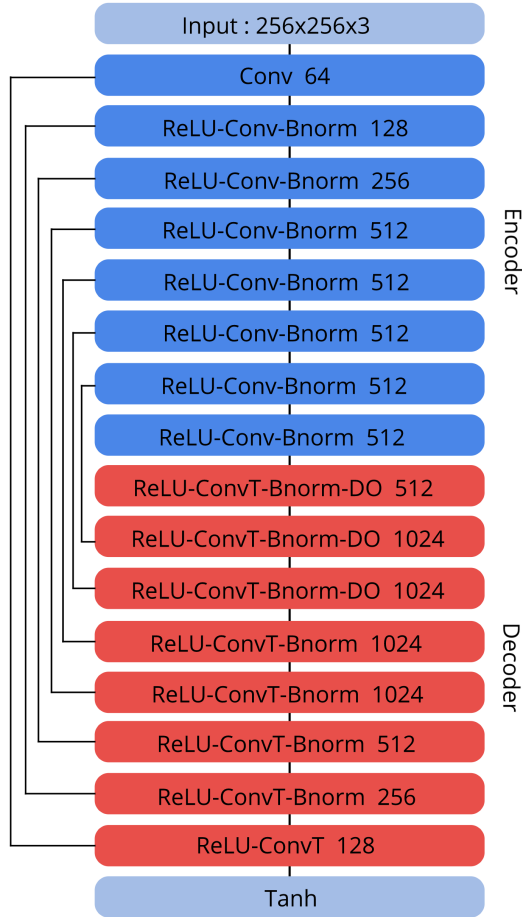


Figure 4.3: Generator architecture is an hourglass shaped model with skip connections. The number associated with each block signifies the number of its convolutional filters. All convolutional layers have the same parameters (with the exception of the number of filters): kernel size is 4×4 , stride is 2, padding is 1. All ReLU layers in the encoder are leaky with the slope of 0.2, whereas in the decoder there is no leak. Dropout rate is 0.5. Connections are indicated by black lines.

² Based on observation of previous experiments.

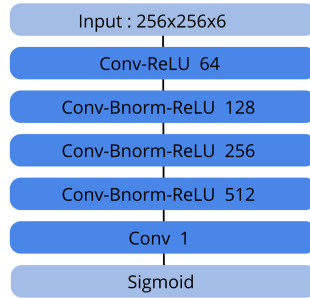


Figure 4.4: Discriminator architecture. The GAN framework is conditioned on LR image. Therefore, the input has 6 channels - concatenated generated/real image and LR image. All convolutions have same parameters: kernel size is 4×4 , stride 2 and padding 1. ReLU layer has a slope 0.2. The output reflects a probability of the input data being from the real data distribution.

4.2.1 Chaining the networks: cGAN + CNN chain

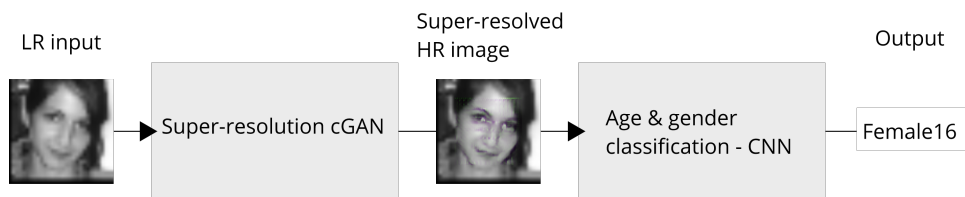


Figure 4.5: Illustration of the super-resolution strategy.

In the super-resolution strategy the cGAN and CNN are connected as shown in Figure 4.5. We tested four options to implement the connection of the two networks:

- (i) **cGAN + CNN - not-trained** - two networks are trained independently and connected only in testing phase.
- (ii) **cGAN + CNN - fixed CNN** - both networks are connected during training but weights of the CNN classifier are not updated.

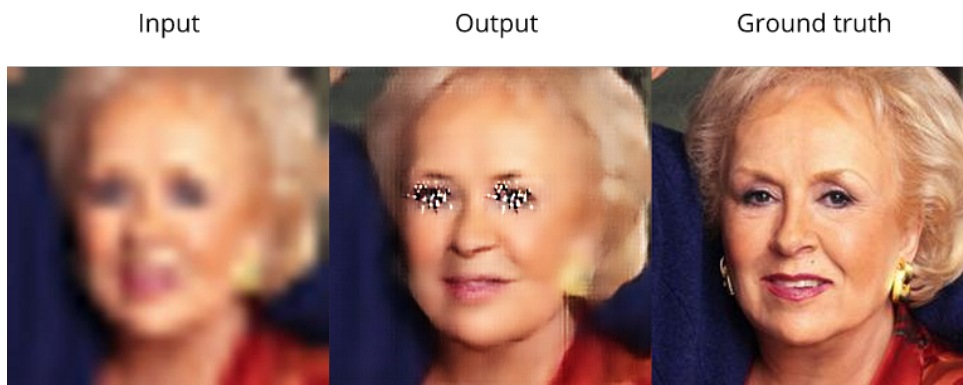


Figure 4.7: Example of artifacts when training the cGAN on images with a larger context around the face.

4.3 Implementation details

To connect both networks, additional operations had to be arranged. Mainly, due to the difference in cGAN output size (256×256 px internal pix2pix [14] network input size) and CNN input size (100×100 px) and the fact that input images needed to be striped off the context, super-resolved and subsequently padded with low-resolution context for age and gender prediction. Operations (such as resizing by bilinear interpolation, converting from RGB to grayscale, padding the super-resolved image with low-resolution context, etc.) of the connection of cGAN and CNN are differentiable, which allows the chain to be trained. Scripts for training were written in pytorch that has an automatic differentiation package. Operations on torch tensors are differentiated automatically.

Training times. The training of the originally trained cGAN with 2419 images took approximately 1 day. The training times of chained cGAN and CNN varied from 2 to 6 days.

Chapter 5

Experiments

First, we introduce the test set, the error statistics for accuracy of age and gender prediction and for super-resolution accuracy. Then we present experiments comparing the baseline with two proposed strategies and variants. Namely: **Data-augmentation, cGAN + CNN chain - not-trained, cGAN + CNN chain - end-to-end, cGAN + CNN chain - fixed cGAN** and **cGAN + CNN chain - fixed CNN**.

5.1 Dataset

Experiments on age and gender were conducted on labeled test dataset (Chapter 4, Dataset) containing 16013 grayscale images with zero overlap with training set. As explained in (Chapter 4, Dataset), detected images were resampled to 100×100 pixels. Images were downsampled by scale $s \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$ and upsampled with bilinear interpolation. Statistics were computed over all 16013 images for each scale s .

5.2 Error Statistics

5.2.1 Age and Gender prediction accuracy

MAE - Mean Absolute Error. Average magnitude of age prediction error. Where \hat{y}_i denotes the predicted age and y_i the true age for the i -th input image. N is the number of test samples.

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i| \quad (5.1)$$

CS k - Cumulative score at k . Ratio of predictions differing from the ground truth by less than or equal to k . Where \mathbb{I} are the Iverson brackets.

$$\text{CS}k = \frac{100}{N} \sum_{i=1}^N \mathbb{I}[|y_i - \hat{y}_i| \leq k] \quad (5.2)$$

gerr - Gender error. Ratio of incorrectly classified samples. Where \hat{g}_i denotes the predicted gender and g_i the true gender for the i -th input image.

$$\text{gerr} = \frac{100}{N} \sum_{i=1}^N \llbracket g_i \neq \hat{g}_i \rrbracket \quad (5.3)$$

5.2.2 Super-resolution

MSE - Mean square error. Average of squared differences between every pixel intensities in output and target images. For images I, K with size $m \times n \times c$ is MSE defined as

$$\text{MSE}(I, K) = \frac{1}{mnc} \sum_{i=1}^m \sum_{j=1}^n \sum_{k=1}^c [I(i, j, k) - K(i, j, k)]^2. \quad (5.4)$$

PSNR. Peak signal to noise ratio. PSNR for image I with reference image K is computed as

$$\text{PSNR}(I, K) = 10 \log_{10} \left(\frac{\text{MAX}_I^2}{\text{MSE}(I, K)} \right). \quad (5.5)$$

Although the PSNR is a widely used statistics to assess the quality of the super-resolution, it does not reflect well subjective human visual perception. Paper [31] studied several error statistics and measured their correlation with human perception. The highest reported correlation was *information fidelity criterion* (IFC) [32]. Although, the conclusion may be different for face images in particular.

5.3 Baseline

The experiment evaluates the impact of low resolution input to prediction accuracy of CNN trained on high-quality images. The CNN [18] was trained for 16 epochs using labeled dataset (Chapter 4, Dataset). As can be seen in Figures 5.1, 5.2, 5.3, 5.4, Baseline method copes with resolutions larger than 60×60 pixels (scale 0.6). On smaller scales, the prediction accuracy gradually diminishes. That is due to the fact that the network was trained only on high-quality images.

5.4 Data-augmentation

The objective of the experiment was to adapt the CNN [18] to low-resolution images and compare its accuracy to the **Baseline method**. The CNN was trained on the same labeled dataset (described in Chapter 4, Dataset) for 58 epochs. Data was augmented by downsampling to scales [1,0.5,0.25,0.125] randomly with uniform distribution and upsampled to the original 100×100 px size via bicubic interpolation. Figures 5.1, 5.2, 5.3, 5.4 show that, as expected,

Data-augmentation surpassed the baseline method on low-resolution images, due to the fact that the Data-augmentation was adapted to LR images during training. Whereas on larger scales, the performance of Data-augmentation worsens only slightly.

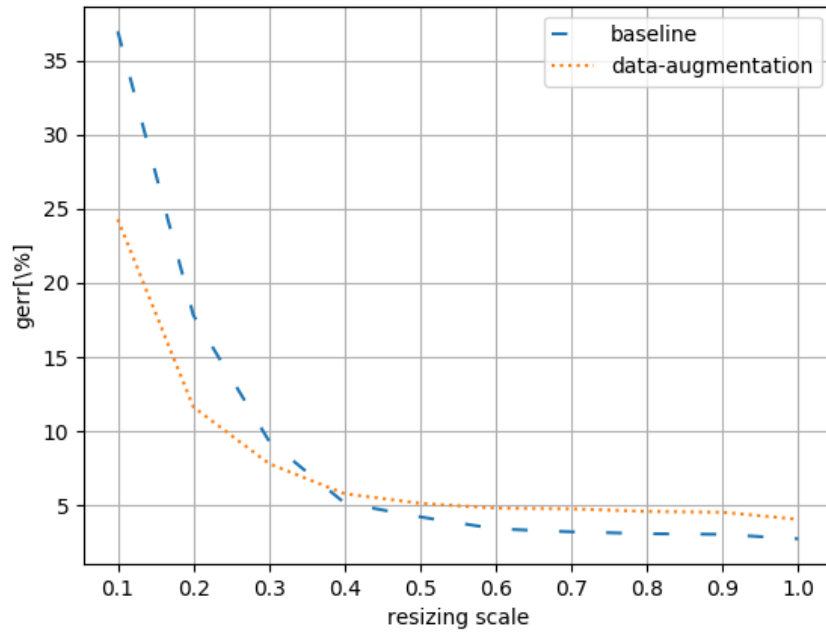


Figure 5.1: Baseline/Data-augmentation comparison - Gender error.

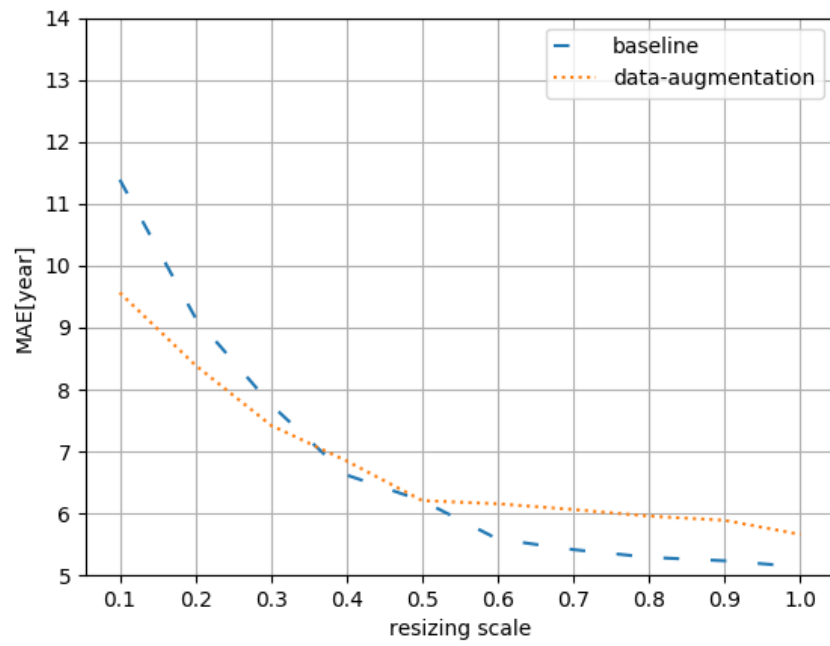


Figure 5.2: Baseline/Data-augmentation comparison - Mean Absolute Error.

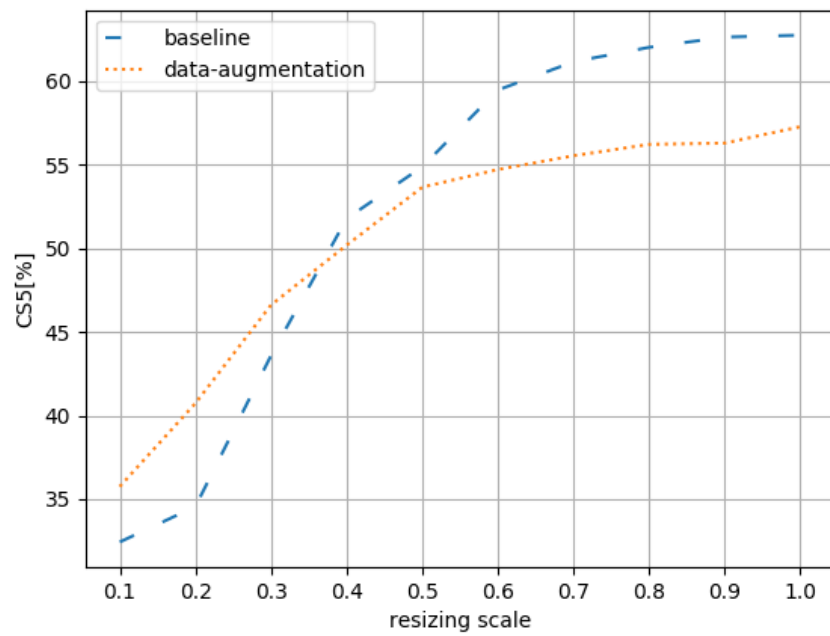


Figure 5.3: Baseline/Data-augmentation comparison - Cumulative Score at 5.

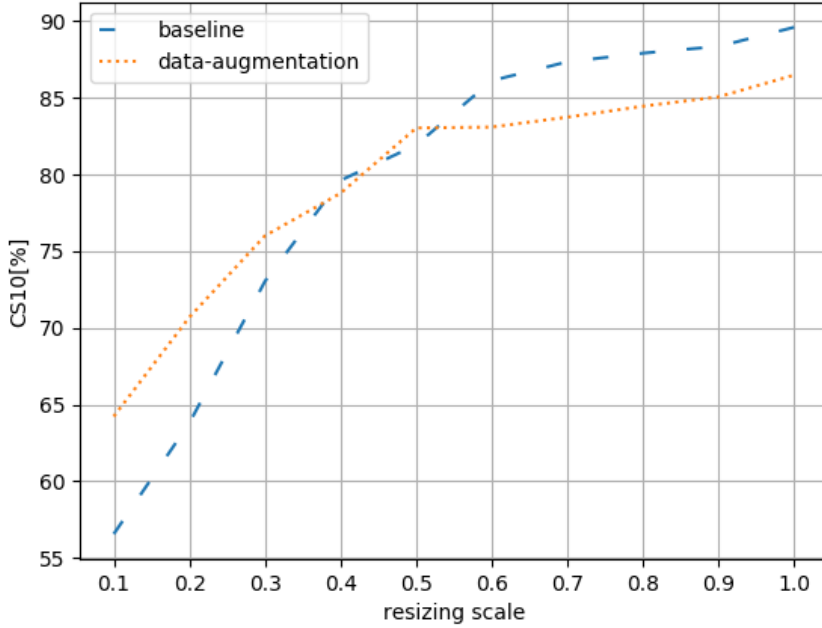


Figure 5.4: Baseline/Data-augmentation comparison - Cumulative Score at 10.

5.5 Super-resolution cGAN

The experiment focuses solely on the assessment of the super-resolution quality of low-resolution images. The cGAN was trained to enhance the resolution of facial images. The training started from scratch on 2419 facial images with resolution 224×224 px for 200 epochs using pix2pix framework [14], explained in detail in section 4.2. None of the subjects were presented simultaneously in both test and train set. Weights were initialized randomly from a normal distribution $\sim \mathcal{N}(0, 0.02)$. Resolution degradation was simulated by downsampling the input image randomly by scales $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$ with probabilities $\{0.2, 0.15, 0.1, 0.1, 0.1, 0.1, 0.1, 0.05, 0.05, 0.05\}$ respectively, and by upsampling to the original size via bicubic interpolation.

Figures 5.5, 5.6 demonstrate that cGAN is capable of producing sharp edges and enhancing visual quality. Although certain imperfections are still present in the reconstructed images i.e. corruptions by generated artifacts, blurry regions. The reconstruction quality would most likely improve with more training images.

In Figure 5.7, PSNR performance of cGAN is worse than of bilinear interpolation. The reason is that PSNR metric, as we noted in section 5.2, is not correlated with human quality perception and does not take in account edge sharpness and image structure.




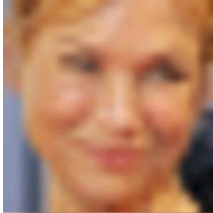


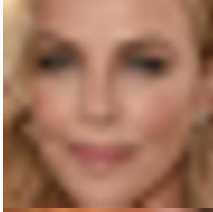





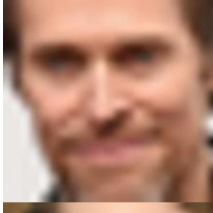


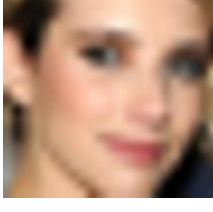
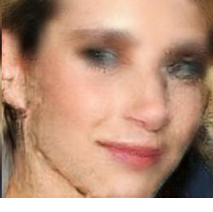

Input	Output	Ground truth	PSNR [dB]
			25,5976
			22,1802
			23,4348
			21,8126
			22,1410
			20,1801

Figure 5.5: Examples of super-resolved images with trained cGAN. PSNR of the output image.












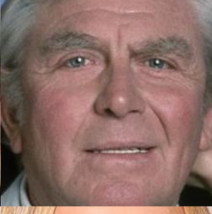
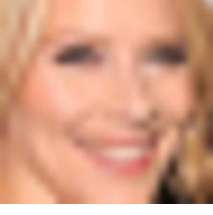


Input	Output	Ground truth	PSNR [dB]
			20,5054
			22,7327
			24,6686
			23,8506
			21,1220

Figure 5.6: Examples of super-resolved images with trained cGAN. PSNR of the output image.

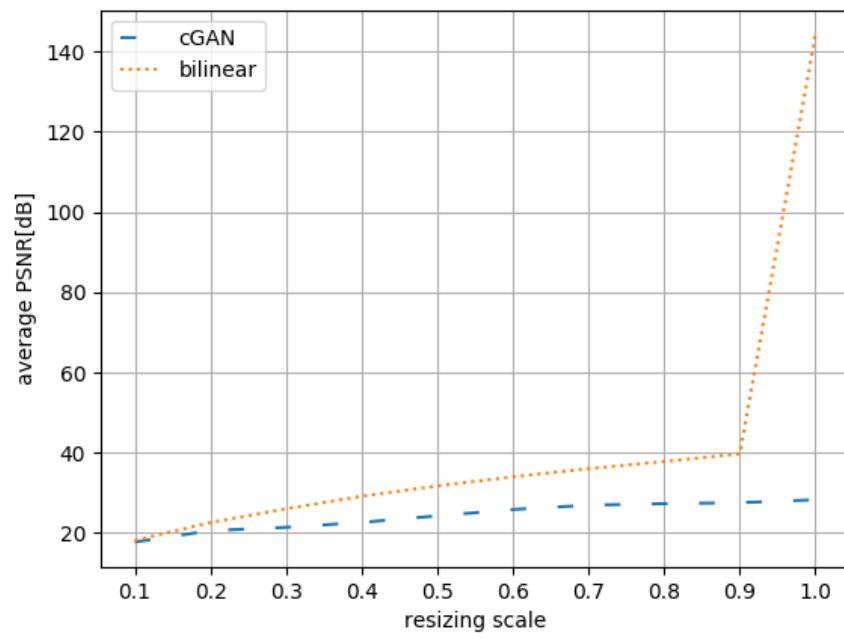


Figure 5.7: Comparison between trained cGAN and the input image upsampled with bilinear interpolation - PSNR.

5.6 Chained networks: cGAN + CNN

The experiment evaluates the second strategy for improvement of prediction accuracy on LR images. We compare it with **Baseline** method. To connect the chain, we used pretrained networks from section 5.5 and from section 5.3. The training was performed on labeled dataset (Chapter 4, Dataset). We tested four alternatives explained in detail in section 4.2.1. The **cGAN + CNN - end-to-end** was trained for 4 epochs, **cGAN + CNN - fixed cGAN** for 10 epochs and **cGAN + CNN - fixed CNN** for 20 epochs.

The results are shown in Figures 5.8, 5.9, 5.10, 5.11, 5.12, 5.13.

Super-resolution quality is shown in Figures 5.8, 5.9. The **cGAN + CNN - not-trained** produces surprisingly worse results than other alternatives of the connection. The reason is most likely the under-fitting (cGAN trained independently was trained on 2419 images while the chain variations were trained on 68,238 images).

Results of age and gender prediction (Figures 5.10, 5.11, 5.12, 5.13) show that the naive connection of the networks trained independently (**cGAN + CNN - not-trained**) performed even worse than the **Baseline**. That is caused by the artifacts produced by cGAN and that probably confounds the CNN classifier.

On **cGAN + CNN - trained end-to-end**, training improved performance over the **Baseline** method on smallest scales but worsened on larger. That might be caused by the fact that smaller scales were prioritized during training (as is the case of super-resolution alone, section 5.5).

In comparison with **cGAN + CNN - trained end-to-end**, the performance of **cGAN + CNN - fixed cGAN** worsened on all scales. However, on the smallest scales, the **cGAN + CNN - fixed cGAN** outperformed the **Baseline**.

The best performance of all chain alternatives brought **cGAN + CNN - fixed CNN**, which performed the best on low-resolution images in particular. The conclusion holds for all age and gender error statistics. The reason for this result is the fact that the originally trained cGAN was most likely under-fitted.

Therefore, we use the **cGAN + CNN - fixed CNN** for the final comparison of strategies for prediction accuracy improvement on LR images.

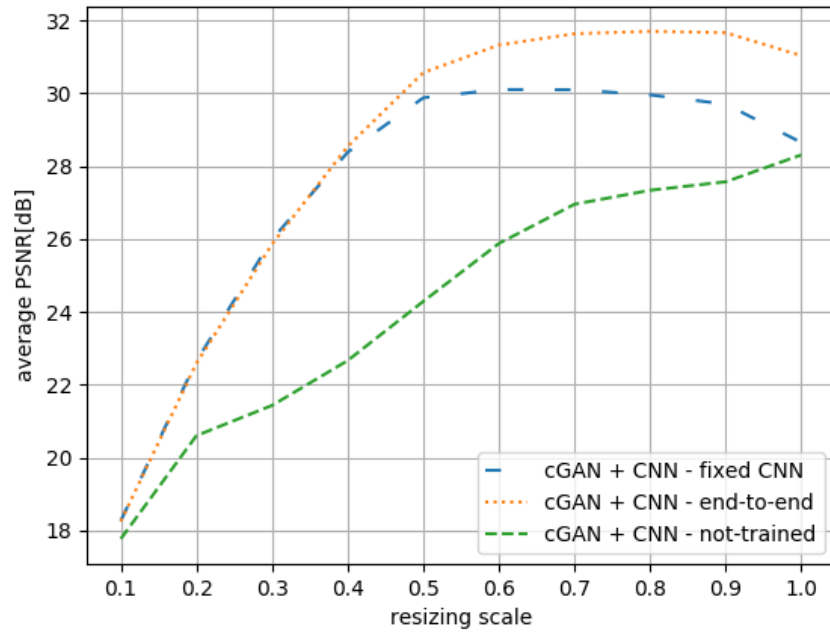


Figure 5.8: Chain comparison - Average PSNR. Note that the **cGAN + CNN - not-trained** is the same as **fixed cGAN**.

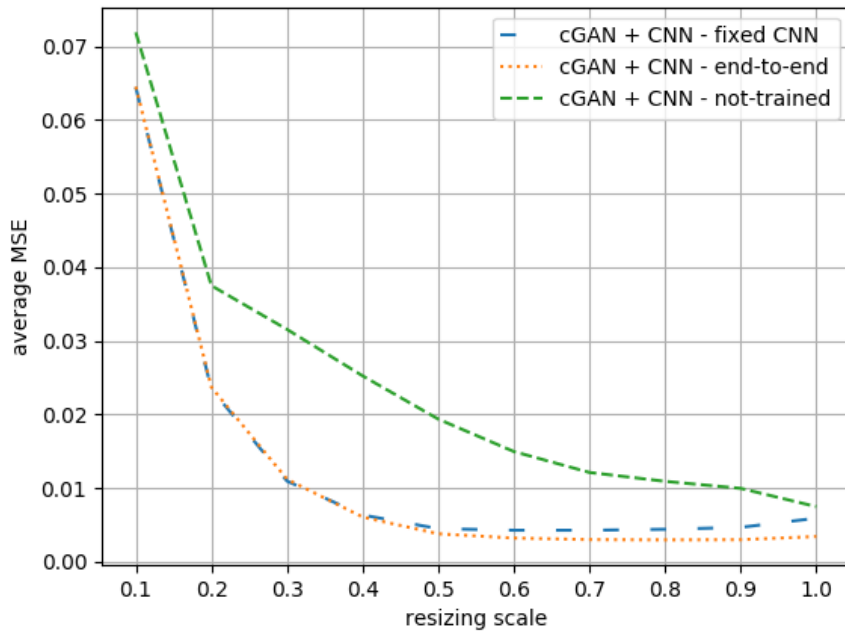


Figure 5.9: Chain comparison - Average MSE. Note that the **cGAN + CNN - not-trained** is the same as **fixed cGAN**.

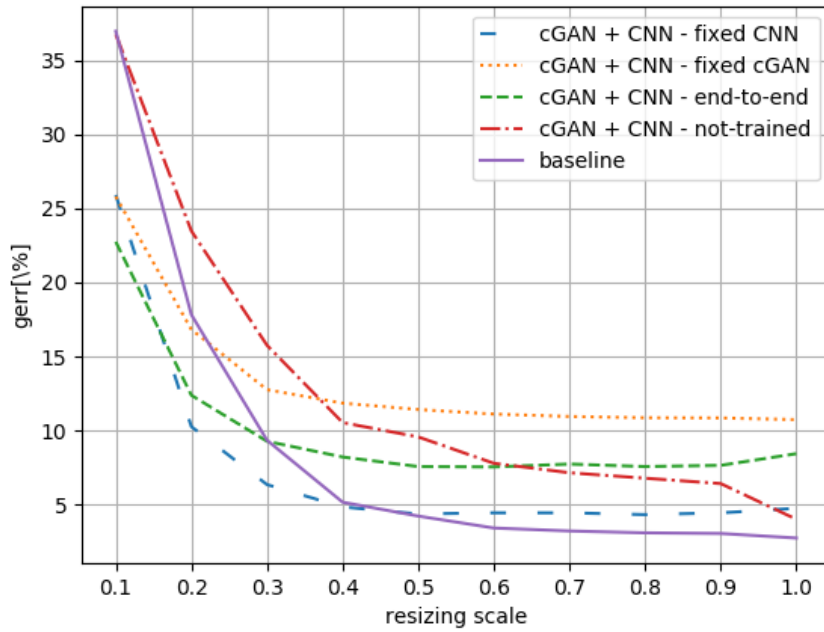


Figure 5.10: Chain comparison - Gender error.

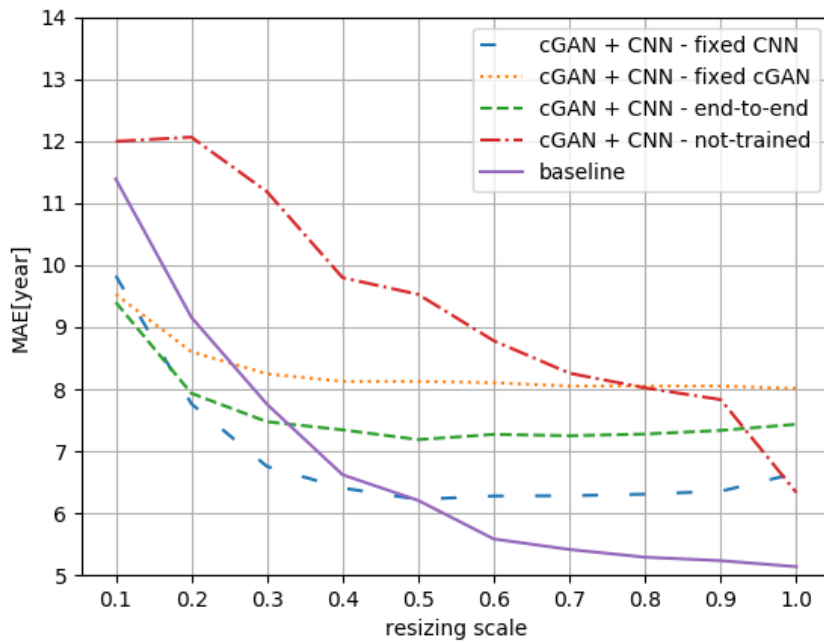


Figure 5.11: Chain comparison - Mean Absolute Error.

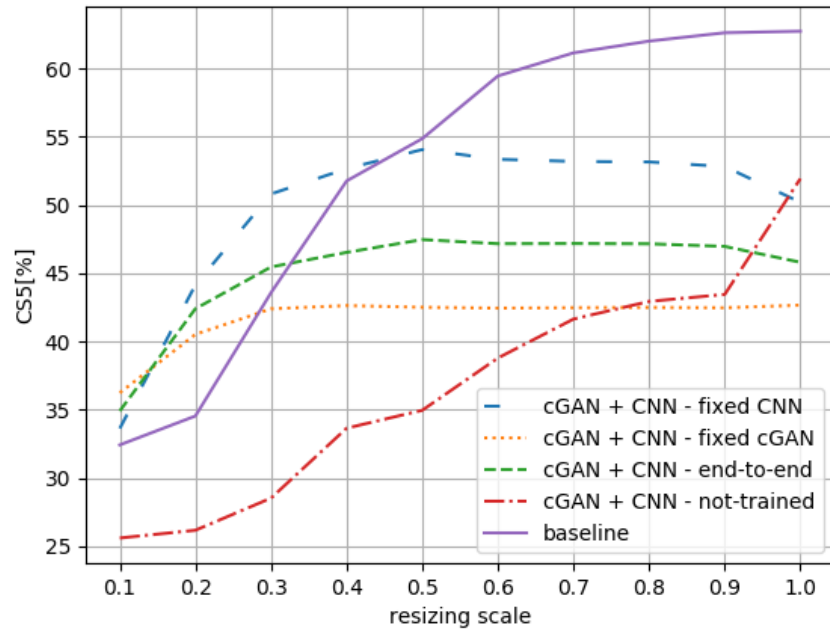


Figure 5.12: Chain comparison - Cumulative Score at 5.

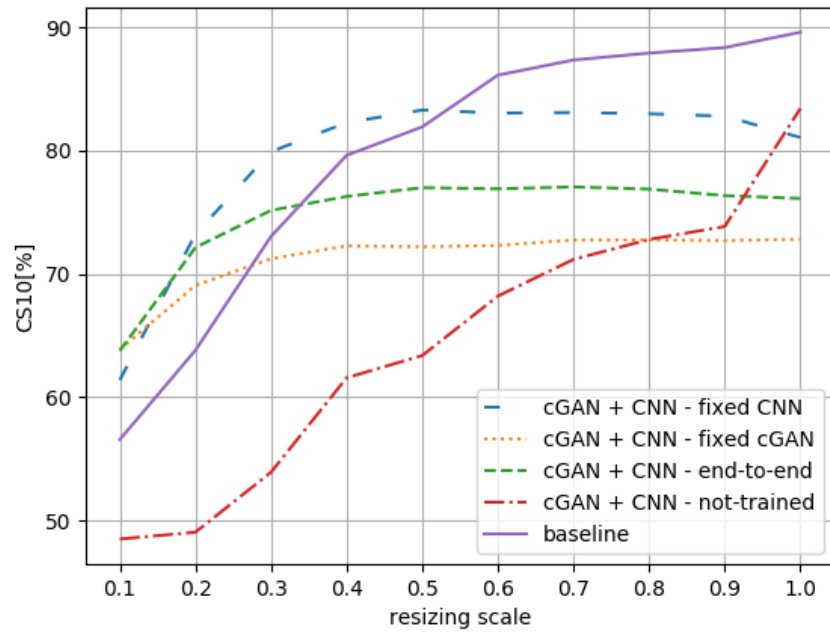


Figure 5.13: Chain comparison - Cumulative Score at 10.

5.7 Final comparison

In this section, we compare strategies for prediction accuracy improvement on LR images - **Data-augmentation** and **Super-resolution** - with the **Baseline** method. Final results are shown in figures 5.14, 5.15, 5.16, 5.17.

Both proposed strategies outperformed the **Baseline** method on low-resolution images and had a comparable accuracy on larger scales. Super-resolution strategy outperformed both **Baseline** and **Data-augmentation** methods on LR images. In larger resolution, there is slightly worse performance, which can be caused by the fact that the smaller scales were prioritized during training phase.

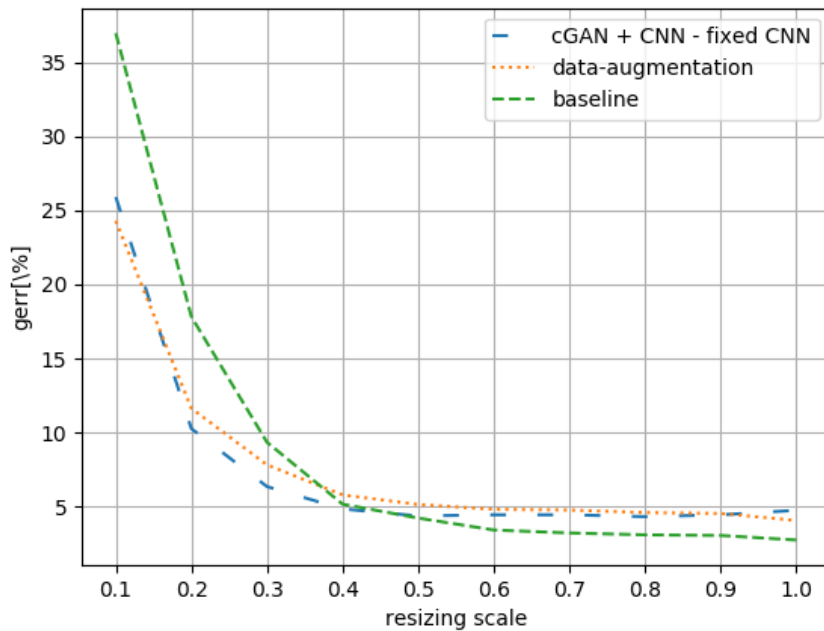


Figure 5.14: Final comparison - Gender error.

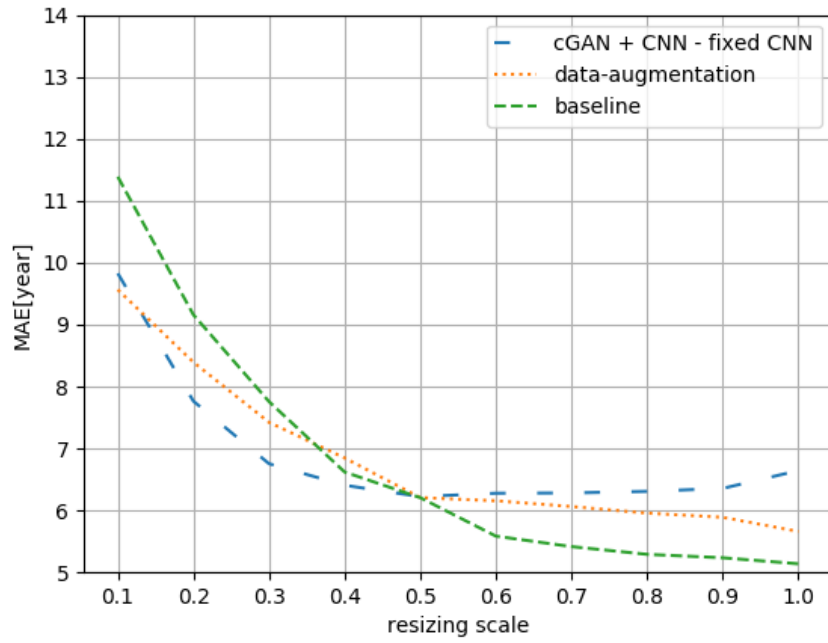


Figure 5.15: Final comparison - Mean Absolute Error.

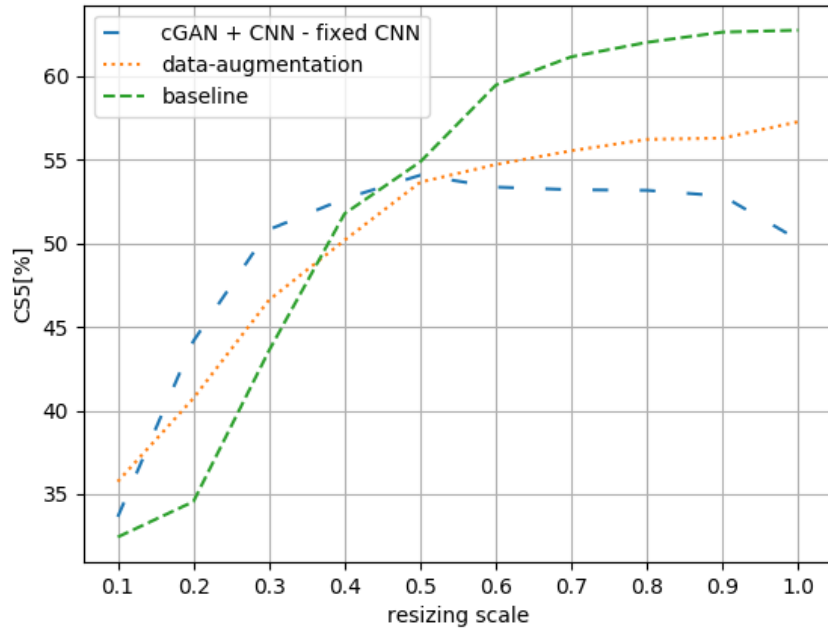


Figure 5.16: Final comparison - Cumulative Score at 5.

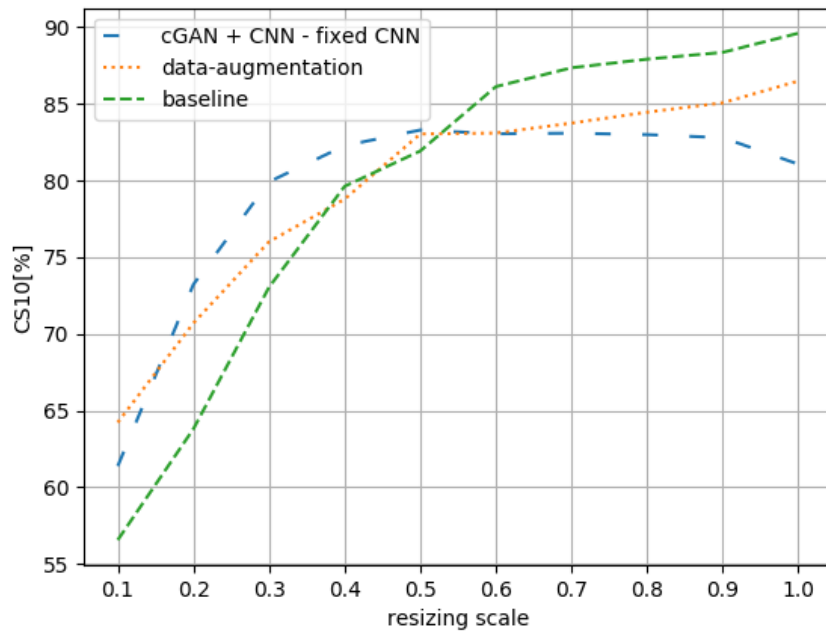



Figure 5.17: Final comparison - Cumulative Score at 10.



Chapter 6

Conclusion

The thesis analyzed and compared proposed strategies for prediction accuracy improvement on low-resolution images: data-augmentation and super-resolution. Data-augmentation adapts a CNN age and gender classifier to LR images. The super-resolution enhances the image resolution using cGAN and subsequently predicts age and gender using a CNN classifier trained purely on high-quality images. Unlike in the case of the data-augmentation, the super-resolution method provides an interpretation of produced attributes (age and gender), because the results can be confronted with high-resolution image.

We tested 4 alternatives of super-resolution strategy. Connecting cGAN and CNN only for testing phase without joined training. Training jointly both networks end-to-end or with fixed weights of either network. The experiments with super-resolution method showed that best performance gave a connection with fixed weights of CNN age and gender classifier.

When comparing the super-resolution, the data-augmentation and the baseline method, it showed, that both proposed strategies for improving prediction accuracy on LR images outperformed the baseline method. Super-resolution had slightly better performance than data-augmentation method.

Despite promising results, we are aware that full potential of the cGAN super-resolution was not fully exploited. The training of cGAN does not need labeled images, however, due to limited computational resources we trained from rather modest dataset of unlabeled 2419 images and of 67k images from the labeled dataset.

In the same spirit, we trained from mixed resolutions instead of training several scale-specialized CNNs. That was probably sub-optimal, since in the most real-world applications, the real resolution is often known.



Bibliography

- [1] Jayasree T.V., Arun Kumar M.N., Multiscale Similarity Learning Single Image Super-resolution with Fast Edge Preserved Reconstruction, *Fourth International Conference on Advances in Computing and Communications (ICACC)*, 2014.
- [2] Kui Jia, Xiaogang Wang, Xiaoou Tang, Image Transformation Based on Learning Dictionaries across Image Spaces, *IEEE Transactions on Pattern Analysis and Machine Intelligence* (Volume: 35, Issue: 2, Feb. 2013).
- [3] Jianchao Yang, Zhaowen Wang, Zhe Lin, Scott Cohen, Thomas Huang, Coupled Dictionary Training for Image Super-Resolution, *IEEE Transactions on Image Processing* (Volume: 21, Issue: 8, Aug. 2012)
- [4] Tingrong Yuan, Fei Zhou, Wenming Yang, Qingmin Liao, Image super-resolution via Kernel regression of sparse coefficients, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.
- [5] Le An, Bir Bhanu, Improved image super-resolution by Support Vector Regression, *The 2011 International Joint Conference on Neural Networks (IJCNN)*.
- [6] Jie Xu, Cheng Deng, Xinbo Gao, Dacheng Tao, Xuelong Li, Image super-resolution using multi-layer support vector regression, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.
- [7] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang, Learning a Deep Convolutional Network for Image Super-Resolution, *arXiv:1501.00092*, 2014.
- [8] Dmitry Ulyanov, Andrea Vedaldi, Victor Lempitsky, Deep Image Prior, *arXiv:1711.10925*, 2017
- [9] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, Wenzhe Shi, Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network, *In CVPR*, 2017.

- [10] Scott Reed, Zeynep Akata, Xinchun Yan, Lajanugen Logeswaran, Bernt Schiele, Honglak Lee, Generative Adversarial Text to Image Synthesis, *arXiv:1605.05396*, 2016
- [11] Grigory Antipov, Moez Baccouche, Jean-Luc Dugelay, Face aging with conditional generative adversarial networks, *arXiv:1702.01983v2*, 2017
- [12] Yaniv Taigman, Adam Polyak, Lior Wolf, Unsupervised cross-domain image generation, *arXiv:1611.02200*, 2016
- [13] Jun-Yan Zhu, Taesung Park, Phillip Isola, Alexei A. Efros, Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks, *arXiv:1703.10593v4*, 2018
- [14] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, Alexei A. Efros, Image-to-Image Translation with Conditional Adversarial Networks, *arXiv:1611.07004v2*, 2017
- [15] W. Zhao, R. Chellappa, A. Krishnaswamy, Discriminant analysis of principal components for face recognition, *Third IEEE International Conference on Automatic Face and Gesture Recognition*, 1998.
- [16] Shumeet Baluja, Henry A. Rowley, Boosting Sex Identification Performance, *International Journal of Computer Vision* 71 (1)
- [17] B. Moghaddam, Ming-Hsuan Yang, Gender classification with support vector machines, *Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, 2000.
- [18] Vojtech Franc, Jan Cech, Learning CNNs for face recognition from weakly annotated images, *IEEE 12th International Conference on Automatic Face & Gesture Recognition*, 2017.
- [19] M. Irani, S. Peleg, Super resolution from image sequences, *10th International Conference on Pattern Recognition, 1990. Proceedings*.
- [20] Chao Dong, Chen Change Loy, Kaiming He, Xiaoou Tang, Learning a Deep Convolutional Network for Image Super-Resolution, *ECCV 2014: Computer Vision – ECCV 2014 pp 184-199*
- [21] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio, Generative Adversarial Nets, *arXiv:1406.2661v1*, 2014.
- [22] Mehdi Mirza, Simon Osindero, Conditional Generative Adversarial Nets, *arXiv:1411.1784v1*, 2014.
- [23] Neeraj Kumar, Alexander C. Berg, Peter N. Belhumeur, Shree K. Nayar, Attribute and Simile Classifiers for Face Verification, *IEEE International Conference on Computer Vision (ICCV)*, 2009.

- [24] Gary B. Huang, Manu Ramesh, Tamara Berg, Erik Learned-Miller, Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments, *Tech. Report 07-49, University of Massachusetts, Amherst*.
- [25] Gabriel Panis, Andreas Lanitis, Nicholas Tsapatsoulis, Timothy F. Cootes, Overview of research on facial ageing using the FG-NET ageing database, *IET Biometrics, Volume: 5, Issue: 2*, 2016.
- [26] Sergio Escalera, Junior Fabian, Pablo Pardo, Xavier Baró, Jordi González, Hugo J. Escalante, Dusan Misevic, Ulrich Steiner, Isabelle Guyon, ChaLearn Looking at People 2015: Apparent Age and Cultural Event Recognition Datasets and Results, *IEEE International Conference on Computer Vision Workshop (ICCVW)*, 2015.
- [27] Grigory Antipov, Moez Baccouche, Sid-Ahmed Berrani, Jean-Luc Dugelay, Apparent Age Estimation from Face Images Combining General and Children-Specialized Deep Learning Models, *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2016.
- [28] Michal Uricar, Radu Timofte, Rasmus Rothe, Jiri Matas, Luc Van Gool, Structured Output SVM Prediction of Apparent Age, Gender and Smile From Deep Features, *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2016.
- [29] Guodong Guo, Yun Fu, Charles R. Dyer, Thomas S. Huang, Image-Based Human Age Estimation by Manifold Learning and Locally Adjusted Robust Regression, *IEEE Transactions on Image Processing* (Volume: 17, Issue: 7, July 2008).
- [30] Yun Fu, Thomas S. Huang, Human Age Estimation With Regression on Discriminative Aging Manifold, *IEEE Transactions on Multimedia* (Volume: 10, Issue: 4, June 2008).
- [31] Chih-Yuan, YangChao, MaMing-Hsuan Yang, Single-Image Super-Resolution: A Benchmark, *ECCV 2014: Computer Vision – ECCV*, 2014 pp 372-386
- [32] H.R. Sheikh, A.C. Bovik, G. de Veciana, An information fidelity criterion for image quality assessment using natural scene statistics, *IEEE Transactions on Image Processing* (Volume: 14, Issue: 12, Dec. 2005)



Appendix A

CD contents

- Project_LaTeX - folder containing LaTeX source codes, images and graphs
- thesis_subrtade.pdf - The thesis in pdf format.