DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

IMAGE RESTORATION BASED ON CONVOLUTIONAL NEURAL NETWORKS

RESTAURACE OBRAZU KONVOLUČNÍMI NEURONOVÝMI SÍTĚMI

THESIS SUMMARY

AUTHOR PAVEL SVOBODA

SUPERVISOR PROF. DR. ING. PAVEL ZEMČÍK

BRNO 2016

ABSTRACT

A merit of this thesis is to introduce a unified image restoration approach based on a convolutional neural network which is to some degree degradation type independent. Convolutional neural network models were trained for two different tasks, a motion deblurring of license plate images and a removal of artifacts related to lossy image compression. The capabilities of such models are studied from two main perspectives. Firstly, how well the model can restore an image compared to the state-of-the-art methods. Secondly, what is the model's ability to handle several ranges of the same degradation type.

An idea of the unified end-to-end approach is based on a recent development of neural networks and related deep learning in a field of computer vision. The existing hand-engineered methods of image restoration are often highly specialized for a given degradation type and in fact, define state of the art in several image restoration tasks. The end-to-end approach allows to directly train the required model on specifically corrupted images, and, further, to restore various levels of corruption with a single model.

For motion deblurring, the end-to-end mapping model derived from models used in computer vision is deployed. Compression artifacts are restored with similar end-to-end based model further enhanced using specialized objective functions together with a network skip architecture.

A direct comparison of the convolutional network based models and engineered methods shows that the data-driven approach provides beyond state-of-the-art results with a high ability to generalize over different levels of degradations. Based on the achieved results, this work presents the convolutional neural network based methods suggesting a possibility having the unified approach used for wide range of image restoration tasks.

KEYWORDS

Convolutional neural networks; deep learning; image restoration; motion deblurring; JPEG artifacts

CONTENTS

1 INTRODUCTION

In 1943 Warren S. McCulloch, a neurophysiologist, together with Walter Pitts, a mathematician, published their work *A logical calculus of the idea immanent in nervous activity* which is in the field of artificial neural networks considered to be one of the first attempts to define and design a model of a very simplified network reflecting the real neural architecture. During more than 70 years, the neural network based models were developed into more complex and in several aspects more by nature inspired architectures. Nowadays, the most visible Artificial Neural Network (NN) impact is in the tasks of speech recognition and computer vision where the ongoing research develops fast and almost continuously reveals new knowledge.

The primary objective of this thesis involves the NN deployment in image restoration which, by its nature, is part of the more general image processing field. Such an idea does not evolve for the first time. However, the presented image restoration is framed by a unified approach based on a data-driven Convolutional Neural Network (CNN) model. This idea is introduced in more detail on two examples of common image restoration tasks such as an image deblurring, i. e. restoring the blurred image into its sharp representation, and an image artifacts removal.

A well-established approach exists to restore the degradation caused by blur which consists of several steps. First, the model of the process blurring the image has to be defined. Based on this model, the so-called Point Spread Function (PSF) is derived. Second, having the PSF, the blurred image can be reversed into its sharp representation using deconvolution. The approach differs in the case of image artifacts removal. The degradation process has to be modeled as well; however, the method to remove or at least to suppress the artifacts is diametrically different from the one for deblurring.

Compared to the traditional engineered methods designed for a particular type of corruption restoration, the NN allows using the same NN based model just trained on different data. A single model used for arbitrary corruption restoration would be the desired outcome, which, considering the capabilities the neural networks have, should not be so much unrealistic. However, this is not the case. This thesis focuses on the utilization of a NN as the primary approach in image restoration, which may differ in training or particular architecture providing significant and state-of-the-art comparable results. There exist various published methods in image processing which make use of NN. The selected restoration tasks are often considered in the examples including traffic surveillance system, the production line monitoring system, or any utiliza-

tion of low-quality image capture devices. In a case of artifacts removal it consists of low-quality bandwidth, i. e. images may be heavily compressed and later restored.

OBJECTIVES The main hypothesis of this thesis and the related objectives can be summarized as follows. *Most of the different image restoration methods is replaceable by a unified approach represented by CNN models which are end-to-end trained and often achieves state-of-the-art or even beyond results.* These models may differ in particular architecture or in the objectives they are trained for. The end-to-end mapping considers the direct transformation from a corrupted representation of a restored image. To provide the evidence showing the validity of such a hypothesis, two various image restoration tasks are selected.

Motion Deblurring The deblurring, namely the motion deblurring, is evaluated on the specific text images including the license plates captured by the surveillance system. In this task, the primary attention will be given on the capability of deblurring itself under the assumption of not known blur parameters, i. e. the model will provide a blind deconvolution. CNN deblurring model will be examined to reveal its capacity which may allow using a single model for a large range of possible blurs.

Artifacts Removal The artifacts comprise a non-linear corruption compared to the linear blur degradation. In this task, the same approach of CNN as in motion deblurring comprise the unified approach. Next to the simple architecture used in the direct end-to-end mapping approach, several different objectives the model is trained for together with an architecture extension are studied. Finally, the CNN based image restoration applied directly on the JPEG coefficients instead of pixels is proposed and described. Considering the deployment of a CNN model in different data domain, the achieved results may support the idea of a single CNN based approach for different tasks of image restoration.

THESIS OUTLINE ENGINEERED IMAGE RESTORATION briefly introduces the motion blur and the high compression artifacts. CNN IMAGE RESTORATION comprises the core hypothesis of this thesis with the detailed description of the objectives framed by the principle idea of a single unified approach. EXPERIMENTS provides the evaluations and results showing the validity of the presented hypothesis and also offers the possible extensions to the introduced models with the hints for further research. CONCLUSION summarizes the whole work, highlights achievements, and with a conclusion based on the results closes this thesis.

2 ENGINEERED IMAGE RESTORATION

Image restoration is generally a transformation of a damaged image on an undistorted image. This chapter introduces the selected degradations and describes various hand engineered widely used methods for their restoration. Image restoration[1], in the scope of this thesis, consists of two different inverse problems. Motion deblurring can be understood as a linear inverse transformation described as deconvolution. In contrary, a restoration of lossy JPEG compression represents the non-linear inverse transformation which, generally, is an ill-posed[2] problem because the transformation can be non-invertible. JPEG compression, with a low-quality setting, produces the blocking artifact and the ringing – Gibbs phenomenon.

In this thesis, an image is understood as a finite matrix. Precisely, an image expressed as a continues function $f(x, y)$ of two coordinates in the plane is sampled into a matrix $M \times N$, where each sample is quantized to an integer value of K intervals [1]. Three types of images are considered, latent[3] image represents an ideal image which does not suffer from any corruption and it is denoted as x. An undistorted image represents the estimation of the latent image and is denoted as $\hat{x}$. Finally, a distorted image is the result of the process modifying a latent image and is denoted as y. In this work, the terms like degradation, corruption, damage, etc., are understood as synonyms for a general process modifying the latent images.

Both types of degradation can be decomposed into an operator applied to a discrete image and additive noise. The model [2] considering the discrete property can be written

$$y = Ux + W\,, \tag{2.1}$$

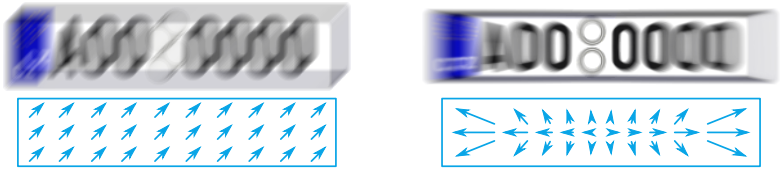
where y is the degraded image, x is the latent image, U represents the discrete operator – motion blur or JPEG artifacts, and W is an additive noise. The discrete operator U can be represented as a linear operation, i. e. convolution, or a non-linear operation, the discrete cosine transform with quantization.

This chapter introduces both of previously mentioned degradations and the methods of its restoration. Motion blur is described with examples of some simple yet typical linear operators and its outcomes. Next, a basic motion blur Point Spread

[1] As both degradations can be well modeled, the inverse transformation is therefore referred as restoration. Image enhancement, on the other hand, does not suppose a strong model.

[2] An incorrectly or improperly posed problem.

[3] The original meaning is related to exposed photosensitive material – photographic film.



(a) (b)

Figure 2.1: An uniform motion blur (a) with a vector field representing the spatial blur and a non-uniform blur (b).

Function (PSF) estimation is introduced to compute the Wiener filter and produce the estimated sharp image. A state-of-the-art text-oriented deconvolution method – L0-regularized intensity and gradient prior [3] is described to be later on compared with the introduced data-driven learned CNN based approach.

JPEG compression based degradation is mentioned with the emphasize on stages of transformation pipeline where the compression artifacts come from. Methods dealing with these artifacts are mentioned with a description of the current state-of-the-art Shape Adaptive Discrete Cosine Transform method [4, 5]. The majority of hand engineered methods usually consist of several steps based on an analytical solution. This chapter briefly introduces several of such methods to highlight the difference between data-driven methods which a CNN is a part of.

2.1 MOTION BLUR

Digital image restoration related, beside others, to the motion blur massively appeared with the space programs in 1950s. The rising amount of aerial pictures taken during the missions were often subject to many photographic degradations – including the motion blur [6]. This is often caused by a shake of a camera or a moving object in the scene. Degraded images can be uniformly or non-uniformly blurred. A convenient example of the easier case – uniform blur can be found in surveillance systems where the camera is fixed and a moving object appears captured with longer exposure. A uniform blur is represented solely by a single PSF applied on the entire image. Non-uniform blur may often be related to an optics distortion, camera rotation, or objects moving in the scene with different speed or in various distances and consists of several PSF describing the blur in a particular part in the image. Both types of blur, uniform and non-uniform is shown in Figure 2.1. Direct solution of (2.2) leads to the inverse filter with all the drawbacks mentioned later. In a case of considering the noise and keeping the assumption of linearity, the Wiener filter is usually used. This

Figure 2.2: The sharp image x blurred with the motion blur PSF g. The result is an uniformly blurred image y.

may be based on known or unknown PSF – in such a case the transformation called a non-blind or blind deconvolution. Often the existing methods of blind deconvolution concentrated in estimating the single blur PSF for the entire image. This is valid for a restricted set of applications but generally, such an assumption is far being satisfied in the case of objects which in the scene move independently.

In case of an uniform motion blur, the equation (2.1) can be derived into a model described as

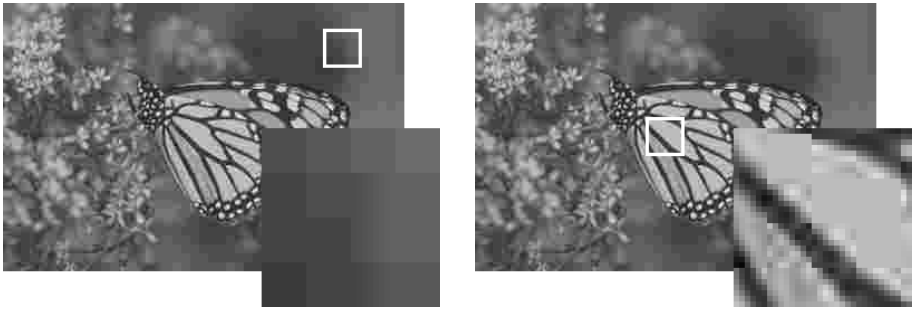
$$y = x * g + w \,, \tag{2.2}$$

where y is the captured motion blurred image, x is the sharp latent image. The operator $\mathcal{U}$ (2.1) becomes the convolution $*$ with a shift invariant PSF g representing a degradation due to motion and optics imperfections, and, finally, w is an additive random noise with zero mean Gaussian distribution. Figure 2.2 shows the example of motion blurred license plate image with the corresponding PSF. The presented model (2.2) rarely, if any, match the realistic conditions, e.g. optics is not exactly shift-invariant, digital imaging sensors do not have the precise Gaussian distribution of noise etc.

2.2 IMAGE COMPRESSION – JPEG

Citing the ITU [7] *Recommendations*, Joint Photographic Experts Group (JPEG) was formed in 1986 to establish a standard for the sequential progressive gray-scale and color images. The abbreviation JPEG used for the file format itself is an informal name for the JPEG File Interchange Format (JFIF) [8] used mainly for images processed by computer software or Exchangeable Image File Format (EXIF) [9] used by imaging cameras. A typical compression ratio of lossy JPEG is approximately 10:1. In the case of the higher compression ratio, the image degradation becomes much more perceptible indicated by the blocking and ringing artifacts Figure 2.3. This section provides the short description of JPEG compression pipeline focusing on the source of artifacts.

JPEG compression artifacts suppression has several considerable applications where data acquisition is expensive, difficult or demanding. For instance, the image or video playing over unreliable or low-bandwidth data connection. Image processing with low compression quality in surveillance systems encompasses application from traffic to production line monitoring. Its massive employment can be in the



(a) (b)

Figure 2.3: The JPEG artifacts in the form of the blocking (a) on the left and ringing (b) on the right which is visible on the edges. The monarch image is here compressed with the quality 10 and it is selected from *LIVE1* image dataset [10].

low-quality images preview in systems where the storage together with bandwidth capacity matters.

2.2.1 *JPEG Compression Pipeline*

The compression pipeline as introduced in [7] consists of various steps which differ according the lossy or lossless compression. The first one, lossy, is Discrete Cosine Transform (DCT) based Figure 2.4 and allows depending on the characteristics of the particular image as well as on desired picture quality to set the required amount of compression. Lossy image compression, generally, achieves high compression ratios through an elimination of information that does not contribute to a human perception of images, or contributes as little as possible. The second one, lossless coding, is based on predictor definition and Huffman or arithmetic coding rather than DCT.

Firstly, the image color space is transformed from RGB to $Y'C_BC_R$ representing the luma Y', C_B blue-difference, and C_R red-difference chroma components. Usually, the chroma components are down-sampled due to lower human sensitivity to colors compared to brightness intensities. Secondly, during encoding, the input image is split into 8×8 blocks which are transformed by the forward DCT into a 64 values referred as the DCT coefficients which represent the particular frequencies the DCT block consist of.

The quantization step in JPEG compression pipeline actually causes a non-linear degradation based on the compression quality. The quantized DC coefficient is then treated separately from the remaining quantized AC coefficients. Its value is based on

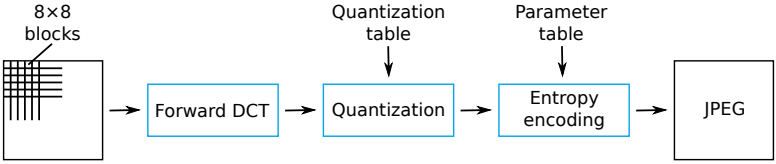


Figure 2.4: The JPEG compression pipeline with the highlighted DCT encoder part.

the difference of the previous DC value i. e. the very first DC coefficient is the reference one for all the subsequent DCs. Next, the coefficients are passed to an entropy encoding process which, losslessly, compress the data. Decoding is proceed in the reverse order, where the dequantized coefficients B are transformed by the Inverse Discrete Cosine Transform (IDCT). The main degradation sources are 8×8 block sampling and related quantization step with following rounding operation.

2.3 SUMMARY ON IMAGE DEGRADATIONS

Two different image degradation types were introduced, motion blur and the JPEG related artifacts. The blur in the image is usually a consequence of a single reason, the long exposure time, which is often caused by several factors. The motion blur is a linear transformation where the image information is not reduced but only transformed. This yields to the straightforward solution, i. e. the deconvolution of the blurred image to restore the latent sharp image. Several related problems can and often do occur like the noise in the image which makes the deconvolution hard and requires specialized approaches. Often the estimation of PSF is performed with the external knowledge represented like the prior as for example the distribution of gradients in the sharp image.

The JPEG artifacts solely caused by the high compression ratio differs from the motion blur primarily in lost image information. The artifact removal is therefore completely different from the methods for deblurring. However, the prior in the form of a regular grid is often used to deal with the blocking artifacts. An important thing to notice is the diversity of approaches the engineered restoration consists of.

3 CNN IMAGE RESTORATION

Image restoration based on CNN representing an unified approach is the core idea of this thesis. The unified method is based on an assumption of a single end-to-end model which directly maps the degraded image on the restored image. This model is purely data driven and in fact, may differ in its architecture which comprises the number of neurons, depth of a model and layers arrangement together with their type. That allows the end-to-end models shift the effort from designing the specific methods towards training objective definitions. The end-to-end approach allows simplifying extend or adapt the model on certain degradation level which involves just to train a model on new data.

The field of image restoration includes various types of degradations. Within the scope of this thesis, two different tasks of restoration were selected. It is the motion deblurring which together with the additive noise represents a linear transformation. The other is artifacts removal approach which deals with a non-linear transformation caused by the quantization step in the JPEG compression pipeline. These two types of degradations are a subject of restoration method based on the data-driven CNN models.

The majority of engineered restoration approaches comprise a particular processing pipeline. The deployment of NN in image restoration is usually associated with a certain step in the pipeline. These are, in fact, the vast majority of image processing NN approaches described in the previous chapter. Namely, the L0 regularized method [3] represents the most recent approach for blurred text image restoration. The Shape Adaptive Discrete Cosine Transform (SA-DCT) [4] is considered to be an up-to-date advanced method for JPEG artifact removal. Both represent the engineered approaches with the first-class results.

However, the recently introduced CNN based end-to-end methods provide significant outcomes often beyond what the widely used engineered approaches can achieve. The recent data-driven methods are represented by the text image denoising CNN [11] or JPEG artifact removal CNN model [12] which is an extension of the super-resolution model [13].

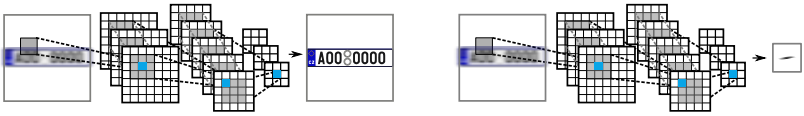
This chapter formulates the CNN based methods for license plate deblurring and JPEG artifacts removal. The presented models are based on almost only on the existing approaches often used in the field of computer vision. Both introduced approaches, compared to the vision related CNN methods, are extended and adapted for the image

restoration requirements, which yields to regression instead of classification models. The main concepts are introduced which were used to train and deploy the network in both image restoration tasks. Follows the description of direct mapping approach. The improvements based on the skip architecture are introduced with the relation to gradient vanishing and neuron exploding problems. Several different objectives of the direct mapping and an initialization proposals are given. The data resampling is proposed to make the objective easier to learn. The chapter is closed by an introduction to the end-to-end approach for the nonimage data restoration focused directly on the JPEG coefficients.

HYPOTHESIS *Most of the different image restoration methods is replaceable by a unified approach represented by CNN models which are end-to-end trained and often achieves state-of-the-art or even beyond results.* These models may differ in particular architecture or in the objectives they are trained for. The term unified covers the data-driven approach which adapts to a particular type of degradation, it does not inherently mean a single model. Different training objectives provide various speeds of convergence and rarely better models as well. The end-to-end mapping considers the direct transformation from a corrupted representation of a restored image. On the other hand, this approach would allow just to obtain a model for a particular type of degradation which needs to be restored. The following text comprises several ideas, assumptions, and considerations framed by the unified CNN based approach for image restoration. Based on the provided experiments, it often does not finally depend on the extensions primarily in the sense of performance, but in particular cases, different train objectives speed up the training in the sense of convergence time.

3.1 END TO END MAPPING

To introduce the end-to-end mapping based on the data-driven learned CNN model, the usual restoration pipelines of deblurring and image artifacts removal are quickly summarized. The common approach of deblurring is to estimate the PSF the image was corrupted with and to use it to restore its sharp representation. The restoration can be computed locally using the deconvolution with the inverse PSF or globally based on some specific global operator. However, estimating the PSF in a case of the blind scenario is an ill-posed problem. Several approaches were presented using the natural image priors, the histogram of gradients in the sharp image distribution, the specific spectrum properties in the frequency domain and other priors. Both steps, the PSF estimation, and the consequent deconvolution are prone to fail due to the noise, significant outliers, and other related causes. Thus, image deblurring is a spe-



(a) (b)

Figure 3.1: The end-to-end mapping (a) is a direct transformation from the degraded image to its restored representation. In contrary, the recent engineered and learned methods (b) usually estimate the PSF to deconvolve the image.

cialized processing pipeline. The methods for JPEG artifacts removal, from the simple yet widely used Simple Postprocessing (SPP) included in FFmpeg up to the SA-DCT utilizing the adaptive shape support to estimate the restoration, present the engineered post-processing approaches. These are entirely different from the methods for deblurring. Being highly specialized is the only common thing they share.

That is not the case of the end-to-end mapping approach considered in this thesis. This data-driven approach is based on the CNN, specifically the Fully Convolutional Network. The general Fully Convolutional Network (FCN) model is trained to process the input image directly. The image is transformed – scattered through the network layers in the feedforward transformation. That consists of gradually applied nonlinearity operators and convolutions. The whole network is trained to estimate the restored image or the error being the difference between the degraded and restored image. The last layer finally outputs the data straight in the pixel format with an arbitrary number of channels. Compared to the majority of previously learned methods, which comprise several steps including PSF estimation and consequent restoration, this approach provides quantitative simplification and simultaneously the qualitative improvement. The direct end-to-end mapping compared to the different approach based on a PSF estimation is shown in Figure 3.1. The definition is written

$$\hat{x} = F_L(\mathcal{W}, y), \tag{3.1}$$

where L defines the number of layers, $\hat{x}$ is the estimated non-degraded latent image x, y represents the input image corrupted by an arbitrary distortion, and $\mathcal{W}$ are the network weights and biases.

The common assumption related to the CNN depth, i. e. number of layers, is that the deeper models provide better results [14, 15]. That is in regards to reviewing the network as a complex data transformation where the layers compose a feature hierarchy representation. This thesis put the emphasize on the end-to-end models considering the ability to generalize over various parameter ranges in a restoration task to provide a single and unified model. The regression model is proposed, which in contrary

to the classification, is generally harder to train[1] together with higher acquirements on the numerical precision. Finally, the end-to-end mapping architecture allows to be quite easily trained for specific parameters in case if needed, i. e. refine the model in case the parameters are roughly known. This approach was initially applied in the text denoising model presented in [11], for superresolution tasks [13, 16], and also for artifacts reduction [12]. Within this thesis, the end-to-end model is studied for two specific yet different image tasks, the motion deblurring, and artifacts removal.

3.2 ARCHITECTURE EXTENSION

Deeper networks may have problems with exploding and vanishing gradients and they may take a long time to learn to propagate information through a large number of layers efficiently. The problems with the gradients can be eliminated by proper initialization [17, 18, 19] which takes effect in the beginning and predicts the overall training speed. The skip architecture influences the network weights during the whole training. This behavior is significant in a case the whole natural image propagates through a deep network in the end-to-end mapping approach.

Training deep models in case of image restoration are still quite a challenge. The problems with propagating information through many layers can be alleviated by bypassing some more deep layers [20]. Such an approach, the skip architecture, can beneficially improve the novel end-to-end methods as it contributes to building a deeper model. The goal of the skip architecture in the image restoration is to allow the network to pass geometric information easily from the input to the output, and to allow for more complex reasoning about the image content in the middle layers, e. g. in case of artifacts removal, what is an artifact and what local context information should be used to restore the image.

An arbitrary CNN model F_L of depth L which utilizes the skip architecture is shown on Figure 3.2 and could be written as

$$\begin{aligned} f_{l\|s}(x) &= h_{l\|s}\left(\mathcal{W}_{l\|s}\left(f_{l-1}(x) \,\|\, f_s(x)\right)\right) \\ F_L(\mathcal{W}, x) &= \left(f_L \circ \ldots \circ f_{l\|s} \circ \ldots \circ f_s \circ \ldots \circ f_1\right)(x) \\ y &= F_L(\mathcal{W}, x), \end{aligned} \tag{3.2}$$

where the operator $\|$ denotes the concatenation and f_s is the skip layer, i. e. the one to be transfered, and $f_{l\|s}$ is the layer to which the skip one is concatenated to. The $f_{l\|s}$ layer is defined as a function which is computed on the concatenated activation maps obtained from f_{l-1} and previous layer f_s. The W denotes the CNN weights,

[1] Classification outcomes are much more limited compared to regression results, namely, compare classifying into two classes and the real number prediction.

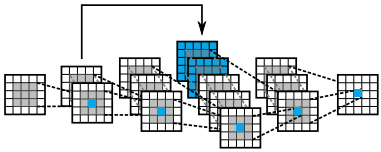


Figure 3.2: The skip architecture allows propagating the low-level image features from the front layer in the network to deeper layers.

trainable parameters including the biases and h_l is an arbitrary activation function. The skip architecture does not have to utilize the concatenation only but can be based on addition as can be found in Long et al. [20] who adds the activations together. The skip architecture utilizing the concatenation of activations from the arbitrary previous layer is proposed to a more challenging task of JPEG artifacts removing.

3.3 SPECIALIZED OBJECTIVES

The end-to-end mapping forces the network to transfer the whole general image through all the convolutional layers interleaved by non-linearities and to restore the degraded image parts while not touching the uncorrupted patches. It shows that such a straight approach requires more training, measured by a number of iteration, however, it does not have to reach the best optimum in a case of restoring complex natural images. Moreover, the learning of such autoencoder-like mapping in situations where the input images are highly correlated with the desired outputs may be wasteful especially for broad and deep networks. It may be one of the main reasons why Dong et al. [12] were not able to scale up their networks and why they required approximately 10^7 iterations to train their AR-CNN. Similar problems were reported by Kim et al. [21].

RESIDUAL In specific tasks, the residual image can be learned instead of a highly variable natural image. Such an idea was first introduced by He et al. [22, 23] for a super-resolution based on the CNN, where the input and output images are highly correlated. The same approach for JPEG artifact removal is almost simultaneously introduced in this thesis which supports layers to learn a residual of their inputs. Instead of training the network to restore the whole image, the task could be defined to only complete the degraded image, i. e. to restore the residual $\triangle x$ between the input corrupted image y and the original latent uncorrupted image x. The residual objective

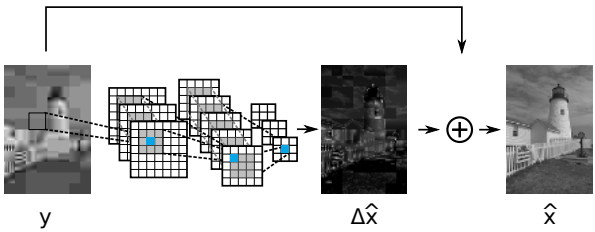


Figure 3.3: Pixel-to-Residual mapping network scheme.

is suitable for the task like JPEG compression artifacts removal, where the repeated blocking artifact occurs. The residual objective is written

$$\arg\min_{\mathcal{W}} \frac{1}{2} \sum_{i=0}^{N-1} \left\| F_L\left(\mathcal{W}, y_i\right) - \triangle x_i \right\|_2^2 , \tag{3.3}$$

where the latent residual is defined as $\triangle x = x - y$. The x corresponds to the ground truth image while the $\hat{x}$ is the result obtained by the CNN processing. The residual learning scheme is shown in Figure 3.3. Kim et al. [21] were able to speed up the training by the factor of up to $10^4\times$ with the residual learning and it allowed them to learn much deeper networks – 20 layers compared to three in [13] and four in [12].

EDGE ENHANCEMENT Mean square error used in many image restoration methods does not necessarily well correlate with the image quality perceived by humans. With convolutional networks, it is relatively easy to use more perceptually valid error measures as long as they can be efficiently differentiated. Therefore, next to the residual objective, the edge enhancement learning is proposed to support the human edge sensitivity perception. The partial first derivatives of the image with the image itself are the inputs into the loss function. The input is in the form of the transformed image x_e defined as

$$x_e = \left[x,\ x * g_x,\ x * g_y\right] , \tag{3.4}$$

where g_x and g_y represent the *Sobel* [24] horizontal respectively vertical operators. The x_e is thus the concatenation of the original image and its horizontal and vertical edge enhanced representation. The objective utilizing the edge priors in y_e and x_e is defined

$$\arg\min_{\mathcal{W}} \frac{1}{2} \sum_{i=0}^{N-1} \left\| F_L\left(\mathcal{W}, y_{e\,i}\right) - x_{e\,i} \right\|_2^2 . \tag{3.5}$$

The scheme of edge enhancement deployed in the network architecture shows Figure 3.4. The assumption is that the addition of the first derivatives should force the

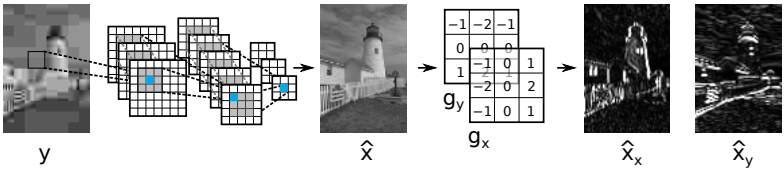


Figure 3.4: Scheme of a restoration network trained with the emphasize on edges.

network to focus specifically on high-frequency structures such as edges, ringing artifacts, and blocking artifacts and it could lead to perceptually better restorations. The combined edge emphasized loss can be easily implemented in all existing convolutional network frameworks by defining the Sobel derivative kernels as a convolutional layer with predefined fixed filters.

PSNR The quality of the restored images is measured is measured in several metrics, e. g. the signal focused Peak Signal to Noise Ratio (PSNR) and more human perception adapted Structural Similarity (SSIM) index [25, 26]. The loss function usually used based on the squared ℓ^2-norm can be with several assumptions swapped to the loss emphasizing function. The network, therefore, can focus on restoring the image to be more visually plausible or to provide better values measured by particularly metric. The loss function based on PSNR is introduced together with its differentiation needed for the backpropagation, i. e. the chain rule. PSNR based on the Mean Square Error (MSE) is defined

$$\operatorname{MSE}(\hat{x}, x) = \frac{1}{MN} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} (\hat{x}_{mn} - x_{mn})^2 \tag{3.6}$$

$$\operatorname{PSNR}(\hat{x}, x) = 10 \log_{10} \left(\frac{\operatorname{MAX}^2}{\operatorname{MSE}(\hat{x}, x)} \right) , \tag{3.7}$$

where $\hat{x} = F_L(\boldsymbol{W}, y)$ is the network restored image and x is the latent uncorrupted image, and MAX represents the maximum intensity value the image can be of, i. e. 1 in the case oh having the image values in the range $[0, 1]$. The loss function based on the PSNR is then defined

$$\arg\min_{\boldsymbol{W}} \left(-10 \log_{10} \left(\frac{\operatorname{MAX}^2}{\operatorname{MSE}(\hat{x}, x)} \right) \right) , \tag{3.8}$$

where the minus sign is present to keep the minimization, i. e. the gradient descent approach. Within the CNN based image restoration, the PSNR objective is proposed. Its differentiation w. r. t. to the input, i. e. the restored image is written

$$\frac{\partial \operatorname{PSNR}(\hat{x}, x)}{\partial \hat{x}} = \frac{\partial 10 \log_{10}\left(\frac{\operatorname{MAX}^2}{\operatorname{MSE}(\hat{x}, x)}\right)}{\partial \hat{x}}, \tag{3.9}$$

which equals to the partial differentiation written in the *Jacobian* matrix yielding to just rescaled error

$$\begin{aligned} k &= 20\left(\log(10)\sum_{m=0}^{M-1}\sum_{n=0}^{N-1}(\hat{x}_{mn} - x_{mn})^2\right)^{-1} \\ \frac{\partial \operatorname{PSNR}(\hat{x}, x)}{\partial \hat{x}} &= \begin{bmatrix} (x_{00} - t_{00})\,k & \dots & (x_{0N} - t_{0N})\,k \\ \vdots & \ddots & \vdots \\ (x_{N0} - t_{M0})\,k & \dots & (x_{MN} - t_{MN})\,k \end{bmatrix}, \end{aligned} \tag{3.10}$$

where the small errors has higher cost compared to the large ones. The interpretation of PSNR loss function in the task of JPEG compression artifacts removal is based on the sensitivity to distortions in the stationary regions of the image like the sky and the clearly visible blocking artifacts in such a region. Finetuning the model could utilize these properties to focus on the ostensibly small errors yet more noticeable compared to high errors in the image areas with heterogeneous structure.

3.4 TASK SPECIFIC MODIFICATIONS

All the mentioned methods operate directly with the image pixels. In a case of a JPEG file, this leads to an additional postprocessing, which is computed after decoding the image. On the other hand, utilizing the technique of JPEG Quality Transcoder (JQT) [27] allows to process the DCT coefficients directly. In this thesis, the new approach of CNN based JPEG file coefficients processing to suppress or remove the high compression related artifacts is proposed. A scheme of such a network which transforms the JPEG coefficients to coefficients representing the restored image is shown in Figure 3.5 where, nevertheless, the loss is computed through the pixels.

The coding and decoding pipeline described in Section 2.2.1 transforms the 8×8 image patches into the 8×8 of DCT coefficients which correspond to specific frequencies in that patch. These coefficients noted as B are sorted based on their frequencies in the *zig-zag* manner. Based on the user specified compression quality the predefined quantization table Q is selected and the DCT coefficients are quantized and rounded. The quantization affects the amount of blocking and ringing artifacts and implicates

two potential types of CNN input, the quantized DCT coefficients $\mathcal{B}$, where the network is forced to learn the quantization table Q as well, and the DCT coefficients $\mathcal{B}$ already per element multiplied by the quantization table, the $Q\mathcal{B}$.

The non-linearity caused by the quantization of otherwise linear DCT transform (**??**) affects the network loss function. The properties of the loss computed on quantized $\mathcal{B}$ or quantization table multiplied coefficients $Q\mathcal{B}$ differ from the loss calculated on the decoded values – the pixels. That means that the network trained on minimizing the loss of coefficients is actually producing different restoration compared to training the network based on the pixel loss. That is given by the different gradients of the loss computed on pixels versus the coefficients $Q\mathcal{B}$. Next, the T.81 recommendation [7] states the IDCT transformed values have to be clipped to fit into the range of the image domain which also influences the loss.

The IDCT layer is defined to being able to compute the loss function directly on the pixels and further backpropagate the loss computed gradients. To follow the chain rule the IDCT differentiation w. r. t. the input dequantized coefficients QB is defined in (3.15). Therefore, the backpropagation through the IDCT layer equals to

$$\frac{\partial \mathcal{F}_c^{-1}(Q\mathcal{B})}{\partial Q\mathcal{B}} \triangle d = \mathcal{F}_c(g) \ , \tag{3.11}$$

where the partial differentiation of the IDCT $\mathcal{F}_c^{-1}$ multiplied by the gradients $\triangle d$ from the layer above is equal to the discrete cosine transform $\mathcal{F}_c$.

The inference of the IDCT differentiation consists of several steps. First, consider to dequantized coefficients $Q\mathcal{B}$ to be denoted as c which is defined as $c = Q\mathcal{B}$. First, the partial differentiation of the $\mathcal{F}_c^{-1}$ w. r. t. c is written

$$\frac{\partial \mathcal{F}_c^{-1}(c)_{mn}}{\partial c} = \begin{bmatrix} \frac{\partial \mathcal{F}_c^{-1}(c)_{mn}}{\partial c_{00}} & \cdots & \frac{\partial \mathcal{F}_c^{-1}(c)_{mn}}{\partial c_{0,N-1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial \mathcal{F}_c^{-1}(c)_{mn}}{\partial c_{M-1,0}} & \cdots & \frac{\partial \mathcal{F}_c^{-1}(c)_{mn}}{\partial c_{M-1,N-1}} \end{bmatrix} , \tag{3.12}$$

where the differentiated element of the Jacobian matrix reduces from the summation to a single expression

$$\frac{\partial \mathcal{F}_c^{-1}(c)_{mn}}{\partial c_{pq}} = \alpha_p \alpha_q \cos\left(\frac{\pi(2m+1)p}{2M}\right) \cos\left(\frac{\pi(2n+1)q}{2N}\right) . \tag{3.13}$$

Second, all the Jacobian matrices (3.12) written in the general expression define the whole 8×8 differentiated patch w. r. t. c in the form of

$$\frac{\partial \mathcal{F}_c^{-1}(c)}{\partial c} = \begin{bmatrix} \frac{\partial \mathcal{F}_c^{-1}(c)_{00}}{\partial c} & \cdots & \frac{\partial \mathcal{F}_c^{-1}(c)_{0,N-1}}{\partial c} \\ \vdots & \ddots & \vdots \\ \frac{\partial \mathcal{F}_c^{-1}(c)_{M-1,0}}{\partial c} & \cdots & \frac{\partial \mathcal{F}_c^{-1}(c)_{M-1,N-1}}{\partial c} \end{bmatrix} . \tag{3.14}$$

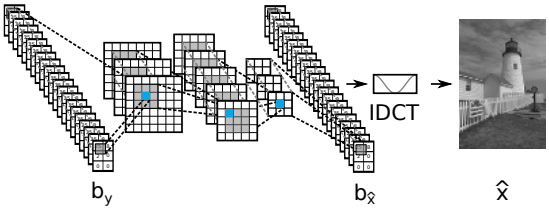


Figure 3.5: DCT-to-Pixel mapping network with a predefined IDCT layer.

Based on this expression, the backpropagation of gradient $\triangle d$ is equal to the summation of per element multiplication of the top layer gradients $\triangle d$ and the corresponding partial differentiations $\partial \mathcal{F}_c^{-1}(c)_{mn}/\partial c_{pq}$. That is written as the equation

$$\mathcal{F}_c(\cdot)_{pq} = \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} \triangle d_{mn} \frac{\partial \mathcal{F}_c^{-1}(c)_{mn}}{\partial c_{pq}}\,, \tag{3.15}$$

which, if expanded, directly equals to the discrete cosine transform $\mathcal{F}_c()$

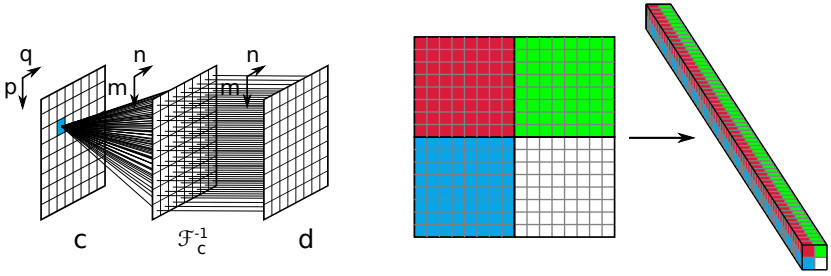
$$\mathcal{F}_c(\cdot)_{pq} = \alpha_p \alpha_q \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} \triangle d_{mn} \cos\left(\frac{\pi(2m+1)p}{2M}\right)\cos\left(\frac{\pi(2n+1)q}{2N}\right). \tag{3.16}$$

The illustration of the gradients backpropagation through the IDCT layer is shown in Figure 3.6a Based on the defined inverse discrete cosine transform layer, the network the decoding layer is deployed in is defined

$$\begin{aligned} F_L(\mathcal{W}, x) &= \left(\mathcal{F}_c^{-1} \circ f_{L-1} \circ \ldots \circ f_1\right)(x) \\ y &= F_L(\mathcal{W}, x)\,, \end{aligned} \tag{3.17}$$

where the loss function is computed directly on the $\mathcal{F}_c^{-1}$ output of a layer, i. e. pixels.

In a case of JPEG artifacts, it is simple to define the prior, e. g. the blocking artifacts occur every 8th pixel. That can be utilized in the form of resampled input which is illustrated in Figure 3.6b. The input patches are resampled from 8×8 into 64D vectors. Meanwhile, the resampled input is proposed to be used with the DCT coefficients. The same technique is introduced for the pixel input data. However, the motivation to resample the data differs in both cases, coefficients and pixels. The resampled input data in the cases of the quantized or dequantized coefficients provides the network the possibility to learn the spatial filters which can utilize the continuity of the related frequencies represented by the coefficients. The resampling, within the pixels based method, is suitable due to the blocking artifact properties, namely its fixed position and repeating structure. Resampling these 8×8 blocks into the 64D channel vectors can directly support the network to utilize the blocking prior.



(a) (b)

Figure 3.6: The illustration of the backward gradient propagation (a) through the IDCT layer from right to left. D are the gradients computed in the loss function. The contributions of all the gradients to every coefficient $Q\mathcal{B}_{pq}$ shows the left part of the figure (a) and is equal to the discrete cosine transform $\mathcal{F}_c(D)$. The 4 pixel blocks of size 8×8 (b) resampled to the 4 64 channel vectors.

3.5 SUMMARY

The general end-to-end mapping Convolutional Neural Network approach has been introduced with several adjustments for the text based image motion deblurring and JPEG artifacts removal. The concept of the end-to-end mapping has been clarified. Nevertheless, it is not an entirely new technique, this thesis puts the emphasize on such an approach because it has an impressive potential to be successfully deployed in a variety of different tasks. The end-to-end data-driven direct mapping has been slightly improved using the skip architecture which concatenates the previous activations to the activations deeper in the network. This skip allows to transfer the features of input data deeper into the network and provide a more complex basis for further reasoning.

A set of specialized objectives has been introduced. These allow the network to focus on a specific subject to learn like the residual learning which is much less model capacity demanding compared to the full image end-to-end approach. An edge enhancement technique based on the Sobel operators has been proposed to support the heterogeneous structures in the images. The loss function based on the PSNR has been introduced to allow the narrowly focused optimization which compared the usually used MSE based loss function forces the network to rate the errors differently.

Finally, in a case of JPEG artifacts removal, the possibility to suppress the artifacts directly in the DCT domain is described. The specialized IDCT layer is proposed to allow the direct end-to-end mapping yet training on the pixel loss function instead of

coefficient loss function which has different properties. The coefficients arrangement allows utilizing the samples-coefficients between connectivity and directly learn the adapted spatial filters. The similar prior and the same approach has been introduced for pixels, where the resampled data organization from 8×8 block to the 64D channel vector allows the network to adapt directly on the fixed blocking artifacts. The extensions and techniques of CNN based model show the applicability and deployment in the tasks of image restoration which is empirically proved later in this work.

4 EXPERIMENTS

The CNN models based on the proposals given in the previous chapter are deployed and studied in the field of image restoration. Namely, it is the motion deblurring of images captured by the surveillance system and the high compression JPEG artifacts removing. Various experiments show the strengths yet also some weaknesses the CNN models have. The presented approach is viewed from two different perspectives. Firstly, the contribution which the proposed methods deliver in comparison with the other widely used approaches is shown. Secondly, the description of how the models behave, which includes the model generalization possibilities, several model extension impact, and other more or less task-specific traits, is presented. Almost all the experiments have very similar structure. This consist of the way data are retrieved, a model specification, a description of the training procedure, and finally the achieved results with their interpretation.

The vast majority of data is artificially produced from the latent undistorted, i. e. ground truth, images. Interestingly, model based on artificial data works very well as it is shown later in this chapter on the image deblurring task. However, it is not so much surprising in the case of artifacts removal, where this is the only way to acquire the training data. It is important to mention that all the experiments were performed using caffe [28] – the fast open framework for deep learning which allowed to concentrate on the model itself instead of the network implementation.

In the beginning, the attention is directed to the deblurring of license plate images [29]. That presents the end-to-end mapping model of 15 layer network. Besides the reported results beyond state of the art, an interesting generalization ability these models have is revealed. Various models are trained for an identical degradation of different levels. That shall provide a perspective how well the CNN approach restores the images of different degradation level compared to blind and non-blind approaches. The part describing the JPEG artifacts removal [30] addresses the majority of the proposed network enhancements including the different objectives, extended architecture, and processing of DCT coefficients instead of pixel.

The last part of this chapter names the possible CNN exploitations in various fields including the surveillance systems, data storage, transfer based services, and user photo-based applications. The future work and possible research directions based on the results of this thesis are outlined. Finally, the very last brief summary closes this chapter.

LIST OF EXPERIMENTS

License plate motion deblurring

- Length range
- Direction range
- Real data deblurring with model trained on artificially blurred data
- Optical Character Recognition (OCR) accuracy comparison with state-of-the-art method

JPEG compression artifacts removal

- Three different architectures, L4, L8, and L5
- Direct, residual, edge enhancement and PSNR objectives
- Comparison with state-of-the art methods
- Generalization over various compression qualities restoration
- Training dataset size impact
- Resampled input
- Coefficient based restoration
- An impact of quantized vs dequantized coeficients

4.1 CNN FOR MOTION DEBLURRING

The majority of methods used for deblurring do not utilize the direct end-to-end mapping. The only exception is the work of Hradis et al. [11] who focused on noise corruption and out of focus blurred text restoration. The other methods deploy the end-to-end mapping but not as an integral solution but more as a subtask [31, 32] which estimates the PSF to be later used in the deconvolution itself. An experiment with the non-blind and blind approach as well is performed on the task of license plate motion deblurring utilizing the 15 layer architecture introduced by [11]. This experiment addresses the model generalization properties and the comparison with blind and non-blind deconvolution approaches .

4.1.1 *Architecture*

This 15 layer fully convolutional network architecture, *L*15 *CNN*, is selected to train the motion deblurring end-to-end mapping model. The reason this model has been

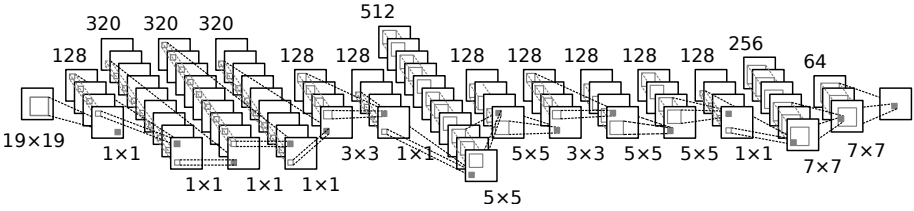


Figure 4.1: L15 architecture with a number of filters per layer, their spatial size, and the preview of grouped channels in the second half of the network. L15 consists solely of convolutional layers followed by the ReLU activation layers (not shown) providing the FCN model.

selected is the success of this model based on the out-of-focus text deblurring results published in [11]. The motion deblurring L15 network definition is written

$$\begin{aligned} L_{15}(\mathcal{W}, y) &= (f_{15} \circ f_{14} \circ \ldots \circ f_1)(y) \\ \hat{x} &= L_{15}(\mathcal{W}, y), \end{aligned} \tag{4.1}$$

where y is the degraded input image, $\boldsymbol{W}$ are the network weights including biases, f_i represents the ith convolutional layer with the consequent activation Rectified Linear Unit (ReLU) function, and $\hat{x}$ is the restored estimation of the latent sharp image x. Besides the formal notation, the Figure 4.1 and **??** show and describe the exact network architecture with several channels grouped together.

The spatial sizes of the network filters and the composition of the layers provides the network with the receptive field of 50 px. The implementation of convolutions yields to 25 px crop of the input image. That is caused by computing the convolutions without any padding. The L15 network architecture consists of grouped data and related filters in its second half. That helps to reduce the total number of parameters. Such an architecture is trained on data generated according to various motion blur parameters, namely the length and direction.

4.1.2 *Data*

All the data the presented network is trained on, are artificially generated. A random blur kernel is computed representing simple linear motion blur PSF. The kernels are generated with the sub-pixel accuracy to cover the generally nondiscrete space, That is achieved by drawing a line representing the motion blur PSF with the 100× scale and finally resampled into the required length using pixel area relation method which gives moiré-free results in image decimation. The final motion blur kernel has odd dimensions. The same technique is used to sample various directions. The drawn line,

representing the motion blur, is rotated based on the sampled direction. This kernel is subsampled into right sized PSF. The motion blurred image is further corrupted by an additive white noise sampled from the user defined parameters. That helps to generate artificially blurred images reflecting the natural images captured in the real-world conditions. Such a data processing allows generating arbitrary linear motion blur PSF used to produce the final blurred image. With the sub-pixel accuracy, the data augmentation allows generating random sized training dataset.

In a case of the end-to-end mapping approach, it is crucial that the ground truth x images are not corrupted. The data used for generating the artificially blurred images are images captured with various imperfections. These are mostly based on the conditions what the real surveillance system operates in. A small fraction of all images was therefore mostly blur distorted or captured in a poor light, i. e. contained high levels of noise. For this reason, the ground truth dataset was processed to filter out the highly corrupted images. The detection of such images was based on an approach based on the high and low-frequency ratio. An ad-hoc threshold was chosen based on the observation to filter out the degraded images. The final dataset consists of 140 k clean and sharp license plate images.

Nevertheless, the disjunct set of naturally blurred data was collected including 721 images of various motion blurred license plates. These were used for verification the model works well on naturally blurred images as well, where the blur PSF usually is not a straight line but reflects some curved trajectory. These images were taken by two static surveillance cameras controlling the road under different angles with the restricted range of directions the vehicles could approach. The cameras were set to capture the images with near uniformly sampled exposition times from 6 ms to 12 ms with the step of 2 ms on the road where the official speed limit is up to 90 $\mathrm{km\,h^{-1}}$. These images were cropped around the license plate and normalized to the size of 264×128 px. They were carefully manually annotated with license plate characters such that OCR accuracy could be evaluated. The approximate direction range the captured cars did approach were 37° to 57° and 59° to 79°, see the Figure 4.2.

4.1.3 *Training*

The pairs of artificial blurred image and its sharp undistorted representation (y_i, x_i) were divided into two disjoint parts. The training set which consisted of 126 k pairs and the testing set which had 14 k pairs of images. All the images were of the same size 264×128 px. The model was trained on fixed size crops with the dimension of 66×66 px, where 5 randomly sampled crops per training image created the set of 630 k input crops. Because the receptive field of the model is 50 px, the output images,

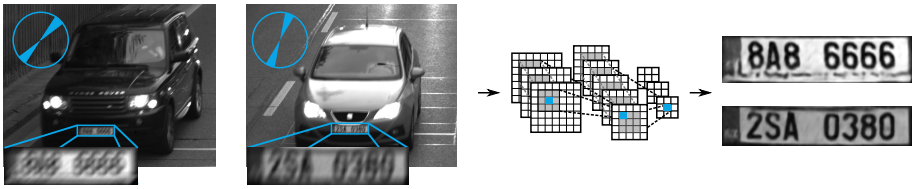


Figure 4.2: The illustration of the surveillance system images and the correspondent results of L15. The blue circle shows the approximate direction range the cars usually approach, where the left range equals to 37° to 57° and the right range to 59° to 79°. The blurred input and restored license plate images are shown.

the model produce, are only 16×16 px central patches of the input cropped images. The pair of training data is shown in Figure 4.3.

The whole network was initialized using the modified[1] *Xavier* initialization [18]

$$\mathrm{Var}\left(W_i\right) = \sqrt{\frac{3}{\text{fan-out}_i}}\,, \tag{4.2}$$

where the variance distribution of the initialized convolutional layer weights W_i is related to the number of filters, precisely on the fan-out$_i$ parameter which is defined as spatial filter size × number of filters in the layer f_i (**??**). The network was trained for 400 k iterations with a mini-batch of 54 samples. The objective was based on minimization the loss function defined as

$$\frac{1}{2N}\sum_{i=1}^{N}\left\|L_{15}\left(\mathcal{W}, y_i\right) - \hat{x}_i\right\|_2^2\,, \tag{4.3}$$

where N is the number of training pairs in the mini-batch of degraded image y_i and its ground truth sharp central patch representation x_i. The network took on average 3 days to train on a single Nvidia GeForce 980 GPU. Initial learning rate was set to 4×10^{-5} and it was reduced five times by a factor of 2. The weight update was performed based on the Stochastic Gradient Descent (SGD) with the momentum equal to 0.9 and the weight decay 5×10^{-4}. All the input data were normalized and centered around zero.

Figure 4.3: The training image pair with the illustrated blurred crop on the left and the equivalent sharp center patch on the right. In the middle is the magnified motion blur PSF.

[1] Based on the implementation in Caffe [28].

4.1.4 *Semi Non-Blind Restoration*

Two experiments were performed to assess the behavior of deblurring CNN on motion blur length and a range of blur directions. These experiments were performed on the artificially blurred images. The restored image quality was measured based on PSNR (3.7). The deblurring model is first trained on specific motion blur parameters defined as the range of the motion blur length and the range of direction.

There were 4 models trained with the fixed direction range to 20° and gradually increasing the motion blur lengths including 0–5 px, 0–9 px, 0–13 px and 0–17 px. The length was always uniformly sampled from the corresponding range. Figure 4.4a shows the results of these networks for different blur lengths. These results indicate that networks trained for shorter blur length range perform better inside these ranges. However, their results degrade rapidly outside the trained range. The restoration quality starts to degrade already at the border of the respective ranges. That is probably because no larger blurs are represented in the respective training sets. The reconstruction quality decreases linearly for longer blur kernels.

The second experiment is shown in Figure 4.4b assess the performance of the networks for different blur direction ranges. Seven models were trained, one model per different direction range, including the uniformly sampled, 10°, 20°, 40°. 60°, 90°, 130°, and 180° wide ranges of possible directions. Note that the blur kernels are symmetric and consequently the largest range of 180° covers all the possible directions. All the directions were blurred with a length uniformly sampled from 0–13 px. The observed results show similar trends as in the experiment with different blur lengths, the networks trained for tighter direction ranges perform better inside these ranges, but their performance degrades rapidly outside the respective direction ranges.

4.1.5 *Blind Restoration of Naturally Blurred Data*

Six models were trained to provide the evaluation on the naturally blurred test images captured by two surveillance cameras. These networks were all trained on blur kernels covering both cameras, i. e. the range of the blur directions was 50° wide, which shall be sufficient according to the possible directions of approaching vehicles. The networks were trained for blur lengths 0–9 px, 0–11 px, 0–15 px, 0–19 px, 0–21 px, and 0–23 px. The L0-regularized blind deconvolution method by Pan et al. [3] was selected as a representative of the traditional blind deblurring methods to serve as the baseline for a model comparison. This method is specifically optimized for images containing text and it should be suitable for the license plate images as well. An opti-

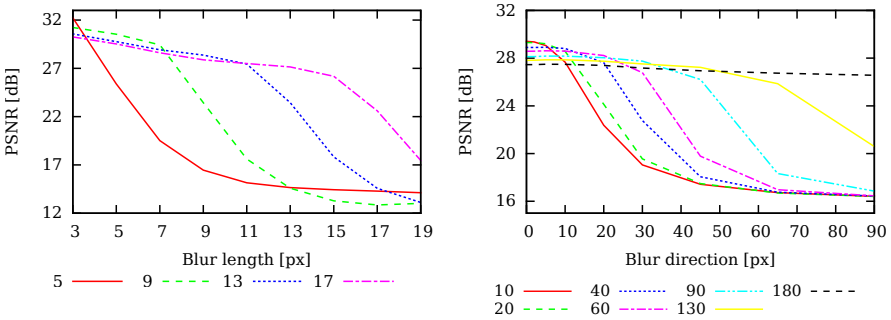


Figure 4.4: The graph on the left shows the result of specific length range trained model on several blur lengths. On the right, the presents results of models trained on specific ranges evaluated on several direction ranges.

mal parameters of L0-regularized were selected using the grid search directly on the test images.

Figure 4.5 shows results on the naturally blurred images as an accuracy of an Optical Character Recognition system. The deployed OCR system[2] is optimized for license plates and is used in commercial traffic surveillance systems. The networks trained for shorter blur perform poorly as the set contains blurs up to 19 px long. The networks trained for sufficiently long blurs significantly outperform the baseline blind deconvolution method of Pan et al. [3]. The improvement is from the character error of 23% down to 9% compared to the L0-regularized which corresponds to relative improvement by a factor more then 2. It is worth to emphasize that the OCR

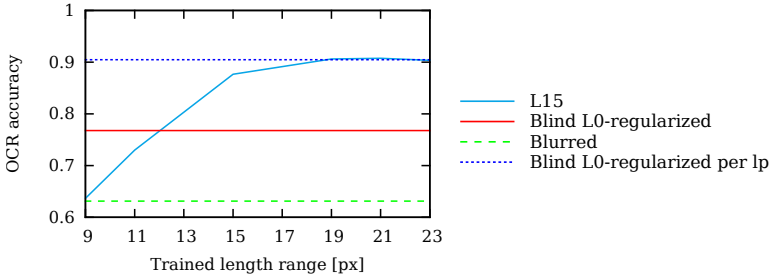


Figure 4.5: The OCR accuracy results of the originally blurred images, L0-regularized blind and non-blind deconvolved images and the L15 restorations.

[2] UnicamLPR, http://www.camea.cz/

Figure 4.6: The naturally blurred license plates sorted from left to right based on the blur amount with the corresponding deconvolved results of L0-regularized method and the restored L15 images.

accuracy keeps approximately the same for the models trained for long blurs. In a case of non blind restoration, the L0-regularized method tuned per license plate to performs similarly to the blind approach based on CNN. However, this requires the known motion blur parameters for each license plate. Figure 4.6 presents the original blurred images, reconstructed license plates by L0-regularized blind deconvolution and the L15 restorations.

4.1.6 *Summary on motion deblurring*

The evaluated L15 architecture contains of 2.3 M unique weight parameters, i. e. the model occupies approximately 9 MiB in memory. Compared to convolutional networks used in computer vision tasks, this network is still small and computationally relatively efficient. It requires 2.3 M multiply-accumulate operations per pixel. The CNN proved to be effective for the naturally blurred images even though they were trained only on images which were blurred artificially with a simple line kernel. The deblurring CNN provided superior accuracy of a consequent OCR compared to the state-of-the-art L0-regularized blind deconvolution tuned for text images [3]. These results show for the first time that CNNs provide quantitatively better deblurring quality compared to engineered state-of-the-art blind methods in a practical application.

The experiments showed that the quality of reconstructed images could be improved by customizing the CNNs for the specific range of blurs. However, the improvement is only modest in the target application, and general networks trained for a wide range of blurs still provide the high-quality results. The reconstruction quality declines linearly, in PSNR, with the increasing length of the blur kernels which makes it easy to predict possible reconstruction quality for larger blurs. Although the networks can reconstruct real images which suggest that the kernels used for training

do not have to match the shape of kernels in a real application too closely, the reconstruction quality degrades quite sharply for blurs which parameters like direction and length range are outside the trained values. The deblurring CNN are well suited for embedded applications due to their flexibility, relatively low computational power requirements, robustness, and the absence of any tunable parameters. The deblurring CNNs can be considered mature and ready to be deployed in the traffic surveillance systems.

4.2 CNN FOR JPEG ARTIFACTS REMOVAL

The end-to-end mapping network architecture is deployed for JPEG compression artifacts removal. Its utilization is principally based on the achieved results of the CNN model in motion deblurring. The artifacts are caused and clearly visible by a low compression quality. That is caused by setting the higher frequency related coefficients during the quantization step to zero. On the other hand, this loss is redeemed by achieving the high compression ratio. The way the coefficients are omitted is related to the human perception where the less sensitivity correlates with the high frequencies and vice versa.

Several metrics exist to assess the perceptual quality of images objectively. In this work, the restoration is measured based on PSNR, PSNR-B, and SSIM metrics. Generally, the most commonly used quality metric is the MSE [26] (3.6). This quantity is computed by averaging squared intensity differences of the distorted image and the reference image. That is often expressed in a logarithmic scale as the Peak Signal to Noise Ratio (PSNR) (3.7). Unfortunately, PSNR and MSE are not necessarily well correlated with the perceptual quality.

The SSIM [25] that compares local patterns of pixel intensities should better correlate with human perceptual quality. Since the attention is focused on the JPEG artifacts, the blocking artifacts, a block-sensitive metric referred to as the PSNR-B [33] is used to provide additional insights. PSNR-B modifies the original PSNR by including an additional blocking effect factor (BEF). Some experiments report IPSNR which is a PSNR increase compared to PSNR of the degraded image. IPSNR is more stable across different dataset and it directly reflects the quality improvement.

In regard to the color space $Y'C_BC_R$ which represents the luma Y', C_B blue-difference, and C_R red-difference chroma components, the most details are covered in the Y' luma channel. That is the primary reason why the main attention in this work is focused almost on the Y' luma channel only. Note, that the JPEG compression is by definition a nonlinear degradation compared to the almost only linear based motion blur.

In contrary to the deep L15 network, several small architectures are introduced including the 4, 5 and 8 layer networks L4, L5, and L8 respectively. The L4 network is a simple model similar to the AR-CNN [12] with the main distinctions in the training and related objective function. The results are compared to AR-CNN, to the widely regarded deblocking oriented SA-DCT [4, 5], and to a simple postprocessing filter SPP included in the FFmpeg framework [34]. The deepest L8 network introduces an extended skip architecture described in Section 3.2. The L5 network is used to compute, besides the pixels, directly mapping on the DCT coefficients as well. Regarding the specific architecture and different training dataset, L5 is not directly comparable with the other architectures.

4.2.1 *Architectures*

The L4 is a shallow network trained regarding direct, edge enhancement, and residual objective. The network size is comparable with the AR-CNN which is actually recognized as the state-of-the-art CNN based method. The entire L4 network receptive field is 19 px where, considering the block size of 8×8, the whole JPEG block and half is covered on each side, which provides the network with possibly sufficient spatial information. The L8, except to be a deeper model, differs mainly in the skip architecture defined as

$$\begin{aligned} f_4(x) &= h_4\Big(W_4\big(f_3(x) \parallel f_1(x)\big)\Big) \\ f_6(x) &= h_6\Big(W_6\big(f_5(x) \parallel f_1(x)\big)\Big) \\ L_8(\mathcal{W}, y) &= (f_8 \circ f_7 \circ \ldots \circ f_1)(y) \\ x &= L_8(\mathcal{W}, y), \end{aligned} \tag{4.4}$$

where the operator $\parallel$ denotes the concatenation. The layers represented by f_4 and f_6 are defined as functions which are computed on the concatenated activation maps obtained from f_1 and previous f_3 and f_5 layers. The receptive field of whole L8 network is 25 px. L4 and L8 include solely the convolutional layers followed by the nonlinear ReLU units. Both architectures are shown in Figure 4.7.

The last architecture, L5, illustrated in Figure 4.8, is slightly deeper compared to the most shallow L4 network, but in the same time much wider than any here presented network. Such a width is closely related to the data the network is fed with as it mainly is the DCT coefficients resampled from the 2D 8×8 blocks into the 1D 64 channels vectors as illustrated in Figure 3.6b. The same L5 architecture is trained for identically resampled pixels with an assumption that the block structure, which is coded directly into the input data arrangement, provides an additional information the CNN can utilize. The L5 model has several modifications related to the type of input data. In

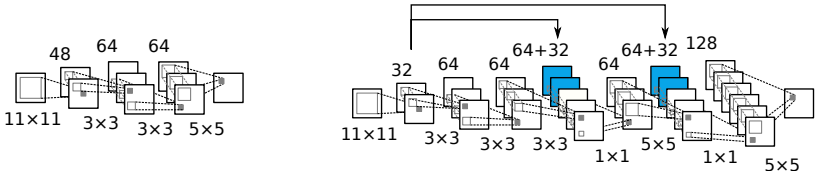


Figure 4.7: The L4 on the left is a simple and shallow FCN network. The L8 on the right deploys the skip architecture which allows transferring the layer representation deeper into the network. The activation maps of the first layer are transfered into 4th and 6th layer.

the case of pixel input data, L5 is a straight end-to-end mapping architecture, while in the case of the coefficients input data, the network is extended by a fixed IDCT layer similarly as in Figure 3.5 which allows computing the loss of pixels instead of coefficients. All the L5 architectures are trained using the residual objective. The L5 model is based on the convolutional layers followed by the trainable Parametrized Rectified Linear Unit (PReLU) [23].

4.2.2 *Data*

The majority of the experiments were computed on images from *BSDS500* [35] and *LIVE1* [10] datasets. The L4 and L8 networks were trained solely on the merged train and validation part of BSDS500 which contains 400 images. The L5 training was based on the *INRIA* holidays dataset [36] where the included images were downsampled to correspond the size of images from the other datasets and to suppress the already occurring JPEG artifacts in the original ground truth data.

The images were transformed, as was stated earlier, to the grayscale representation using the $Y'C_BC_R$ color model keeping the luma Y' component only. Only the grayscale images were considered because the attention was solely focused on the ringing and blocking artifacts while the chromatic distortions were left out. The grayscale images were compressed with the MATLAB JPEG encoder into five disjoint sets based on the JPEG quality. Specifically, the images were compressed with the quality 10, 20, 40, 50, and 60. The DCT coefficients were extracted and stored together with the related quantization tables.

The networks were evaluated on the test set from BSDS500 which includes 100 high-quality compressed images and on the LIVE1 dataset containing 29 color images of uncompressed BMP format. All the evaluation images were transformed to grayscale the same way as the training images and also compressed using the same encoder.

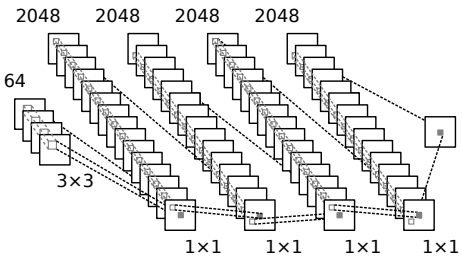


Figure 4.8: L5 architecture compared to L4 or L8 has much wider layers – number of filters to process the input of 64 channels.

It is important to use the same encoder because the quantization tables may differ between different encoder implementations.

4.2.3 *Training*

The training of presented models differs according to the objective, architecture, and data. The formerly presented L4 and L8 networks were trained the same way except for the several objective experiments which were evaluated with L4 architecture only. The L5 based models differ already in the solver itself. Namely L5 used the Adam solver instead of SGD with momentum.

The importance of the network initialization has been formerly emphasized in several publications [17, 18, 19]. In this work, the assumption of zero mean of the network initialization is recognized as helpful as it prevents mean offsets of activations to propagate through the layers. In case the mean was not zero, any mean offset in input values would result in the non-zero mean of output activations which could force the ReLU non-linearities to get fully stuck either in the positive linear interval or, even worse, in the negative interval where gradients are not propagated rendering the unit useless.

This problem is eliminated by explicitly forcing individual filters to have zero mean during initialization. Such initialization allows to use significantly higher initial learning rates, especially together with residual learning, and it results in trained networks with significantly fewer saturated neurons. The L4 and L8 based models were initialized using the Xavier approach (4.2) and shift to have the zero mean per filter.

All the filters can be forced to have zero mean during the whole training. Such constraint almost entirely eliminates any potential for unit saturation, but it prevents networks to utilize the DC component of input signals. Although reasonably good results were achieved with this constraint in the preliminary experiments, it was not

decided to use the offset suppression in the presented experiments. The L4 and L8 based models were trained using the SGD with the momentum with the minibatch of 64 64×64 px patches and 4 128×128 px patches respectively. Solver related parameters are collected in Table 4.1. The patches were randomly sampled from the training images.

In all the experiments, the loss was normalized by the number of output pixels

$$\frac{1}{N \times x_w x_h x_{ch}} \sum_{i=1}^{N} \left\| F(\mathcal{W}, y_i) - x_i \right\|_2^2 , \tag{4.5}$$

where y_w is the output patch width, y_h the height and y_{ch} number of channels. Such scaling influences the scale of gradients and results in some cases in relatively high learning rates and low weight decay parameters. The number of L4, L8, and L5 training iterations was fixed to 250 k which is significantly less compared to AR-CNN's 10^7 iterations.

The L5 based models were trained based on the residual objective and using the *ADAM* solver [37]. The specific solver parameters are given in Table 4.1. The learning rate was five times decreased by the factor of 3. The L5 models were all equally trained using 250 k iterations, where the minibatch per iteration consisted of 24 samples. The models were initialized per layer with the Gaussian distribution with zero mean and the standard deviation equal to 10^{-1} for the first layer, 10^{-2} for all the middle layers, and 0.5×10^{-2} for the last 5th layer.

4.2.4 *Artifacts Removal Quality*

The results are compared to AR-CNN [12], to the widely regarded deblocking oriented SA-DCT [4, 5], and to a simple postprocessing filter SPP included in the FFm-

Table 4.1: L4, L8 and L5 training parameters including solver type, learning rate (lr), momentum (m), and weight decay (wd).

Network	solver	lr	m	wd
L4 Direct	SGD	0.4	0.97	5×10^{-7}
L4 Residual	SGD	8	0.97	5×10^{-7}
L4 Edge enh.	SGD	0.05	0.97	5×10^{-4}
L8 Skip arch.	SGD	4	0.95	5×10^{-7}

Network	solver	lr	β_1	β_2	ϵ	wd
L5	ADAM	5×10^{-4}	0.9	0.999	10^{-8}	0

Table 4.2: Image restoration quality on LIVE1 test dataset for JPEG quality 10 and 20.

	Q10			Q20		
method	PSNR	PSNR-B	SSIM	PSNR	PSNR-B	SSIM
distorted	27.77	25.33	0.791	30.07	27.57	0.868
spp	28.37	27.77	0.806	30.49	29.22	0.877
SA-DCT	28.65	28.01	0.809	30.81	29.82	0.878
AR-CNN	28.98	28.70	0.822	31.29	30.76	0.887
L4 Residual	**29.08**	**28.71**	**0.824**	31.42	30.83	0.890
L5 Pixel	–	–	–	31.42	30.63	0.890
L8 Residual	–	–	–	**31.51**	**30.92**	**0.891**

Table 4.3: Image restoration quality on BSDS500 test dataset for JPEG quality 10 and 20.

	Q10			Q20		
method	PSNR	PSNR-B	SSIM	PSNR	PSNR-B	SSIM
distorted	27.58	24.97	0.769	29.72	26.97	0.852
spp	28.13	27.49	0.782	30.11	28.68	0.859
AR-CNN	28.74	**28.38**	0.796	30.80	30.08	0.868
L4 Residual	**28.75**	28.29	**0.800**	30.90	30.13	0.871
L5 Pixel	–	–	–	30.94	29.91	**0.873**
L8 Residual	–	–	–	**30.99**	**30.19**	0.872

peg framework [34]. While L4 architecture was used in most experiments and it was trained for various compression quality levels, L8 was trained only for JPEG quality 20. If not stated otherwise, the residual version of networks was used. The results of L5 are included with the note that it was trained on the INRIA Holiday dataset instead of BSDS500 used for L4 and L8.

The evaluation of removing the artifacts on LIVE1 dataset with JPEG quality 10 and 20 is presented in Table 4.2. The results achieved on BSDS500 test dataset are written in Table 4.3. L8 model outperforms all the other methods with significantly higher scores in all three quality metrics with the exception on BSDS500 test dataset, where the L5, trained completely on different data, achieved a higher SSIM result. Although L4 model performs worse compared to L8, it still surpasses the other methods in most cases even though it is much smaller and computationally efficient compared to both L8 and L5. Interestingly, the L5 performance is between the L4 and L8 having good results based on the SSIM meanwhile surprisingly worse on the B-PSNR. Examples of resulting images are presented in Figure 4.13. There are still visible blocking artifacts of L4 and L8 models trained with residual objective while the L5 model with the

worse results based on the PSNR metric seems to restore such a type of artifact very well. That is seen on the monotonic parts of the image like for example the sky.

JPEG QUALITY GENERALIZATION The attention was focused on the generalization ability of the trained networks regard to a different compression quality. The ability of CNNs to handle various compression qualities is assessed by the experiment which consisted of training the single L4 model for one particular quality and consequently evaluating such a model on all the other qualities. The results in Figure 4.9 show that L4 trained on a range of qualities, from Q10 up to Q60, provides stable results across the equal quality range. However, the quality-specific networks perform better for their respective qualities which yield to a possibility to train the high specialized models in case of the quality of degraded images is known. On the other hand, the quality-specific networks generalize only to similar qualities. In practice, a single network should easily be able to handle smaller quality ranges, e. g. from 10 up to 20 quality points wide, when trained on data from such a range.

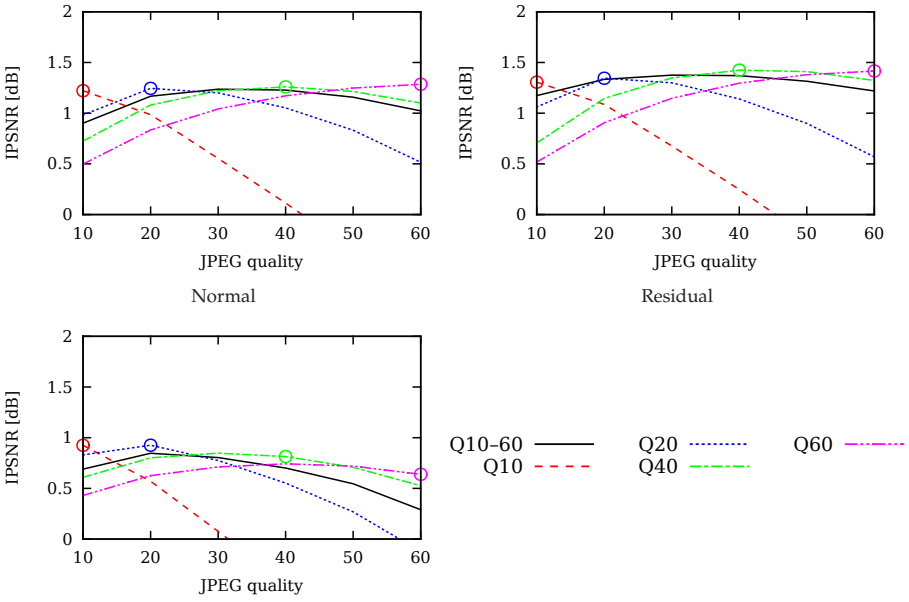


Figure 4.9: Generalization ability of L4 networks trained with normal, residual, and edge preserving objectives for different JPEG quality levels.

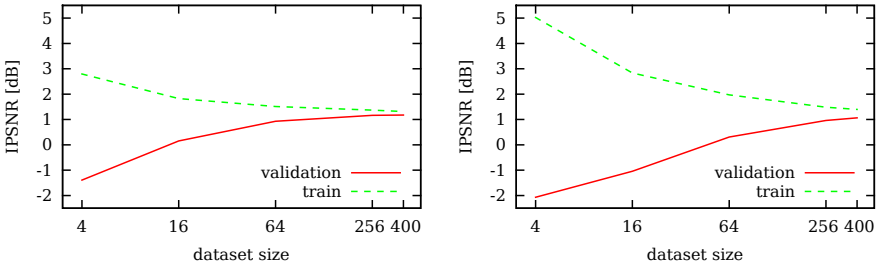


Figure 4.10: Generalization for different sized train set.

DATASET SIZE The quality of restoration achieved by larger networks may suffer due to inadequate size of a training set. In order to assess how the L4 and L8 models behave with respect to training dataset size, the residual versions of the networks were trained on 4, 16, 64, 256, and 400 images from BSDS500 training set. The L4 and L8 models contain approx 70 k and 220 k learnable parameters respectively which suggest that L8 model should require a larger training set for the same generalization. Figure 4.10 shows results of models trained and evaluated on differently sized training datasets together with the evaluation on the corresponding independent test dataset. Both networks clearly overfit on the smaller datasets. L8 model overfits significantly more, and it would require more images to reach proper generalization, while L4 seems to reach its maximal generalization already on the relatively small dataset of 400 images.

4.2.5 *Impact of The Objective*

All the L4 models were trained for direct mapping, residual, and edge enhancement objectives to evaluate the contribution of each. Although the architecture and initialization of all the L4 networks were the same, the suitable learning rates (lr) and weight decay coefficients (wd) had to be selected based on the parameters grid search for each learning objective separately. The solver parameters are noticed in Table 4.1. All the parameters were selected regarding JPEG quality 10 and they were used for all the other qualities as well.

The learning progress is shown in Figure 4.11. The residual network converges much faster compared to the both direct and edge enhancement objectives. The results on LIVE1 based on PSNR, PSNR-B and SSIM metrics are presented in Table 4.4.

The results show that the residual based model converges faster and achieves the best restoration quality compared to other objectives. The edge enhancement objective converges a slightly faster in the beginning but stops to develop quite soon letting the direct mapping to overcome its results. It could be expected that the direct objective-based training may achieve a similar restoration quality compared with the residual objective-based training with a clear disadvantage in the form of time needed to converge.

The progress of training the filters of the first layer during training in different objective based models is shown in Figure 4.12. All the networks formed reasonable-looking filters. The residual objective trained model formed more complex higher frequency filters compared to the other networks. The edge preserving network learned some low-pass filters which are probably needed to transfer the general image appearance through the network. These filters are missing in the residual objective trained model. The filters of the direct objective trained model remain noisy, which could be due to different weight decay coefficient the low learning rate, or their combination. It also implies that the direct mapping would get slightly better results if trained for more iterations which are indicated regarding the IPSNR shown in Figure 4.11.

The results indicate that the residual learning is beneficial for JPEG artifact removal regarding restoration quality and training speed. On the other hand, the edge preserving objective does not improve the quality as is shown in the case of L4.

DCT coefficient based restoration was computed using the L5 architecture with an atypical layers width providing much more filters per layer compared to the L4 or L8 models. The L5 can not be compared directly to both L4 and L8 pixel based models because L5 models were trained on the different training set, the INRIA Holiday [36]. The input data were normalized by a single fixed value to be approximately in the interval from -1 up to 1. The L5 models operating with quantized DCT coefficients – $\mathcal{B}$, JPEG dequantized coefficients – $Q\mathcal{B}$, and directly with pixels were evaluated with the results presented in Table 4.5.

Table 4.4: Results of L4 networks with different objectives on LIVE1 dataset with quality 10.

Objective	PSNR	PSNR-B	SSIM
Distorted	27.58	24.97	0.769
Direct mapping	28.99	28.66	0.820
Edge preserving	28.69	28.40	0.813
Residual learn.	**29.08**	**28.71**	**0.824**

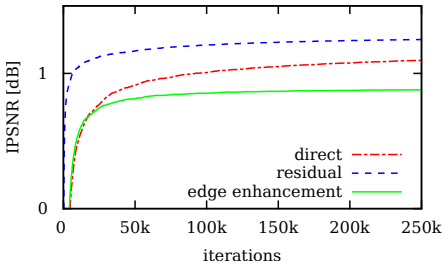


Figure 4.11: Development of L4 with different training objectives.

DCT COEFFICIENTS The training dataset was later on augmented by shifting the images by a uniformly sampled shift size in the range from 0 up to 7 pixels in both directions. These shifted original images were then encoded into the JPEG. 64× more data became available leaving the blocking artifact in the same position within the image. The L5 pixel–pixel mapping model trained on the augmented dataset achieved 16% higher IPSNR compared with the same model trained on the original smaller amount if training data but with the same amount of training iterations. Despite the lower achieved PSNR compared to the L8 network, the result images are blocking free while both L4 and L8 models, unfortunately, preserve surpassed still visible blocking artifacts.

The different type data based L5 models, coefficients, dequantized coefficients and pixels, show very similar results, where the differences are apparently related to the model initialization. The exception is the case in which the JPEG DCT coefficients $\mathcal{B}$ are multiplied by the quantization table Q. The results of L5 model operating with such data are slightly better compared to other L5 models. It is apparent that the DCT based restoration models can be successfully deployed without the requirement of any post-processing of the decoded JPEG image. That allows keeping the existing decoders and just use the networks in a preprocessing step being similar to JPEG Quality Transcoder (JQT) approach. The blocking artifacts are well removed by models operating with the resampled input pixels from the 8×8 blocks into the 64 channel vectors. Regarding the results, it is highly probable that such resampled input data explicitly helps the model to train focus on the blocking fixed size and periodicity.

4.2.6 *Summary on Artifacts Restoration*

The CNN based models, namely L4, L5, and L8 were presented. All three outperformed state of the art with most significant results achieved by the L8 model based

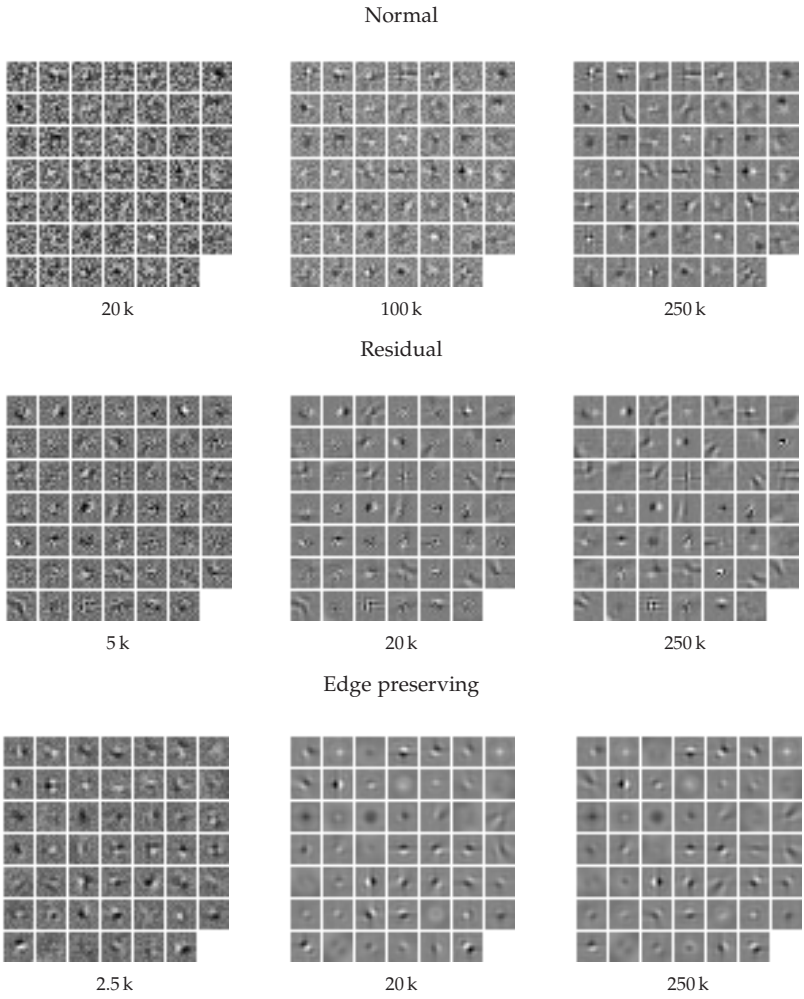


Figure 4.12: Filters from the first layer of L4 networks with normal/residual/edge preserving objective at different stages of training. Iterations are showed below the images.

on the residual training and skip architecture. The residual objective proved to be appropriate for JPEG artifacts restoration and allowed to train the model faster regard to the number of iterations and achieved results. However, the edge enhancement objective did not show any benefits compared to the direct mapping which would provide any reason to prioritize such a learning objective. The importance of the dataset size in regard to the model capacity showed both experiments, the observed L4 and L8 mod-

Table 4.5: The different input data and loss function based L5 architecture results. The structure of the model name describes the settings, i. e. the input data and the loss-computed-data. The $\mathcal{B}$ is the JPEG quantized DCT coefficient, $Q\mathcal{B}$ is the $\mathcal{B}$ multiplied by the quantization table, pix stands for pixel data. L5 $\mathcal{B}$–pix represents the L5 model with the JPEG DCT quantized coefficient input data and the loss computed on pixels.

	LIVE1			BSDS500		
method	PSNR	PSNR-B	SSIM	PSNR	PSNR-B	SSIM
distorted	30.07	27.57	0.868	29.72	26.97	0.852
L5 $\mathcal{B}$–$\mathcal{B}$	31.25	30.51	**0.890**	30.81	29.82	0.871
L5 $Q\mathcal{B}$–$Q\mathcal{B}$	**31.31**	**30.52**	**0.890**	**30.85**	**29.84**	**0.872**
L5 $\mathcal{B}$–pix	31.25	30.51	**0.890**	30.81	29.82	0.871
L5 pix–pix	31.23	30.49	0.889	30.78	29.81	0.871
L5 pix–pix C-PSNR	31.44	30.63	0.892	30.94	29.92	0.873
L5 pix–pix aug	31.42	30.63	0.892	30.94	29.90	0.873

els trained on several dataset sizes and the L5 model trained on the 64× augmented training dataset which provides more than 16% of IPSNR increase compared to the same model on the original training dataset size. The CNN based models ability to generalize was investigated with the results showing the single model covering a wide range of compression qualities with restoration level. However, the specialized model for specific quality can deliver slightly better results measured by the PSNR metric. The experiments provided support for the JQT which transforms the low-quality JPEG coefficients to the coefficients representing higher quality restored image. The input image blocks resampled from 8×8 spatial size into the 64 channel vectors provided the L5 architecture with the subsidiary information feasible to compute high quality blocking artifacts restoration.

The pixel based architectures, L4 and L8, are with their 70 k and 220 k weight parameters significantly smaller compared to the motion deblurring L15 model with 2.3 M weights. Using cuDNN[3] v3 implementation of convolutions on GeForce GTX 780, the 1 Mpx image takes approximately 220 ms with network L4 and roughly 1052 ms with L8 to be restored. The L4 and L8 networks require approximately 140 k and 440 k floating point operations per pixel.

[3] Nvidia GPU-accelerated library of primitives for deep neural networks.

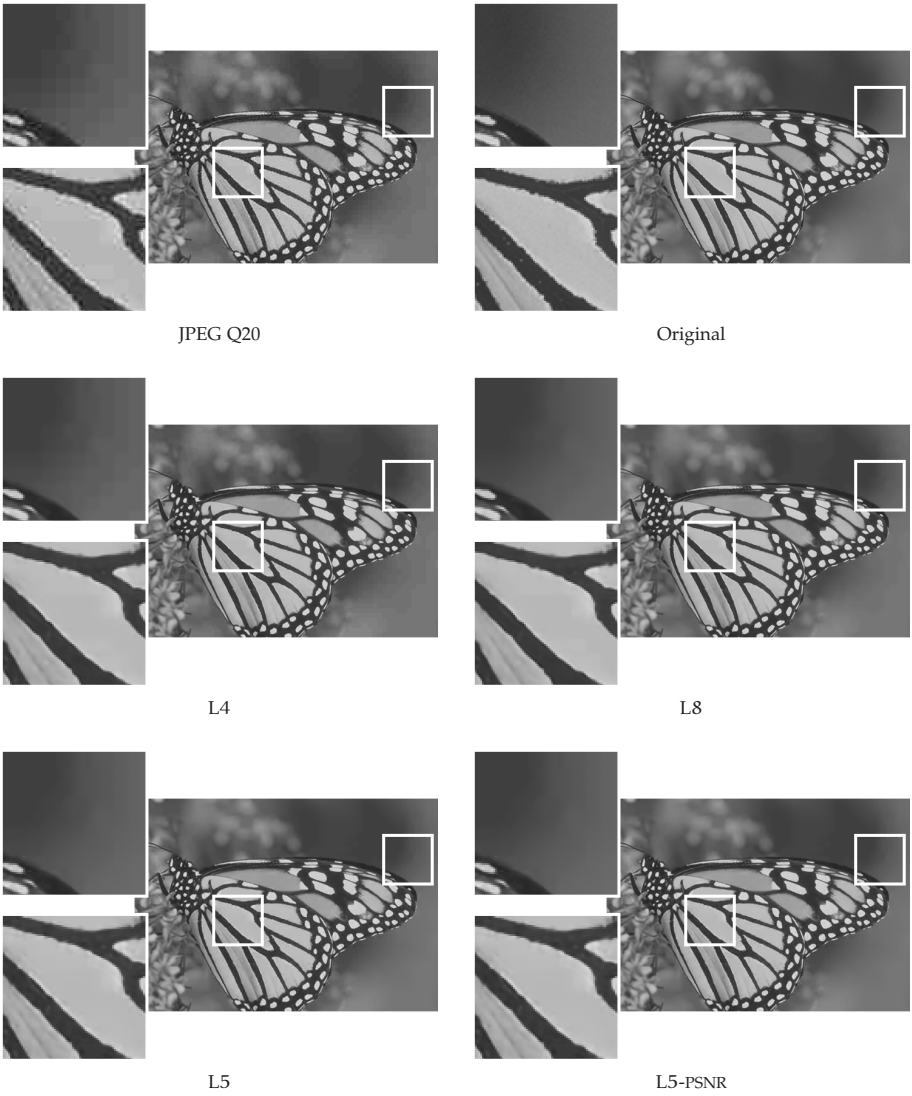


Figure 4.13: Visual comparison of restored monarch image from LIVE1 [10] dataset originally compressed with JPEG quality 20. L4 restores the ringing artifacts but the blocking is to a certain extent preserved. L8-skip compared to L4 provides obviously better results restoring the blocking artifacts. Note, that both L5 smooth the blocking artifacts but performs slightly worse on the ringing artifacts compared to L8 and L4 as well.

4.3 SUMMARY OF CONTRIBUTIONS

The core of this thesis is framed by a unified method of image restoration based on Convolutional Neural Network. The data-driven approach has been deployed for particular tasks of image restoration. It is the deconvolution, namely the motion deblurring, which is well described and where the hard part is to estimate the unknown blur parameters such as the length and direction. Further, it is the compression artifacts removal task, which instead of deconvolution restores an image by suppressing the artificial boxing and ghost edges of ringing artifacts. In both cases, i. e. license plate motion deblurring and JPEG artifacts removal, the presented CNN provides beyond state-of-the-art results. Compared to the engineered methods, which significantly differ from each other according to the task they focus on, the CNN approach allows to quickly train a specialized or a universal model solely dependent on the training data. The case of a specific model is related to a limited range of parameters the degradation can be modeled with, e. g. the limited range of lengths the motion blur can consist of. In contrary, the universal model can be used for various levels of particular degradation. That is the case of the single model used for an arbitrary range of motion blur lengths and directions. *Considering the results presented in this thesis, the hypothesis is fulfilled.*

This work extends the approach of text deblurring based on CNN introduced by Hradis et al. [11], which is based on the 15 layer CNN model trained purely on artificially blurred data. This model performs well for various ranges of motions blur lengths and directions. The results on artificially blurred data show model ability to recover an arbitrary range of blur parameters. Simultaneously, the end-to-end model easily outperforms the blind deconvolution L0-regularized method and competes very well compared to the non-blind variation of the same text image specialized L0-regularized method. Further, the L15 model can restore the naturally blurred images as well. Based on the OCR accuracy, L15 CNN model delivers significantly better results compared to L0-regularized method which is considered to be state of the art. The motion deblurring based on CNN reveals how simple it is to obtain a model with the beyond state-of-the-arts outcomes. Model, which generalizes very well and which can handle a wide range of possible blur parameters.

The CNN approach for image compression artifacts restoration presented in this work significantly improves the-state-of-the-art results. Similarly to the L15 model for license plate motion deblurring, the introduced models besides the beyond state-of-the-art results provide a significant generalization ability over various JPEG compression qualities. The analysis of the architectures and the objectives the networks are trained for is given. The residual objective used for artifacts restoration is presented allowing to speed up the training process together with better outcomes compared

to the direct objective. The experiments pointed out the contribution of input data reorganization referring to the deblocking. The work shows that the CNN model used for image restoration in the pixel domain is suitable for transforming the highly compressed JPEG coefficients to the coefficients representing the image, which decoded, becomes artifacts free. The JPEG compression artifacts removal supports the idea of a unified approach to image restoration. There are many others tasks of image restorations. Nevertheless, even these are not reviewed in this thesis, here presented results indicates a possible performance increase in the sense of accuracy and quality based on the data-driven CNN models.

4.4 FUTURE WORK

The combination of all described approaches including the skip architecture, residual objective with further relatively smaller kernel stacking, e. g. like the inception network [38, 39], may provide the results yet far beyond state of the art. Unfortunately, the amount of computational time is directly proportional to the model complexity. Therefore, the recently used architectures take several days to train which makes the exhaustive architecture state space search quite difficult.

The image restoration CNN based models were and yet significantly are influenced by the computer vision research. Based on results in computer vision, the next steps shall lead to architectures of stacked filters comprising the model build from relatively small kernels interleaved with a higher amount of non-linearities, like ReLU, PReLU as used in the L5 models, or recently introduced Exponential Linear Unit (ELU). Further, classification instead of regression may provide the network with a much easier problem to learn, i. e. the output would be one of 256 possible values representing the image intensity. In such a case an ensemble of models in a form as presented in [40] or just utilizing the dropout in a network can be simply used to achieve better results.

In the case of JPEG artifacts restoration, the transposed convolution can offer interesting outcomes. It is worth considering networks utilizing the transposed convolution – sometimes noted as deconvolution which spatially scatters the data. That includes various scenarios like deconvolution, in the beginning, gradually stacked deconvolution, and deconvolution at the end of the network. The deconvolution, precisely transposed convolution, is understood as the reverse convolution where the single input value, the result of a convolution, is partially distributed to its source values **??**. Here, the possible future research regarding JPEG DCT coefficients is likely to provide interesting results.

Although the PSNR based objective did not directly show any significant benefit over the simple MSE loss function, the SSIM loss function is worth a try. The related

idea of inpainting the corrupted image to obtain the perceptually plausible image could be used in situations where the scene fidelity is not necessary. Apart from the restoration tasks, the CNN can be deployed in other image processing challenges including in robotics often used visual based parameters estimation. These may include the image matching for the loop closer detection extending the work [41] in the mapping and environment reconstruction applications, the rotation-translate estimation between several consequent images, the scene segmentation, the depth from an image estimation, several sensors fusion [42], or descriptors learning [43].

Plethora of degradations and corruptions types exist, where the CNN utilization may improve the restoration results compared to the engineered methods, e. g. the whole family of deconvolution methods. In this thesis, the deconvolution CNN is utilized for license plate images deblurring, which is a very narrow image domain compared to the natural images. The end-to-end mapping for such tasks may be much too hard for recent network models. Nevertheless, no such known research has been yet done in this field. A regression CNN models introduced to compute the image restoration are likely to be suitable for similar tasks related to inpainting. An inpainting model can be used to estimate the shape and a texture of partially occluded objects in an image or generate details which may provide better perceptual image quality. An interesting approach to image generation is based on adversarial networks, where the generator network tries to fool the discriminator network with generated images instead of real images. Last but not least, the inpainting may by used for several objects anonymization including the human faces, license plates, advertisements and generally anything in the image.

5 CONCLUSION

This work focuses on an image restoration based on models of convolutional neural networks. Particularly, two different tasks were chosen, motion deblurring of license plate images taken by a surveillance system and artifacts removal caused by low quality of JPEG compression. Usually, the methods of image restoration are hand-engineered. That yields to a variety of approaches which are comprised of certain processing pipelines related to a type of degradation. Specifically, in motion deblurring, the pipeline consists of PSF estimation and a subsequent deconvolution to restore the latent sharp image. Compression artifacts restoration methods try to smooth the discontinuities made by blocking or suppress ringing on edges.

In this work, in contrary, a direct end-to-end mapping based on convolutional neural networks is presented. Restoration relies on a data-driven trained model which directly transforms a degraded image to an undistorted image. Recently introduced convolutional neural network architecture, i. e. AlexNet, inspired a model deployed for license plate motion deblurring. Several experiments show that a single model is sufficient for various motion blurs differing in lengths and directions which allow the comparison with blind deconvolution methods. The model trained solely on artificially blurred data outperforms the considered state-of-the-art method deployed on naturally blurred images where the achieved OCR based error accuracy is 9 % compared to 23 % error accuracy of L0-regularized method.

Further, a nonlinear degradation based on the JPEG compression is restored exploiting the same end-to-end approach of data-driven trained models. Compared to motion deblurring, restoration of compression artifacts is a harder problem due to the missing image information. While the approach is the same, various training and architecture related extensions are introduced including the residual objective, skip architecture, and loss computed on an image data in a case of JPEG coefficient transformation. These extensions contribute to train model which achieved in artifacts suppression state-of-the-art results. Particularly, L8 model achieved 31.51 PSNR compared to 31.29 PSNR of recently introduced AR-CNN and 30.81 PSNR of hand-engineered SA-DCT. In the case of JPEG artifacts restoration, direct transformation of JPEG coefficients based on the convolutional network is proposed. Such transformed coefficients allow restoring the artifacts degraded image before decoding itself.

The results achieved in both tasks contribute to the idea of utilizing CNNs as a unified approach to image restoration. It is worth to try to follow the ongoing research in

computer vision, where the majority of CNN related trends come from. That includes stacking the spatially small filters interleaved by more nonlinearities providing even better models. An interesting yet challenging deblurring of natural images should be investigated further. In the case of JPEG coefficients transformation, the fixed IDCT layer can be substituted by the trained transposed convolution allowing the network to adapt the decoding step. Considering the fact that all the presented models are regression based CNN, they can be therefore deployed for a task of inpainting as well. That would allow restoring incomplete data in an image, i. e. occluded objects or simply too much-degraded image regions. The impact of CNN models in various research domains is high. There is a lot of other applications the deployment of data-driven models is worth to try.

This thesis begins with an introduction reminding a year the research on NN is considered to begin. After more than 70 years later the actual state of the art of CNN dynamically evolves providing a significant impact in various domains including the image restoration as well.

BIBLIOGRAPHY

[1] M. Sonka, V. Hlavac, and R. Boyle. *Image Processing, Analysis, and Machine Vision*. Thompson Learning, 3 edition, 2008. ISBN 9780495082521.

[2] S. G. Mallat. *A wavelet tour of signal processing: the sparse way*. Elsevier Inc., 3 edition, 2008. ISBN 9780080922027.

[3] J. Pan, Z. Hu, Z. Su, and M. H. Yang. Deblurring text images via l0-regularized intensity and gradient prior. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2901–2908, June 2014. doi: 10.1109/CVPR.2014.371.

[4] A. Foi, V. Katkovnik, and K. Egiazarian. Pointwise Shape-Adaptive DCT for high-quality deblocking of compressed color images. In *Proc. of the European Signal Processing Conf.*, September 2006.

[5] A. Foi, V. Katkovnik, and K. Egiazarian. Pointwise Shape-Adaptive DCT for high-quality denoising and deblocking of grayscale and color images. *IEEE Trans. Image Processing*, 16 (5):1395–1411, May 2007.

[6] M. R. Banham and A. K. Katsaggelos. Digital image restoration. *IEEE Signal Process. Mag.*, 14(2):24–41, March 1997. ISSN 1053-5888. doi: 10.1109/79.581363.

[7] ITU. Recommendation T.81, 1993.

[8] E. Hamilton. JPEG File Interchange Format, September 1992.

[9] Technical Standardization Committee on AV & IT Storage Systems and Equipment. Exchangeable image file format for digital still cameras: Exif version 2.2, April 2002.

[10] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik. LIVE image quality assessment database release 2, 2015.

[11] M. Hradis, J. Kotera, P. Zemcik, and F. Sroubek. Convolutional neural networks for direct text deblurring. In *British Machine Vision Conf. (BMVC)*. The British Machine Vision Association and Society for Pattern Recognition, 2015. ISBN 1-901725-53-7.

[12] Ch. Dong, Y. Deng, Ch. Loy Change, and X. Tang. Compression artifacts reduction by a deep convolutional network. In *Intl. Conf. on Computer Vision (ICCV)*, pages 576–584, December 2015.

[13] Ch. Dong, Ch. Ch. Loy, K. He, and X. Tang. Learning a deep convolutional network for image super-resolution. In *Eur. Conf. on Computer Vision (ECCV)*, Lecture Notes in Computer Sci., pages 184–199. Springer International Publishing, September 2014. ISBN 978-3-319-10593-2. doi: 10.1007/978-3-319-10593-2_13.

[14] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015. ISSN 0028-0836.

[15] S. Mallat. Understanding deep convolutional networks. *Philosoph. Trans. of the Royal Soc. of London A: Math., Phys. and Eng. Sci.*, 374(2065), 2016. ISSN 1364-503X. doi: 10.1098/rsta.2015.0203.

[16] Ch. Dong, Ch. Ch. Loy, and K. He. Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Machine Intell.*, 38(2):1–14, 2015. ISSN 0162-8828. doi: 10.1109/TPAMI.2015.2439281.

[17] Y. Bengio. Practical recommendations for gradient-based training of deep architectures. In *Neural Networks : Tricks of the Trade, Lecture Notes in Computer Sci.*, pages 437–478, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-35289-8. doi: 10.1007/978-3-642-35289-8_26.

[18] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. *Intl. Conf. on Artificial Intelligence and Statistics (AISTATS)*, 9:249–256, 2010. ISSN 15324435. doi: 10.1.1.207.2059.

[19] Y. LeCun, L. Bottou, G. B. Orr, and K. R. Muller. Efficient backprop. In *Neural Networks : Tricks of the Trade, Lecture Notes in Computer Sci.*, volume 1524, pages 9–50, London, UK, UK, 1998. Springer-Verlag. ISBN 9783540494300. doi: 10.1007/3-540-49430-8.

[20] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[21] J. Kim, J. K. Lee, and K. M. Lee. Accurate image super-resolution using very deep convolutional networks. *CoRR*, abs/1511.04587, 2015.

[22] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *XoRR*, 7(3):171–180, December 2015. ISSN 1664-1078. doi: 10.3389/fpsyg.2013.00124.

[23] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Intl. Conf. on Computer Vision (ICCV)*, pages 1–11, 2015. ISBN 978-1-4673-8391-2. doi: 10.1109/ICCV.2015.123.

[24] I. Sobel. History and definition of the so-called Sobel operator, 2014.

[25] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Processing*, 13(4):600–612, 2004. ISSN 1057–7149. doi: 10.1109/TIP.2003.819861.

[26] Z. Wang and A. C. Bovik. Mean squared error: Love it or leave it? A new look at Signal Fidelity Measures. *IEEE Signal Process. Mag.*, 26:98–117, January 2009. ISSN 1053-5888. doi: 10.1109/MSP.2008.930649.

[27] J. Lazzaro and J. Wawrzynek. Jpeg quality transcoding using neural networks trained with a perceptual error measure. *Neural Computation*, 11(1):267–296, January 1999. ISSN 0899-7667. doi: 10.1162/089976699300016917.

[28] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *XoRR*, 2014.

[29] P. Svoboda, M. Hradis, L. Marsik, and P. Zemcik. CNN for license plate motion deblurring. In *Intl. Conf. on Image Processing (ICIP)*, pages 1–4. IEEE Signal Processing Society, 2016.

[30] P. Svoboda, M. Hradis, D. Barina, and P. Zemcik. Compression artifacts removal using convolutional neural networks. *J. of WSCG*, 24(2):63–72, 2016. ISSN 1213-6972.

[31] Ch. J. Schuler, Ch. H. Burger, S. Harmeling, and B. Scholkopf. A machine learning approach for non-blind image deconvolution. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2013.

[32] L. Xu, J. SJ. Ren, C. Liu, and J. Jia. Deep convolutional neural network for image deconvolution. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 27, pages 1790–1798. Curran Associates, Inc., 2014.

[33] Ch. Yim and A. C. Bovik. Quality assessment of deblocked images. *IEEE Trans. Image Processing*, 20(1):88–98, January 2011. ISSN 1057-7149. doi: 10.1109/TIP.2010.2061859.

[34] A. Nosratinia. Embedded post-processing for enhancement of compressed images. In *Proc. Data Compression Conference (DCC)*, pages 62–71, March 1999. doi: 10.1109/DCC.1999.755655.

[35] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Intl. Conf. on Computer Vision (ICCV)*, volume 2, pages 416–423, July 2001.

[36] H. Jegou, M. Douze, and C. Schmid. Hamming Embedding and Weak Geometric Consistency for Large Scale Image Search. In *Eur. Conf. on Computer Vision (ECCV)*, pages 304–317. Springer Berlin Heidelberg, Berlin, Heidelberg. doi: 10.1007/978-3-540-88682-2_24.

[37] D. Kingma and J. B. Adam: A Method for Stochastic Optimization. *CoRR*, abs/1412.6980: 1–15, 2014.

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

[39] Ch. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. *XoRR*, 2015. ISSN 08866236. doi: 10.1002/2014GB005021.

[40] M. Hradis, M. Kolar, P. Svoboda, and P. Smrz. Large scale image classification by Brno University of Technology. Poster at ILSVRC 2014 in conjunction with ECCV 2014, September 2014.

[41] V. Ila, L. Polok, M. Solony, and P. Svoboda. Slam++. A highly efficient and temporally scalable incremental SLAM framework. *Intl. J. of Robotics Research*, 2016(123):1–22, 2016. ISSN 1741-3176.

[42] P. Svoboda and P. Zemcik. Applications of LIDAR and camera fusion. In *Proc. of the DT workshop*, pages 105–106, 2010. ISBN 978-80-554-0304-5.

[43] P. Schaffroth and P. Svoboda. Fast corner point detection through machine learning. In *Proc. of the 17th Conf. STUDENT EEICT 2011*, Volume 3, pages 537–541, 2011. ISBN 978-80-214-4273-3.