# Generic 3D Convolutional Fusion for Image Restoration

Jiqing Wu(✉), Radu Timofte, and Luc Van Gool

Computer Vision Laboratory, D-ITET, ETH Zurich, Zürich, Switzerland
{jwu,radu.timofte,vangool}@vision.ee.ethz.ch

**Abstract.** Also recently, exciting strides forward have been made in the area of image restoration, particularly for image denoising and single image super-resolution. Deep learning techniques contributed to this significantly. The top methods differ in their formulations and assumptions, so even if their average performance may be similar, some work better on certain image types and image regions than others. This complementarity motivated us to propose a novel 3D convolutional fusion (3DCF) method. Unlike other methods adapted to different tasks, our method uses the exact same convolutional network architecture to address both image denoising and single image super-resolution. Our 3DCF method achieves substantial improvements (0.1 dB–0.4 dB PSNR) over the state-of-the-art methods that it fuses on standard benchmarks for both tasks. At the same time, the method still is computationally efficient.

## 1 Introduction

Image restoration is concerned with the reconstruction/estimation of the uncorrupted image from a corrupted or incomplete one. Typical corruptions include noise, blur, down-sampling, hardware constraints (*e.g.* Bayer pattern) and combinations of those. After decades of research there is a large literature [1] dedicated to restoration tasks, whereas the literature studying the fusion of restoration results is thin [2]. In this paper we tackle such fusion as a means for further performance improvements. Particularly, we propose a 3D convolutional fusion (3DCF) method and validate it on image denoising and single image super-resolution.

### 1.1 Image Denoising (DN)

Natural image denoising aims at recovering the clean image given a noisy observation. The most often studied case is when the image corruption is caused by additive white Gaussian (AWG) noise with known variance. Also, the images are assumed to be natural, capturing every-day scenes, and the quantitative measure for assessing the recovery result is the peak signal-to-noise ratio (PSNR), which stands in monotonic relation to the mean squared error (MSE).


C.-S. Chen et al. (Eds.): ACCV 2016 Workshops, Part I, LNCS 10116, pp. 159–176, 2017.
DOI: 10.1007/978-3-319-54407-6_11

The most successful denoising methods employ at least one of the following

**Denoising principles** as listed in [3]: Bayesian modeling (coupled with Gaussian models for noiseless patches), transform thresholding (assumes sparsity of patches in a fixed basis), sparse coding (sparsity over a learned dictionary), pixel or block averaging (exploits image self-similarity).

Most denoising methods work at a single image scale, the finest one, and often a small image patch is the basic processing unit. The patch captures local image information for a central pixel and a statistical amount of uncorrupted pixels. Zontak *et al.* [4] recently opened up a fresh research direction by proposing a method based on patch recurrence across scales (PRAS). Another partition of the methods is based on whether only the noisy image is used, or also learned priors and/or extra data from other (clean) natural images. This leads to *internal* and *external* methods. Some well known examples of each are:

**Internal denoising methods:**
**NLM** (non-local means) [5] reconstructs a noisy patch with a weighted average of similar patches from the same image. It uses the image self-similarity and the fact that the noise is usually uncorrelated.
**BM3D** (block matching 3D) [6] extends NLM and the DCT denoising method [7]. BM3D groups similar patches into a 3D block, applies 3D linear transform thresholding, and inverses the transform.
**WNNM** (weighted nuclear norm minimization) [8] follows the self-similarity principle, and applies WNNM to recover the noiseless patch from a matrix of stacked non-local similar patch vectors.
**PRAS** (patch recurrence across scales) [4] creates (an)isotropic image scale pyramids and extracts the estimated (noiseless) patch from the same corresponding position but at a different scale.
**PLE** (piecewise linear estimation) [9] is a Bayesian restoration model, including denoising, deblurring, and inpainting. PLE employs a set of 19 Gaussian models obtained from synthetic edge images (as priors) and an estimation-maximization iterative procedure.

**External denoising methods:**
**EPLL** (expected patch log likelihood) [10] can be seen as a shotgun extended version of PLE. It learns a Gaussian mixture model with 200 components for 2 million clean patches sampled from external natural images, and tries to maximize the expected log likelihood of any randomly chosen patch in the image.
**LSSC** (learned simultaneous sparse coding) [11] adapts a sparse dictionary learned over an external database by adding a grouping step to the noise image.
**MLP** (multi-layer percepton) [12] learns from an external database with clean and noisy images, and was among the first to introduce neural networks to low level image restoration tasks.
**CSF** (cascade of shrinkage fields) [13] proposes shrinkage fields, combining the image model and the optimization algorithm as a whole. The time complexity is greatly reduced by inherent parallelism.

**opt-MRF** (Loss-Specific Training of Filter-Based MRFs) [14] revisits loss-specific training and uses bi-level optimization to solve the image restoration problem.
**TRD** (trained reaction diffusion) [15] extends the solving process of nonlinear reaction diffusion to a deep recurrent neural network, outperforms many of the aforementioned methods, while offering the lowest time complexity for now.

It is quite surprising that most of the recent top denoising methods (such as BM3D, LSSC, EPLL, PRAS, and even WNNM) face a plateau. They perform equally well for a large range of noise, despite that they are quite different in their formulations, assumptions, and information used. This is the reason behind the recent work that fuses them, pushing the limits by combining different approaches [16,17]. We refer the readers to [2] for a study of image fusion algorithms of the past decades. Others investigated the theoretical limits for denoising with natural image patch priors [18], and at least for the lower noise levels, the gap between the most successful methods and the predicted limits seems to rapidly diminish.

**Fusion methods:**
**PatchSNR** (patch signal-to-noise ratio). Mosseri *et al.* [19] propose a patch-wise signal-to-noise-ratio to distinguish whether an internal or an external denoising method should be applied. Their fused result slightly improves over the stand-alone methods.
**RTF** (regression tree fields). Jancsary *et al.* [16] observe that depending on the image content some methods perform better than other. They consider RTFs based on a filterbank ($\text{RTF}_{plain}$), also additional exploitation of BM3D's output ($\text{RTF}_{BM3D}$), or a setting exploiting all the outputs of their benchmarked methods ($\text{RTF}_{all}$). The more methods the better their fusion result. The RTFs are learned on large datasets. It is also worth mentioning that following [16], Schmidt *et al.* [20] propose a cascade of regression tree fields (CRTF) working on deblurring and denoising and obtain good performances in both cases.
**NN** (neural nets/multi-layer perceptron). Burger *et al.* [17] pursue the success of MLP [12] in denoising, to learn the best fusion. They found the internal denoising methods to suit better images with artificial (human-made) contents, and external ones to work better for natural scenes. They argue against PatchSNR and consider that there is no trivial rule to decide among internal or external method on a patch-by-patch basis, and indeed their NN fusion produces the best denoising results to date. Unfortunately, the learning is quite intensive.
**AF** (anchored fusion). Timofte [21] clusters the patch space and for each cluster learns an anchored regressor from fused methods' patches to the fusion output.

### 1.2 Single Image Super-Resolution (SR)

Single image super-resolution (SR) is another active area [22–27] of image restoration aiming at recovering a high-resolution (HR) image from a low-resolution (LR) input image by inferring missing high frequency contents. We can roughly categorize the recent methods in:

**Non-neural network methods:**
**SR** (sparse representation) [28] generates a sparse representation/coding of each LR image patch, and then applies the coefficients of this representation to generate the HR image.
**A+** (adjusted anchored neighborhood regression) [29], considered to be an advanced version of ANR (anchored neighborhood regression) [30], learns sparse dictionaries and regressors anchored to the dictionary atoms.
**RFL** (super-resolution Forests) [31] maps low to high-resolution patches using random forests and anchored regressors as in A+.
**selfEx** (transformed self-exemplars) [32] introduces a self-similarity based image SR algorithm by applying transformed self-exemplars.

**Neural network methods:**
**SRCNN** (convolutional neural network) [33] learns an end-to-end mapping between the low/high-resolution images by a deep convolutional neural network.
**CSCN** (cascade of sparse coding network) [24] combines the key ingredients of deep learning with those of the sparse coding model.

### 1.3 Contributions

In this paper, we study the patch-by-patch fusion of image restoration methods with particular focus on recent top methods for both DN and SR tasks. To this end, we propose a generic 3D convolutional fusion architecture (3DCF) to learn the best combination of existing methods. Our three main contributions are:

1. We show the complementarity of different methods (*e.g.* internal vs. external).
2. We demonstrate that our method learns sophisticated correlation details from top methods to achieve the best reported results on a wide range of images.
3. The generality of our 3DCF method for both DN and SR.

The paper is organised as follows. Section 2 provides some insights and empirical evidence for the complementarity of the DN/SR methods and analyses oracle bounds for fusion. Section 3 motivates and introduces our novel 3DCF method with the necessary details and mathematical formulations. Section 4 presents the experiments and discusses the results. Section 5 concludes the paper.

## 2 Insights

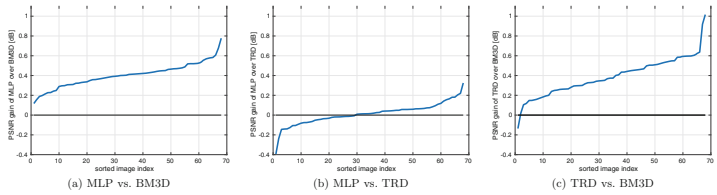
Our focus is fusion for improved image restoration results and particularly for denoising in the presence of additive white Gaussian noise (AWG), with validation on single image super-resolution. Here we analyse the complementarity of the restoration methods and fusion strategies.

### 2.1 Complementarity of Top Methods

Jancsary *et al.* [16], Burger *et al.* [17], and Zontak and Irani *et al.* [4], among others, already observed that each method works best for some particular image contents while being worse than others for other image regions.

First, we pair-wise compare the PSNR performances of BM3D (internal method), and MLP and TRD (external methods) on 68 images from the Berkeley dataset for AWG noise with $\sigma = 50$. The relative improvements (PSNR gain) are reported in Fig. 1. MLP is better than BM3D on all images but is worse than TRD on ∼40% of them. Also, BM3D is better than TRD on some images. We conclude there is no absolute winner at image-level.

Second, we compare pixel-wise or patch-wise and see that within the same image there is no absolute winner always getting the best result either. In Fig. 2 for one image altered with AWG noise, $\sigma = 50$, we report pixel-wise selections from BM3D (25.77 dB PSNR) and MLP (26.19 dB) to best match the ground truth image. Despite MLP being significantly better (+0.41 dB) on denoising this image, at pixel-level the results are almost equally divided between the methods. At patch-level (sizes $5 \times 5$ and $17 \times 17$ pixels) we have a similar pattern.




(a) MLP vs. BM3D (b) MLP vs. TRD (c) TRD vs. BM3D

**Fig. 1.** No absolute winner. Each method is trumped by another on some image.

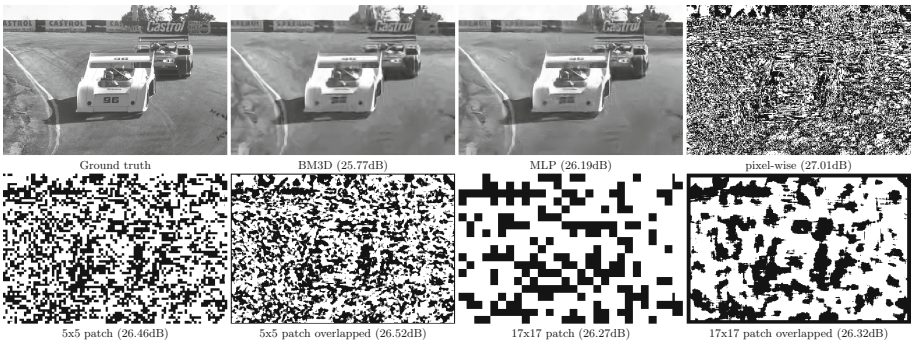



**Fig. 2.** An example of oracle pixel and patch-wise selections from BM3D and MLP outputs and the resulting PSNRs for AWG with $\sigma = 50$.

### 2.2 Average and Selection Fusion and Oracle Bounds

As shown in Fig. 1 for images and in Fig. 2 for patch or pixel regions, the denoising methods are complementary in their performance. Now we study a couple of fusion strategies at image level.

**Average fusion** directly averages the image results.
**Selection of non-overlapping patches** assumes that the fusion result contains non-overlapping (equal size) patches with the best image results of the fused methods (see Fig. 2). One needs to learn a patch-wise classifier.
**Selection of overlapping patches** is similar to the above one in that a patch-wise decision is made, but this time the patches overlap. The final fusion result is obtained by averaging the patches in the overlapped areas (see Fig. 2).

We work with BM3D and MLP, partly because BM3D is an internal while MLP is an external method, and partly because of the result in Fig. 1 where at image level MLP performs better than BM3D. Therefore, the results from fusing BM3D and MLP at patch-level are interesting to see.

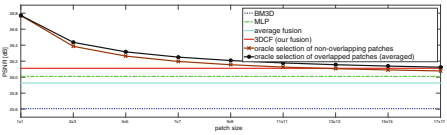



**Fig. 3.** Average PSNR [dB] comparison of BM3D [6] and MLP [12], average fusion, oracle selection of (overlapping or non-overlapping) patches, and our 3DCF fusion on 68 images, with AWG noise, $\sigma = 50$.

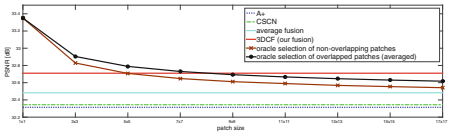



**Fig. 4.** Average PSNR [dB] comparison of A+ [29] and CSCN [24], average fusion, oracle selection of (overlapping or non-overlapping) patches, and our 3DCF fusion on Set14, upscaling factor $\times 2$.

In Fig. 3 we report how the chosen patch size affects the performance of a selection strategy, on the same Berkeley images corrupted with AWG noise, $\sigma = 50$. We report oracle results, an upper bound for such a strategy. In comparison we report the performance of the fused BM3D and MLP methods, as well as the results of the average fusion and our proposed 3DCF method. We note that (i) overlapping patches lead to better results (while significantly slower) than non-overlapping patches; (ii) the smaller the patch size the better the oracle results become; (iii) the average fusion leads to poorer performance than the fused MLP method; (iv) our 3DCF fusion results are comparable with those from the oracle selection strategies for patch sizes above $9 \times 9$.

Complementary, in Fig. 4 we start from the A+ and CSCN methods for the super-resolution (SR) task, where we use the Set14 images and an upscaling factor $\times 2$ (we use the settings described in the experimental section). As in the denoising case, (i) the smaller the patch size is, the better the oracle selection results get; (ii) the overlapped patches lead to better fusion results. However, for SR, (iii) the average fusion improves over both fused methods; (iv) our 3DCF

fusion is significantly better than the fused methods, the average fusion, and compares favorably to the oracle selection fusion for patch sizes above $5 \times 5$.

From these experiments we can conclude that the average and (patch) selection strategies for fusion - while conceptually simple - are either not leading to consistently improved results (case of average fusion) or their oracle upper bounds are quite tight given the difficulty of accurately classifying patches (case of selection strategy). Note that PatchSNR [19] is an example of a selection strategy and that NN [17], a neural network fusion method, reported better results than PatchSNR.

We therefore followed the combination paradigm for image fusion and design and trained an end-to-end 3D convolutional network from the results of two methods to the targeted restored image.

## 3 Learning Fine Features by 3D Convolution

### 3.1 Motivation and Related Work

Most of the existing neural network architectures apply spatial filters which address inputs such as 2D images. When it comes to videos, thus 3D inputs, these 2D convolutional neural networks (2DCNN) do not employ crucial information such as the temporal correlation. For example, in human action recognition, the motion information is not captured by 2DCNNs and Ji *et al.* [34] introduced a 3D convolutional neural network (3DCNN) method (see Fig. 5). The 3DCNN architecture has 1 hardwired layer, 3 convolutional layers and 2 subsampling layers. The spatial dimension of inputs $60 \times 40$ are gradually reduced to $1 \times 1$ by going through the network, *i.e.* 7 input frames have been converted into a 128-dimensional feature map capturing also the motion information. In the end, each element of the 128-dimensional feature map is fully connected to each unit in the last layer, then the action class is determined.

For performance improvements a brute force approach that proved successful is to deepen the (neural network) architecture [15,24]. Yet, the improvements decline significantly with the depth while the training time and the demand of hardware (GPU) resources increase. For example, experiments reported in [15] demonstrate that the bulk of the performance is achieved by the first stages in their denoising TRD method while the last 3 stages (from 8) bring merely 0.01 dB to it. In [22] it is shown for SR methods that the first stages are the most

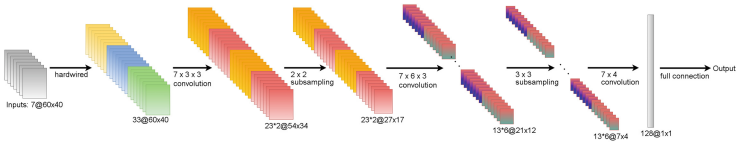



**Fig. 5.** 3DCNN proposed in [34] for human action recognition.

important and that adding more stages only slightly improves the performance (of A+) further.

On the other hand, for image restoration tasks such as SR it is common to recover the corrupted luminance component instead of the RGB image directly, and to interpolate the chroma. However, exploiting the correlation between corrupted RGB or even extra channels such as depth (D) or near-infrared (NIR) should be beneficial to the restoration task at the price of increased computation. For example, for denoising, Dabov *et al.* [35] apply the same grouping method on chroma channels as on the luminance, and they achieve better PSNR performances than by using BM3D [36] independently on three channels. To sum up, given several highly correlated (corrupted) channels/images, we have a better chance to high quality recovery.

It follows that we can consider the outputs of state-of-the-art methods as highly correlated images, which can be treated as the starting point of our proposed novel 3D convolutional fusion (3DCF) architecture.

### 3.2 Proposed Generic 3D Convolutional Fusion (3DCF)

**General Architecture.** As the starting point, we obtain several recovered outputs $\{\mathbf{I}_i\}_{i=1,\ldots,n}$ from the same corrupted image, with different methods. We stack those highly correlated images along the channel dimension, which brings us a multichannel image $\mathbf{I}_a = [\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_n]$ (see Fig. 6).

Furthermore, since directional gradient filters are sensitive to intensity changes and edges, and our task is about recovering fine image details based on the results of existing methods, hence the correlation between the recovered output image and its gradients can be exploited. To this end, we firstly have the naive average input image $\bar{\mathbf{I}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{I}_i$, then filter it with the first- and second-order gradients, in both the x and y direction,

$$\begin{aligned} \mathbf{F}_{1x} &= \begin{bmatrix} 1 & -1 \end{bmatrix} = \mathbf{F}_{1y}^T, \\ \mathbf{F}_{2x} &= \begin{bmatrix} 1 & -2 & 1 \end{bmatrix} /2 = \mathbf{F}_{2y}^T, \end{aligned} \tag{1}$$

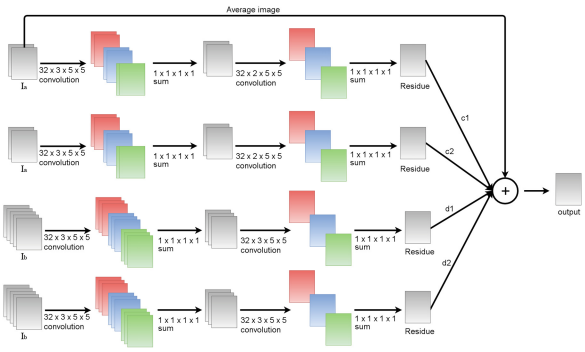



**Fig. 6.** Proposed 3D convolutional fusion method (3DCF).

followed by stacking those gradient filtered- and average images along the channel dimension, we have another input $\mathbf{I}_b$ as our second starting point,

$$\mathbf{I}_b = [\mathbf{F}_{2x} * \bar{\mathbf{I}}, \mathbf{F}_{1x} * \bar{\mathbf{I}}, \bar{\mathbf{I}}, \mathbf{F}_{1y} * \bar{\mathbf{I}}, \mathbf{F}_{2y} * \bar{\mathbf{I}}]. \tag{2}$$

Next, we intensively explore the correlation within $\mathbf{I}_a, \mathbf{I}_b$ by introducing the 3D convolutional layer. Related recent works such as [15,24,33] mainly exploit deep features with spatial filters. In that case, given the image has multiple channels, they are independently filtered and eventually summed up as the input for the next layer, while the correlations among the channels may not be accurately captured. That is the main reason behind our idea – to fully explore the fine details along the channel dimension. As far as we know, this is the first time that a 3D layer is introduced to address low level image tasks.

Our next step is to update the input images $\mathbf{I}_{a,b}$[1] with a 3D hidden layer,

$$\mathbf{H}_1^{a,b}(\mathbf{I}_{a,b}) = \tanh(\mathbf{W}_1^{a,b} * \mathbf{I}_{a,b} + \mathbf{B}_1^{a,b}), \tag{3}$$

where $\mathbf{W}_1^{a,b}$ correspond $n$ 3D filters with $c \times h \times w$ kernel size and $\mathbf{B}_1^{a,b}$ are biases. In our design, due to a tradeoff between the memory constraint and speed, we recommend $n$ and $c \times h \times w$ to be 32 and $3 \times 5 \times 5$ for $\mathbf{I}_b$, along with setting pad to be 0, so that we have the output with same size as input. The default size of filters regarding $\mathbf{I}_a$ showed in Fig. 6 are also determined for the same reason. Besides we use hyperbolic tangent (tanh) as activation function because we allow negative value updates to pass through the network rather than ignore them as ReLU [37] does. In the following step, we use a naive convolutional layer with a single $1 \times 1 \times 1$ filter, which is equivalent to sum up the input $\mathbf{H}_1^{a,b}$

$$\mathbf{H}_2^{a,b}(\mathbf{H}_1^{a,b}(\mathbf{I}_{a,b})) = \tanh(w_2^{a,b} \sum_k \mathbf{H}_{1,k}^{a,b}(\mathbf{I}_{a,b}) + b_2^{a,b}\mathbf{1}), \tag{4}$$

where $w_2^{a,b}$, $b_2^{a,b}$ are the scalar weights and biases, resp. We consider the above two steps as one inference stage. Another important difference between our proposed method and many other neural network methods is that we reconstruct the image residue instead of the image itself (see Fig. 6). Normally, the perturbation on image residues during the optimization is smaller than the one on image values, which increases the odds that the learning process eventually converges. Secondly, residue reconstruction substantiates the robust performance of our general architecture for distinct image restoration tasks. After going through $n$ inference stages, we come to the reconstruction stage,

$$\mathbf{R}_{a,b}(\mathbf{I}_{a,b}) = (w_{2n+2}^{a,b} \sum \mathbf{H}_{2n+1}^{a,b} \circ \mathbf{H}_{2n}^{a,b} \ldots \mathbf{H}_2^{a,b} \circ \mathbf{H}_1^{a,b}(\mathbf{I}_{a,b}) + b_{2n+2}^{a,b}\mathbf{1}), \tag{5}$$

where $\mathbf{R}_{a,b}(\mathbf{I}_{a,b})$ are the image residues we want to predict. In order to robustify the performance of our network, we simply duplicate the above mentioned process for each input image array $\mathbf{I}_a$ and $\mathbf{I}_b$ $n$ times, which gives us $2n$ separate

[1] Here we abuse of notation, $\mathbf{I}_{a,b}$ indicates two inputs $\mathbf{I}_a, \mathbf{I}_b$.

networks with the same architecture. In the end we sum up the residues and the average image to obtain our output image $F(\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_n)$,

$$F(\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_n) = \frac{1}{n}\sum_k \mathbf{I}_k + \sum_k (c_k \mathbf{R}_a^k(\mathbf{I}_a) + d_k \mathbf{R}_b^k(\mathbf{I}_b)), \tag{6}$$

where $c_k, d_k$ are the coefficients to weight the residues.

**Training.** Our main task is to learn the parameters $\mathbf{\Theta} = (\mathbf{W}, \mathbf{B})$ of the non-linear map $F$. To this end, we minimize the loss function $l(\mathbf{\Theta})$, which computes the Euclidean distance (mean square error (MSE)) between the output image $F(\mathbf{I}_1^i, \mathbf{I}_2^i, \ldots, \mathbf{I}_n^i)$ and ground truth image $\mathbf{I}_g^i$ contained in our training set, *i.e.*,

$$l(\mathbf{\Theta}) = \sum_i \|F(\mathbf{I}_1^i, \mathbf{I}_2^i, \ldots, \mathbf{I}_n^i; \mathbf{\Theta}) - \mathbf{I}_g^i\|_2^2. \tag{7}$$

The choice of the cost function is appropriate since PSNR is the main evaluation method of image restoration tasks and stands in monotonic relation with MSE. During the training stage, we update the weights/biases with standard back propagation [38,39].

Currently, the optimization of the loss function is dominated by the stochastic gradient descent (SGD) method [40], for example in [15,24,33]. Basically, at the $t+1$-th iteration they update the parameters $\mathbf{\Theta}_{t+1}$ with the previous parameter update $\mathbf{\Lambda}_t$ and negative gradient $\nabla l(\mathbf{\Theta})$,

$$\begin{aligned} \mathbf{\Lambda}_{t+1} &= a\mathbf{\Lambda}_t - b\nabla l(\mathbf{\Theta}_t), \\ \mathbf{\Theta}_{t+1} &= \mathbf{\Theta}_t + \mathbf{\Lambda}_{t+1}, \end{aligned} \tag{8}$$

where $a, b$ are the momentum and learning rate, resp. One weakness of SGD is that the improvements gained from the optimization decrease rapidly with growing iteration steps. In such case, SGD may not be able to recover accurate details from highly corrupted images. This is the main reason why we prefer adaptive moment estimation (Adam) [41] as our optimization method. The Adam method is stated as follows,

$$\begin{aligned} \mathbf{\Lambda}_t &= a_1\mathbf{\Lambda}_{t-1} + (1-a_1)\nabla l(\mathbf{\Theta}_t), \\ \mathbf{K}_t &= a_2\mathbf{K}_{t-1} + (1-a_2)\nabla l(\mathbf{\Theta}_t)^2, \end{aligned} \tag{9}$$

where $a_1, a_2$ are moments and $\mathbf{\Theta}_{t+1}$ is updated based on $\mathbf{\Lambda}_t, \mathbf{K}_t$,

$$\mathbf{\Theta}_{t+1} = \mathbf{\Theta}_t - b\frac{\sqrt{1-(a_2)^t}}{1-(a_1)^t}\frac{\mathbf{\Lambda}_t}{\sqrt{\mathbf{K}_t} + \epsilon}, \tag{10}$$

here $b$ is the learning rate and $\epsilon$ is used to avoid explosion. At the beginning of the iterations, the cost of $l(\mathbf{\Theta})$ converges considerably faster than SGD. Moreover, Eq. (10) shows that the magnitudes of parameter updates are independent of the rescaling of the gradient, therefore it provides a relatively fast convergence speed even after a large amount of iterations.

## 4 Experiments

In the following we describe the experimental setup and datasets used to validate our 3DCF approach on both the SR and DN tasks, then discuss the results.

### 4.1 Experimental Setup and Datasets

**DN.** Like most DN-related papers we add white Gaussian (AWG) noise to ground truth images to create our corrupted images. 3 standard deviations $\sigma \in \{15, 25, 50\}$ are chosen to measure the performance of 3DCF. Under such conditions, we compare our 3DCF with state-of-the-art DN methods as described in the introductory Sect. 1: BM3D [6], LSSC [11], EPLL [10], opt-MRF [14], CRTF [20], WNNM [8], CSF [13], TRD [15], MLP [12], as well as the NN [17] fusion method.

We use the same training data mentioned in [15], *i.e.*, 400 cropped images with $180 \times 180$ size from the training part of the Berkeley segmentation dataset (BSD) [42]. We evaluate our method on the 68 test images as in [43], a standard benchmark employed by top methods like [13,15].

**SR.** For SR we use the same 3DCF architecture as for DN and test it on the standard benchmarks Set5 [44], Set14 [45] (as proposed in [30]) and B100 [29] with 5, 14, 100 images resp., which are widely adopted by the recent literature. To obtain the LR images, according to many of the SR works, we firstly convert the ground truth image into YCbCr color space, then downscale the luminance channel with bicubic interpolation. Our training data is formed by the 200 training BDS images of size $321 \times 481$ from which we extract millions of LR-HR image pairs. We report PSNR and SSIM results for the latest methods with top performances: A+ [29], SRCNN(L) [33], RFL [31], SelfEx [32], CSCN [24].

### 4.2 Implementation Details

We implement our 3DCF method with Caffe [46]. 3DCF is used in the same form for both DN and SR. For clarity and ease of understanding and deployment we prefer stacking two top methods along the channel dimension as our one starting point $\mathbf{I}_a$. For DN we use MLP [12], an external neural network method, and BM3D [36], an internal method. Thus, such combination of two top methods increases our chance to take advantage of the strengths and overcome the weaknesses of both worlds. For SR, the CSCN [24] and A+ [29] are our favorite because of similar reasons – one from CNN and another from non-CNN type of methods. The starting point $\mathbf{I}_b$ is simply obtained by the average image of two methods as well as its corresponding first- and second order gradients along x/y direction. To enable 3DCF to recover more accurate details, we use two networks for each starting point $\mathbf{I}_a, \mathbf{I}_b$ (See Fig. 6), while slightly perturbing the value as the input of each activation, by multiplying $-1$. For the same reason we fix the

coefficients $c_1, c_2$ to be 1 and 0.1. So are the coefficients $d_1, d_2$. Now Eq. (6) looks as follows:

$$F(\mathbf{I}_1, \mathbf{I}_2) = \frac{1}{2}(\mathbf{I}_1 + \mathbf{I}_2) + \mathbf{R}_a^1(\mathbf{I}_a) + 0.1\mathbf{R}_a^2(\mathbf{I}_a) + \mathbf{R}_b^1(\mathbf{I}_b) + 0.1\mathbf{R}_b^2(\mathbf{I}_b). \quad (11)$$

For the sake of time complexity and memory saving, each network showed in Fig. 6 has 4 layers, and the filter size $n \times c \times h \times w$ is set to be $(32 \times 3 \times 5 \times 5, 1 \times 1 \times 1 \times 1, 32 \times 3 \times 5 \times 5, 1 \times 1 \times 1 \times 1)$ for $\mathbf{I}_a$, while $\mathbf{I}_b$ has the almost same settings except for the 3rd layer with $32 \times 2 \times 5 \times 5$. We also set the channel-, height- and width stride to be 1 for all layers. It is expected that our output is a single image with the same spatial size as the input image. To this end, the channel-, height- and width padding size are determined to be $(1 \times 2 \times 2, 0 \times 0 \times 0, 0 \times 2 \times 2, 0 \times 0 \times 0)$ for $\mathbf{I}_a$, and for $\mathbf{I}_b$ we follow the same setup except the first layer parameters are determined to be $0 \times 2 \times 2$. We also initialize the weights by a Gaussian distribution with standard deviation 0.05 for convolutional layers, and put the weight to 1 for sum layers, and the bias to 0 for all cases.

Meanwhile, we simply use the default learning- and decay rate 1 when learning the weights/biases for each layer. In the end, for Eq. 10 the learning rate $b$ for the whole network is considered to be 0.001, the moments $a_1, a_2$ have the default value 0.9, 0.999, and $\epsilon$ is also set to the default $10^{-8}$. It is worth mentioning that all the parameters are exactly the same for the two tasks, DN and SR.

### 4.3 Denoising Results

We demonstrate our 3DCF method on 68 standard images [43] from BSD [42]. We apply the best setup for the compared methods, already described in the introductory Sect. 1. CRTF [20] has 5 cascades, CSF [13] employs the $7 \times 7$ filter, the same as TRD [15] with 8 stages. Table 1 shows that our 3DCF method achieves top performances compared to other methods for 3 different standard deviations. For example, if we start our method with BM3D [6] and MLP [12], we are 0.11 dB and 0.1 dB better than the top standalone method MLP for $\sigma \in \{25, 50\}$. Due to the lack of an MLP model trained for $\sigma = 15$, we use BM3D+TRD instead. Still, the performance of our 3DCF is consistent with the other cases, 0.09 dB higher than TRD, the currently best method. Interestingly, if we compare 3DCF with the NN fusion method under the same conditions, that is, with the same starting methods BM3D and MLP, the proposed method outperforms NN with 0.15 and 0.07 dB for $\sigma \in \{15, 25\}$. Such observation confirms the non-trivial improvements achieved by 3DCF. Moreover, Fig. 7 indicates that the naive average of MLP and BM3D is even worse than MLP. Besides, it is also notable from Fig. 7 that the PSNR gradually increases with the growth of back propagation. 3DCF is robust to the fused methods, TRD + MLP leads to relative improvements comparable with those achieved starting from BM3D + MLP or BM3D + TRD.

**Table 1.** Average PSNR values [dB] on 68 images from BSD dataset as in [43] for $\sigma \in \{15, 25, 50\}$. The best is with bold. The results with (*) are obtained from [15].

| Method | $\sigma$ | | |
|---|---|---|---|
| | 15 | 25 | 50 |
| BM3D [6] | 31.08 | 28.57 | 25.61 |
| *LSSC [11] | 31.27 | 28.70 | 25.72 |
| *EPLL [10] | 31.19 | 28.68 | 25.67 |
| *opt-MRF [14] | 31.18 | 28.66 | 25.70 |
| *$\text{CRTF}_5$ [20] | | 28.75 | |
| *WNNM [8] | 31.37 | 28.83 | 25.83 |
| $\text{CSF}_{7\times7}$ [13] | 31.24 | 28.71 | |
| $\text{TRD}^8_{7\times7}$ [15] | 31.42 | 28.93 | 25.99 |
| MLP [12] | | 28.96 | 26.01 |
| NN (BM3D+MLP) [17] | | 28.92 | 26.04 |
| 3DCF (BM3D+TRD) | **31.51** | 29.03 | 26.10 |
| 3DCF (BM3D+MLP) | | **29.07** | 26.11 |
| 3DCF (TRD+MLP) | | **29.07** | **26.12** |

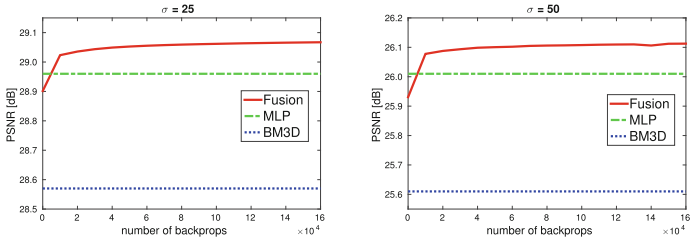



**Fig. 7.** PSNR versus backprops on 68 images for $\sigma \in \{25, 50\}$.

### 4.4 Super Resolution Results

The PSNR and SSIM results are listed in Table 2. Here our 3DCF fuses A+ [29] with CSCN [24]. Note that we modify the steps of downscaling the image for CSCN to be consistent with other methods including A+ and SRCNN(L). That is the reason why we obtain different PSNR results for CSCN than in the original work [24]. As in the case of DN, our 3DCF shows significant improvements over the starting methods. The PSNR improvements vary from 0.11 dB on (B100, $\times3$) to 0.35 dB on (Set5, $\times2$) over the best result from SRCNN (L,with largest model). The SSIM improvements follow the same trend. Note that for SR, the naive average fusion of A+ and CSCN results improves over both fused methods. However, our 3DCF results are on average 0.2 dB higher than the average fusion, as shown in Fig. 8.

**Table 2.** Average PSNR/SSIMs for upscaling factors ×2, ×3, and ×4 on datasets Set5, Set14, and B100. The best results are with bold.

| Dataset | Scale | A+ [29] | SRCNN(L) [33] | RFL [31] | SelfEx [32] | CSCN [24] | **3DCF (CSCN+A+)** |
|---|---|---|---|---|---|---|---|
| | | PSNR/SSIM | PSNR/SSIM | PSNR/SSIM | PSNR/SSIM | PSNR/SSIM | PSNR/SSIM |
| Set5 | x2 | 36.56/0.9612 | 36.68/0.9609 | 36.52/0.9589 | 36.50/0.9577 | 36.55/0.9605 | **37.03/0.9631** |
| | x3 | 32.67/0.9199 | 32.83/0.9198 | 32.50/0.9164 | 32.63/0.9190 | 32.68/0.9197 | **33.11/0.9255** |
| | x4 | 30.33/0.8749 | 30.52/0.8774 | 30.17/0.8715 | 30.32/0.8728 | 30.44/0.8779 | **30.82/0.8865** |
| Set14 | x2 | 32.32/0.9607 | 32.52/0.9612 | 32.30/0.9599 | 32.27/0.9584 | 32.36/0.9593 | **32.71/0.9623** |
| | x3 | 29.16/0.8869 | 29.35/0.8886 | 29.07/0.8842 | 29.19/0.8873 | 29.19/0.8850 | **29.48/0.8907** |
| | x4 | 27.33/0.8277 | 27.53/0.8285 | 27.23/0.8251 | 27.43/0.8279 | 27.41/0.8256 | **27.69/0.8334** |
| B100 | x2 | 31.16/0.8857 | 31.32/0.8874 | 31.13/0.8842 | 31.15/0.8860 | 31.20/0.8836 | **31.48/0.8899** |
| | x3 | 28.25/0.7824 | 28.37/0.7853 | 28.20/0.7814 | 28.25/0.7821 | 28.28/0.7804 | **28.48/0.7881** |
| | x4 | 26.76/0.7073 | 26.86/0.7089 | 26.70/0.7068 | 26.81/0.7078 | 26.83/0.7072 | **26.99/0.7147** |

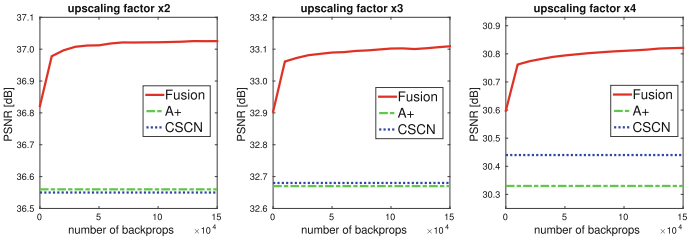



**Fig. 8.** PSNR versus backprops on Set5 dataset for upscaling factors ×2, ×3, ×4.

### 4.5 Other Aspects

**Visual assessment.** In general, the visual results are consistent with PSNR results. Some image results are shown in Fig. 9 for DN and in Fig. 10 for SR. We can observe that the 3DCF results have generally fewer artifacts and sharper edges in comparison with the other methods.

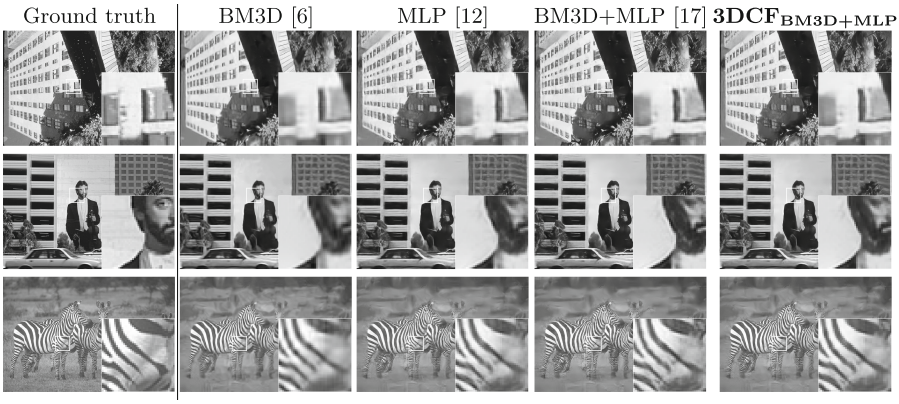



**Fig. 9.** Denosing results for $\sigma = 50$. Best zoomed on screen.

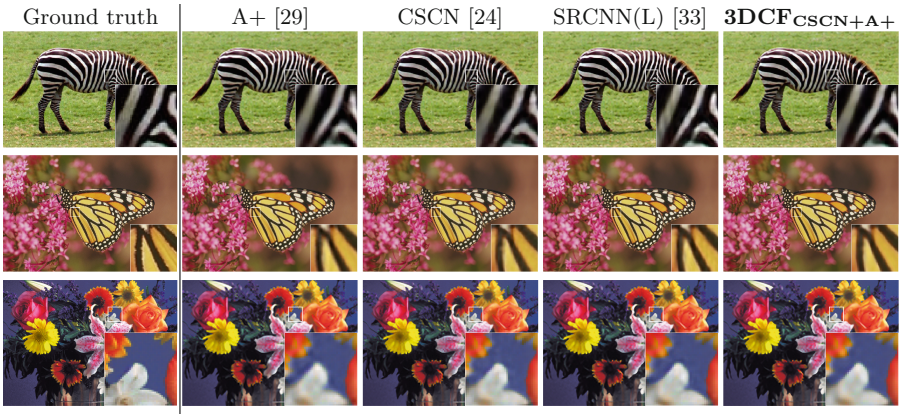



**Fig. 10.** Super-resolution results (×4). Best zoomed on screen.

**Running time.** 3DCF runs on roughly 0.04 second per 321 × 480 image on nVidia TitanX GPU, which is quite competitive and shows that at the price of slight increase in processing time one could fuse available image restoration results. 3DCF needs about 5 h training time to obtain meaningful improvements over the fused methods, and this is mainly due to the Adam method.

**General.** To summarize, our 3DCF method shows wide adaptability for two important image restoration tasks, DN and SR, with non-trivial improvements. Also, the training and running times of 3DCF are competitive in comparison with other neural network architectures. For certain combinations of existing methods our proposed fusion method only shows mild progress, for example for the case of TRD+MLP (see Table 1). This sensitivity to the starting point drives us to be careful of the choice of starting methods.

## 5 Conclusions

We propose a novel 3D convolutional fusion (3DCF) network for image restoration. With the same settings, for both single image super resolution and image denoising, we achieve significant improvements over the fused methods and other fusion methods on several standard benchmarks. For speeding up the training, we apply an adaptive moment estimation method. The testing and training times are also competitive to other recent deep neural networks.

**Acknowledgments.** This work was supported by the ERC project *VarCity (#273940)*, the ETH General Fund (OK) and by an Nvidia GPU grant.

## References


1. Katsaggelos, A.K.: Digital Image Restoration. Springer Publishing Company, Incorporated, Heidelberg (2012)
2. Stathaki, T.: Image Fusion: Algorithms and Applications. Academic Press, Amsterdam (2011)
3. Lebrun, M., Colom, M., Buades, A., Morel, J.: Secrets of image denoising cuisine. Acta Numerica **21**, 475–576 (2012)
4. Zontak, M., Mosseri, I., Irani, M.: Separating signal from noise using patch recurrence across scales. In: CVPR (2013)
5. Buades, A., Coll, B., Morel, J.M.: A non-local algorithm for image denoising. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 2, pp. 60–65. IEEE (2005)
6. Dabov, K., Foi, A., Katkovnik, V., Egiazarian, K.: Image denoising by sparse 3d transform-domain collaborative filtering. IEEE Trans. Image Process. **16**, 2080–2095 (2007)
7. Yu, G., Sapiro, G.: DCT image denoising: a simple and effective image denoising algorithm (2011)
8. Gu, S., Zhang, L., Zuo, W., Feng, X.: Weighted nuclear norm minimization with application to image denoising. In: CVPR (2014)
9. Yu, G., Sapiro, G., Mallat, S.: Solving inverse problems with piecewise linear estimators: from gaussian mixture models to structured sparsity. IEEE Trans. Image Process. **21**(5), 2481–2499 (2012)
10. Zoran, D., Weiss, Y.: From learning models of natural image patches to whole image restoration. In: IEEE International Conference on Computer Vision, pp. 479–486 (2011)
11. Mairal, J., Bach, F., Ponce, J., Sapiro, G., Zisserman, A.: Non-local sparse models for image restoration. In: IEEE 12th International Conference on Computer Vision, pp. 2272–2279 (2009)
12. Burger, H., Schuler, C., Harmeling, S.: Image denoising: can plain neural networks compete with bm3d? In: IEEE Computer Vision and Pattern Recognition, pp. 2392–2399 (2012)
13. Schmidt, U., Roth, S.: Shrinkage fields for effective image restoration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2774–2781 (2014)
14. Chen, Y., Pock, T., Ranftl, R., Bischof, H.: Revisiting loss-specific training of filter-based MRFs for image restoration. In: Weickert, J., Hein, M., Schiele, B. (eds.) GCPR 2013. LNCS, vol. 8142, pp. 271–281. Springer, Heidelberg (2013). doi:10.1007/978-3-642-40602-7_30
15. Chen, Y., Yu, W., Pock, T.: On learning optimized reaction diffusion processes for effective image restoration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5261–5269 (2015)
16. Jancsary, J., Nowozin, S., Rother, C.: Loss-specific training of non-parametric image restoration models: a new state of the art. In: IEEE European Conference of Computer Vision (2012)
17. Burger, H.C., Schuler, C., Harmeling, S.: Learning how to combine internal and external denoising methods. In: Weickert, J., Hein, M., Schiele, B. (eds.) GCPR 2013. LNCS, vol. 8142, pp. 121–130. Springer, Heidelberg (2013). doi:10.1007/978-3-642-40602-7_13

18. Levin, A., Nadler, B., Durand, F., Freeman, W.T.: Patch complexity, finite pixel correlations and optimal denoising. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7576, pp. 73–86. Springer, Heidelberg (2012). doi:10.1007/978-3-642-33715-4_6
19. Mosseri, I., Zontak, M., Irani, M.: Combining the power of internal and external denoising. In: IEEE International Conference on Computational Photography (ICCP), pp. 1–9 (2013)
20. Schmidt, U., Jancsary, J., Nowozin, S., Roth, S., Rother, C.: Cascades of regression tree fields for image restoration (2014)
21. Timofte, R.: Anchored fusion for image restoration. In: ICPR (2016)
22. Timofte, R., Rothe, R., Van Gool, L.: Seven ways to improve example-based single image super resolution. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
23. Kim, J., Kwon Lee, J., Mu Lee, K.: Accurate image super-resolution using very deep convolutional networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
24. Wang, Z., Liu, D., Yang, J., Han, W., Huang, T.: Deep networks for image super-resolution with sparse prior. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 370–378 (2015)
25. Agustsson, E., Timofte, R., Van Gool, L.: Regressor basis learning for anchored super-resolution. In: ICPR (2016)
26. Dai, D., Timofte, R., Van Gool, L.: Jointly optimized regressors for image super-resolution. Comput. Graph. Forum **34**, 95–104 (2015)
27. Timofte, R., De Smet, V., Van Gool, L.: Semantic super-resolution: when and where is it useful? Comput. Vis. Image Underst. **142**, 1–12 (2016)
28. Yang, J., Wright, J., Huang, T.S., Ma, Y.: Image super-resolution via sparse representation. IEEE Trans. Image Process. **19**, 2861–2873 (2010)
29. Timofte, R., De Smet, V., Van Gool, L.: A+: adjusted anchored neighborhood regression for fast super-resolution. In: Cremers, D., Reid, I., Saito, H., Yang, M.-H. (eds.) ACCV 2014. LNCS, vol. 9006, pp. 111–126. Springer, Cham (2015). doi:10.1007/978-3-319-16817-3_8
30. Timofte, R., Smet, V., Gool, L.: Anchored neighborhood regression for fast example-based super-resolution. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1920–1927 (2013)
31. Schulter, S., Leistner, C., Bischof, H.: Fast and accurate image upscaling with super-resolution forests. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3791–3799 (2015)
32. Huang, J.B., Singh, A., Ahuja, N.: Single image super-resolution from transformed self-exemplars. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5197–5206. IEEE (2015)
33. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks (2015)
34. Ji, S., Xu, W., Yang, M., Yu, K.: 3d convolutional neural networks for human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **35**, 221–231 (2013)
35. Dabov, K., Foi, A., Katkovnik, V., Egiazarian, K.: Color image denoising via sparse 3d collaborative filtering with grouping constraint in luminance-chrominance space. In: IEEE International Conference on Image Processing, ICIP 2007, vol. 1, p. I-313. IEEE (2007)
36. Dabov, K., Foi, A., Katkovnik, V., Egiazarian, K.: Image denoising by sparse 3-d transform-domain collaborative filtering. IEEE Trans. Image Process. **16**, 2080–2095 (2007)

37. Nair, V., Hinton, G.E.: Rectified linear units improve restricted Boltzmann machines. In: Proceedings of the 27th International Conference on Machine Learning (ICML 2010), pp. 807–814 (2010)
38. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. Cogn. Model. **5**, 1 (1988)
39. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**, 2278–2324 (1998)
40. Bottou, L.: Large-scale machine learning with stochastic gradient descent. In: Lechevallier, Y., Saporta, G. (eds.) Proceedings of COMPSTAT 2010, pp. 177–186. Springer, Heidelberg (2010)
41. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
42. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Proceedings of the 8th International Conference on Computer Vision, vol. 2, pp. 416–423 (2001)
43. Roth, S., Black, M.J.: Fields of experts. Int. J. Comput. Vis. **82**, 205–229 (2009)
44. Bevilacqua, M., Roumy, A., Guillemot, C., Alberi-Morel, M.L.: Low-complexity single-image super-resolution based on nonnegative neighbor embedding (2012)
45. Zeyde, R., Elad, M., Protter, M.: On single image scale-up using sparse-representations. In: Boissonnat, J.-D., Chenin, P., Cohen, A., Gout, C., Lyche, T., Mazure, M.-L., Schumaker, L. (eds.) Curves and Surfaces 2010. LNCS, vol. 6920, pp. 711–730. Springer, Heidelberg (2012). doi:10.1007/978-3-642-27413-8_47
46. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: convolutional architecture for fast feature embedding. In: Proceedings of the ACM International Conference on Multimedia, pp. 675–678. ACM (2014)