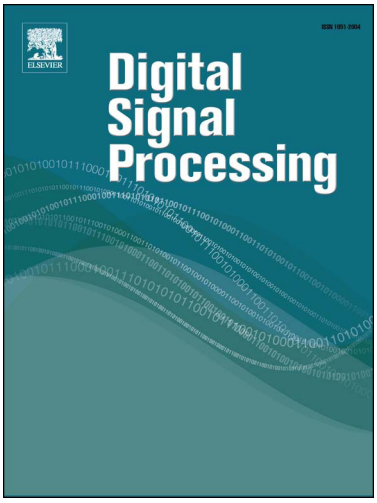
Accepted Manuscript



Multimedia super-resolution via deep learning: a survey

Khizar Hayat

PII: S1051-2004(18)30526-8
DOI: https://doi.org/10.1016/j.dsp.2018.07.005
Reference: YDSPR 2366

To appear in: *Digital Signal Processing*

Please cite this article in press as: K. Hayat, Multimedia super-resolution via deep learning: a survey, *Digit. Signal Process.* (2018), https://doi.org/10.1016/j.dsp.2018.07.005

Highlights

- An extensive survey of deep learning based multimedia super-resolution (SR) methods.
- Encompasses the three main aspects of multimedia *viz*. image, video and 3D, especially the depth map.
- An exhaustive coverage of the SR benchmarking methods and databases and eventual benchmarking of the surveyed methods on a popular database.

Multimedia Super-Resolution via Deep Learning: A Survey

Khizar Hayat

Computer Science Section (DMPS),
College of Arts and Sciences,
University of Nizwa,
Sultanate of Oman

Abstract

The recent phenomenal interest in convolutional neural networks (CNNs) must have made it inevitable for the super-resolution (SR) community to explore its potential. The response has been immense and in the last three years, since the advent of the pioneering work, there appeared too many works not to warrant a comprehensive survey. This paper surveys the SR literature in the context of deep learning. We focus on the three important aspects of multimedia - namely image, video and multi-dimensions, especially depth maps. In each case, first relevant benchmarks are introduced in the form of datasets and state of the art SR methods, excluding deep learning. Next is a detailed analysis of the individual works, each including a short description of the method and a critique of the results with special reference to the benchmarking done. This is followed by minimum overall benchmarking in the form of comparison on some common dataset, while relying on the results reported in various works.

Keywords: Super-resolution, Deep learning, Convolutional neural network

1. Introduction

For the last decade or so, interest in deep learning has skyrocketed. Deep learning had been around for many years but got little interest from researchers; may be due to low computing powers and network speeds or may be huge unstructured data were not commonplace? Now everybody is talking about it and want to contribute in the form of tutorials, special courses, articles, blogs, source codes, APIs and original research. And that has helped a lot in developing a general understanding about the underlying concepts. One can't recall any other issue got that much help available in such a short span of time, be it iPython(Jupyter)notebooks on GitHub, Matlab implementations (MatConvNet [1]), C++ API in the form of Caffe [2] or R-language based source codes or even patents. [3]. If you are interested, there's no dearth of resources and they will make sure you learn it.

As far as super-resolution is concerned, the pioneering work on the role of deep learning is as fresh as 2014 [4]. Since then, there has been a mushroom growth and several works have appeared focusing not only images but also videos and higher dimensional multimedia data, especially depth maps or range images, digital elevation models (DEMs) and multispectral images. The cornerstone of all the relevant research is the single image super-resolution method [5], called Super-Resolution Convolutional Neural Network (SRCNN), which is an extension

Email address: khizar.hayat@unizwa.edu.om (Khizar Hayat)

of the pioneering work [4] called SRCNN-Ex by its authors. The importance of SRCNN can be gauged by the fact that since its appearance, you will hardly find a super-resolution work not using it as one of its benchmarks.

In this paper we attempt to survey the deep learning literature in the context multimedia super-resolution. The main focus is on three areas, *viz.* still images, videos and higher dimensions, especially the range data. For each of the three, we first introduce the relevant benchmarks before reviewing the contemporary literature on deep learning based super-resolution, which is followed by a comparison of important methods on the basis of a common dataset. In case of higher dimensions, however, a common dataset is elusive. The description of benchmarks is in the form of publicly available datasets and important super-resolution methods that are not deep learning based; the latter will be introduced in the subsequent discussion, none the less. For benchmarking among methods, we rely on the results reported in the literature and have therefore compiled the tabular data from relevant resources.

The rest of the paper is arranged as follows. Section 2 presents essential background concepts with reference to deep learning and super-resolution. Section 3 surveys the contemporary deep learning works on image SR. A special part of this section concerns the importance of SRCNN with references from literature. The next two sections, i.e. Section 4 and Section 5, follow the same approach for videos and 3D/depth maps, respectively. Section 6 highlights the need to adopt a difference of approaching to multimedia, while Section 7 concludes the paper.

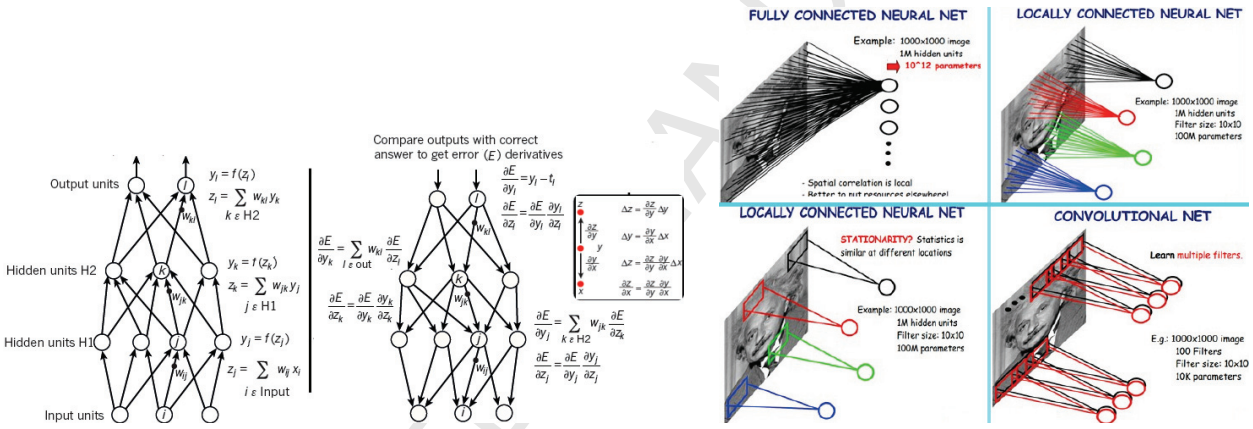
2. Background

Before visiting the deep learning based super-resolution literature, it is expedient to give a brief description of the background concepts needed for the understanding of this work.

2.1. Deep Learning

Classical machine learning (ML) techniques are characterized by the application of the underlying algorithm to a select group of features extracted after a laborious and intelligent processing and pre-processing. The key is to fine tune and select the best features which is normally time consuming and not possible without considerable expertise and knowledge of the domain. These techniques are thus handicapped by their limited ability to process raw natural data; such data are always huge [6]. Representation learning pertains to the auto-discovery of representations from input raw data. "Deep-learning methods are representation-learning methods with multiple levels of representation having simple but non-linear modules" to get "higher and abstract representation". The goal is to use the composition of enough such transformations in order to learn very complex functions. The peculiarities are, a) the non-involvement of humans in designing layers of features and b) employment of a generic learning procedure to get features from data. "Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction ... Deep learning discovers intricate structure in large data sets by using the backpropagation algorithm to indicate how a machine should change its internal parameters that are used to compute the representation in each layer from the representation in the previous layer" [7].

According to Hinton [8], the use of backpropagation algorithm for learning multiple layers of non-linear features can be traced back to 1970's. It, however, did not get attention of ML community for its perceived poor exploitation of multiple hidden layers, time-delay and convolutional nets being the exceptions; in addition, its performance was poor in recurrent networks. It was employed with supervised approaches and was thus aptly criticized for its slowness, requirement of labeled data and the high chance of getting stuck in poor local optima. By 2006 [9], with the advent of high processing speeds, these limitations of backpropagation were overcome by using unsupervised learning [10] in place of supervised one and thus applying it directly on raw rather than the labeled data. The idea is to "keep the efficiency and simplicity of using a gradient method for adjusting the weights, but use it for modeling the structure of the sensory input." In other words, "adjust the weights to maximize the probability that a generative model would have produced the sensory input", i.e. learn $p(data)$ not $p(label|data)$. As can be observed from Fig. 1.a, backpropagation is just a practical application of the chain rule of derivatives.



(a) Backpropagation (after [7])

(b) CNN as seen from NN perspective [6]

Figure 1: CNN basics.

The power of neural networks (NNs) lies in their ability to approximate any continuous function [11], as demonstrated in [12, 13]. A n-layer NN is characterized by an input layer and $n-1$ hidden layers. In a fully connected version, neurons of a given layer are not connected with each other but are fully connected to those of its adjacent layer. The "forward pass of a fully-connected layer corresponds to one matrix multiplication followed by a bias offset and an activation function" [11]. The commonly used activation functions are, sigmoid, *tanh* and the Rectified Linear Unit (ReLU [14]). Parameter initializations for the first forward pass of a NN can be carried out in a variety of ways (like all zero weights, random weights etc). The outputs of each forward pass, during the training phase, are compared with the ground truth to get the error signal. Backpropagation of this error signal yields the derivative/gradient for learning that serves to readjust the weights of the parameters for the next forward pass. The process repeats itself till acceptable level of convergence whereby the optimized parameters should ideally classify each subsequent test case correctly.

The birth of Convolutional neural networks (CNN) or ConvNets can be traced back to 1988 [15][1] wherein backpropagation was employed to train a NN to classify handwritten digits. Subsequent works by LeCun evolved into what was later known as LeNet5 [17]. After that there's virtual lull till late noughties [18] when GPUs were efficient enough to culminate in the work [19]. Since then a floodgate has opened and we hear of various architectures in the form of AlexNet [20], ZFNet [21], GoogLeNet [22] DenseNet [23] etc.; for a detailed overview one can consult [18, 24].

The metamorphosis from fully connected NN to locally connected NN to CNN is illustrated in Fig. 1.b. As can be seen, rather than being fully connected, the CNN employs convolutions leading to local connections, where each local region of the input is connected to a neuron in the output. The input to a CNN is in the form of multiple arrays, such as a color image with three 2D arrays (length $\times$ width) in accordance to RGB or YCbCr channels. The number of channels is called depth and constitutes the 3rd D; note that more than three channels are not uncommon, e.g with hyperspectral images. A CNN is made up of Layers with each layer transforming an input 3D volume to an output 3D volume [11], typically, via four distinct operations [25], *viz.* convolution, a non-linear activation function (ReLU), sub-sampling (pooling) and classification (fully connected Layer). A simplified CNN is illustrated in Fig. 2[2]. A CNN can be described as several convolution layers with nonlinear activation functions (e.g. ReLU or sigmoid) applied to each layer. Each convolution layer applies several (may be thousands) distinct filters[3] (also called feature maps) and combines their results. These filters are automatically learnt during the training part based on the task in hand, e.g. if the task is image classification the learning concerns, a) detecting edges from raw pixels in the first layer, b) then use the edges to detect simple shapes in the second layer, c) followed by the use of these shapes in higher layers to detect higher-level features (such as facial shapes) and d) using the latter in the last layer for classification [26]. For further details on CNN, readers

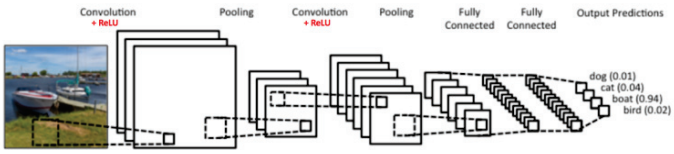


Figure 2: A Simplified CNN.

are encouraged to consult [7, 11, 25, 26, 18, 24, 27, 28].

2.2. Super-Resolution Basics

In fields - like astronomy, remote sensing, microscopy and tomography etc. - the acquired images may be handicapped by a variety of factors. These may include flaws or limitations of the measuring devices and "instability of the observed scene - object motion or media turbulence" [29]. The resultant images may suffer

[1] A claim contested by Schmidhuber [16] who traces it to as early as 1965.

[2] https://www.clarifai.com/technology

[3] The filter is applied in spatial domain using the usual sliding window approach, typical with digital images, and hence the name convolution.

from noise, blur and poor resolution. The remedy could be one or both of blind deconvolution (to remove blur) and super-resolution.

Super-resolution (SR) refers to an estimation of high resolution (HR) image/video from one or more low resolution (LR) observations [30] of the same scene, usually employing digital image processing and ML techniques. Being an inverse problem in most cases, there may be more than one solution, each requiring the construction of a forward observation model [31]. Probably, the first effort on the subject can be traced back to as early as 1984 [32]; the explicit use of the term 'super-resolution' was a bit later in 1990 [33]. In [30], a very detailed taxonomy is given but for the sake of brevity it is better to rely on the three tier classification given in [31] . The first tier classifies on the basis of both input and output, "as single input single output (SISO), multiple input single output (MISO) and multiple input multiple output (MIMO)". MIMO pertains to video SR and can be easily merged with the second one, which makes the first tier redundant. Hence it is better to directly classify according to the second tier, i.e. into two main categories, namely *single image* super-resolution (SISR) and multiple image or *multi-frame* super-resolution. From here onwards the rest of the Section 2.2 is almost entirely based on [31].

Given a single LR image of the scene, the SISR concerns the estimation the corresponding HR image, under the assumption that the original imaging set up is not available. Being an ill-posed problem, since several HR may correspond to the input LR image, SISR can be likened to ordinary "analytical" interpolation - like linear, bicubic, and cubic splines. The task may thus be to compute the missing pixel intensities in the HR grid as averages of known pixels, which may work well in smooth parts but wrought with dangers in case of discontinuities, in the form of edges and corners, that may lead to ringing and blurring artifacts. Hence more sophisticated insight, in addition to interpolation, is needed to super-resolve the input. There are two types of SISR algorithms:

1. **Learning methods** employ ML techniques estimate the HR details locally. These may be *pixel-based*, involving statistical learning [34, 35], or *patch-based* involving dictionary based LR to HR correspondence [36] of squared pixel blocks (called patches). The latter ones, also called *example-based* methods [37, 38], exploit internal similarities within the same image and may adopt various approaches, e.g. neighbor embedding [39], sparse coding [40], or a blend of these [41].
2. **Reconstruction methods** usually require explicit prior information (in the form of distribution, energy function etc.) while defining constraints for the target HR image. This may be carried out in a variety of ways, like sharpening of edge details [42], regularization [43] or deconvolution [44].

Some methods [45] may be termed as a melange of ML and reconstruction methods. Note that most "of the recent SISR methods fall into the example based methods which try to learn prior knowledge from LR and HR pairs, thus alleviating the ill-posedness of SISR. Representative methods include neighbor embedding regression [46, 47, 48], random forest [49, 50] and deep convolutional neural network(CNN) [4, 5, 51, 52]" [53].

In multi-frame SR, the input usually consists of more than one LR images, usually from slightly different perspectives of the scene in question. It is assumed that each input image is a degraded version of an underlying HR scene spoiled by blurring, down-sampling and affine transforms. For the latter case, integral shifts are considered trivial, carrying no useful information; ideally, fractional or sub-pixel shifts have greater information

value. There are three types of multi-frame methods:

1. **Interpolation methods** [54, 55] usually consisting of three steps, namely registration, interpolation and deblurring.
2. **Frequency-domain methods** [56, 57] gather disparate clues about high frequencies of the underlying HR from the DFT, DCT, DWT or any other frequency domain representation of the input LR frames. Due to their localized nature, DWT domain is better suited.
3. **Regularization methods** [58, 59] are useful in case of limited "number of LR images or ill-conditioned blur operators", and try to use either deterministic or stochastic "regularization strategy to incorporate some prior knowledge of the unknown HR image" [31].

For a detailed treatment on the subject, the readers are recommended to consult [30, 60, 31, 61, 62, 63, 29].

3. Deep Networks for Image Super-resolution

According to [64], while employing a CNN for restoration tasks (like SR and denoising), pooling or subsampling may be counter-productive as important image details may be discarded. Hence pooling layers are usually avoided in SR tasks which again has its downside; each additional convolutional layer means a new weight layer and hence more parameters with the consequences in the form of overfitting and too huge a model to store/retrieve.

3.1. Image Benchmarks

3.1.1. Image databases

In Table 1, we list many of the image datasets that are popular with the SR community. Some of the datasets have already been partitioned, by their proponents, to training/validation and testing sets; there is, however, no hard or soft rule and many works use them without restricting to these partitions. Sometimes the researchers improvise on the datasets, e.g. in [5], the 91 Timofte dataset is decomposed into 24,800 sub-images for training along with 395,909 images (over 5 million sub-images) from ImageNet. Similarly some authors combine more than one datasets, e.g. training 291 (91 Timofte +200BSD) from [49] is a popular choice for training [52, 88].

3.1.2. Non-CNN super-resolution methods for benchmarking

An attempt [89] at benchmarking the state of the art SISR methods, focusing mainly on [38, 90, 91, 40, 92, 47, 93] in addition to bicubic interpolation, stops short of giving any verdict because of the mercurial performance of the methods on BSD200 and LIVE1 dataset [94]. For obvious reasons, no CNN based method is there; indirect comparison may still be possible as most of the included references have also been employed for the benchmarking of the CNN based methods appeared since. In this section we attempt to describe the non-CNN methods mainly used for benchmarking emerging methods. The major CNN based benchmarking methods are presented anyhow in the next sections.

Following are some of the principal non-CNN methods favored for contemporary benchmarking in the image super-resolution literature:

Table 1: Publicly Available Image datasets.

S/No.	Name with reference	Details
1	ImageNet [65]	The detection dataset by ILSVRC, which consists of around 400,000 images.
2	Timofte dataset [48]	Widely used by SISR researchers, it consists of "91 training images [40] and two test datasets; Set5 and Set14 with 5 and 14 images" [66], respectively.
3	Berkeley segmentation dataset [67]	BSD300 and BSD500 - 200 images for testing (B[SD]200), the rest for training and validation.
4	CIFAR-10 [68]	After the Canadian Institute for Advanced Research - 6000 images in each of 10 classes of which 5000 are in training set and the rest in test set. CIFAR-100: 600 images in each of 100 classes.
5	L20 [69]	Has very large images, between 3m pixels to up to 29m pixels, while the other datasets have images below 0.5m pixels [70].
6	The General-100 dataset [71]	Contains 100 uncompressed bmp-format images, since the proponents believe the BSD500 are not optimal for SR tasks, due to JPEG format. The image sizes range from 710x704 (large) to 131x112 (small). "They are all of good quality with clear edges but fewer smooth regions (e.g., sky and ocean), thus are very suitable for the SR training." [72]
7	Urban 100 [73]	having 100 HR images diverse in real-world structures; famous for its self-similarities.
8	The Kodak PhotoCD dataset [74]	Consists of 24 lossless true color images without compression artifacts, and is used as a standard testing set for many image processing works [50].
9	The super texture dataset [75]	Provides 136 texture images.
10	291 from [49]	A combination of No. 2 above and BSD200
11	MNIST [76]	Modified National Institute of Standards and Technology database: a large database of hand-written digits.
	- Binary version [77]	
	- MNIST corners dataset by [78]	Constructed by randomly placing an MNIST digit in either the top-left or bottom-right corner
12	CelebA [79]	Centrally cropped faces
13	LSUN Bedrooms [80]	Bedroom images
14	The van Hateren dataset [81]	4167 gray scale images of 12 bit depth (mostly nature or buildings) with 1536 X 1024 pixels each and a gray scale.
15	MS-COCO [82]	91 easily recognizable objects types - 2.5 million labeled instances in 328k images.
16	YFCC100M [83]	The Yahoo Flickr Creative Commons 100 Million Dataset containing "100 million media objects"' (around "99.2 million photos and 0.8 million videos") under a Creative Commons license.
17	LIVE [84]	A database variously distorted images with accompanied subjective assessments from human observers. The images were originally acquired for a project on generic shape matching and recognition.
18	Manga109 [85]	A publicly available dataset of 109 Japanese comic books with numerous comic sketches.
19	Open Images Dataset [86]	About 9 million annotated (labels dealing 6000 categories) HD images.
20	DIV2K [87]	The DIVerse 2K resolution image dataset is a new addition and served as a benchmark for NTIRE 2017 Challenge. It is a high quality (2K resolution) set partitioned to 800 training, 100 validation and 100 test images.

1. **Bicubic interpolation:** one of the widely used classical interpolation methods; others being nearest neighbor and bilinear.
2. **NE+ [46]:** a set of neighbor embedding methods that selects several LR candidate patches in the dictionary by using a nearest neighbor search and employs their HR version for the reconstruction of HR output patches; may be via least squares (NE+LS) or local linear embedding (NE+LLE) or even Non-Negative Least Squares (NE+NNLS [95]).
3. **SC or SrSC [40]:** approximates the input LR patch via sparse representation and then applies the resultant coefficients to sparsely generate the corresponding HR output patch.
4. **KK [91]:** directly learns a LR to HR mapping based on kernel ridge regression (KRR) using a sparse approach that combines kernel matching pursuit with gradient descent.
5. **K-SVD [96]:** refers to a combination of K-SVD (from [97]) and Orthogonal Matching Pursuit (OMP) for efficient dictionary learning in order to improve upon the sparse method SC.
6. **A+ [48], ANR and GR [47]:** Anchored Neighborhood Regression (ANR) is an effort to improve upon K-SVD and SC by introducing a ridge regression (solvable offline and storable per dictionary atom/anchor). A less accurate but more efficient variation employs global regression; hence the name GR. A+ (advanced ANR) is a later improvement that, unlike ANR, does not solely learn from dictionary atoms, but involves all the training patches in the local anchor neighborhood. Although having similar time complexities, A+ has been shown by the authors outperforming ANR and GR.
7. **Self-Ex [73]:** self-similarity algorithm that includes rubber-band transformations, after estimating the expected "deformation of recurring patches", in order to expand the internal "patch search space". The method outperforms state of the art methods, especially A+, on a synthetic database developed by the authors.
8. **SRF [49]:** (super-resolution forests) relies on direct mapping from LR to HR patches using random forests (RFs). The authors demonstrate a relation of contemporary SISR to locally linear regression and try to fit in RFs into this framework. The method has many variants in the form of RF linear (RFL), RFL+ and its advanced version ARFL+ (also called ASRF).
9. **NBRSF [50]:** employs a *Naïve-Bayes SR forest* with bimodal trees for example-based SR using a hierarchical external learning strategy, that provides a fast local linearization search, followed by a fast "Local Naïve-Bayes strategy for per-patch tree selection".
10. **IA [69]:** or improved A+ is the result of a generic 7-way strategy proposed by its authors for the amelioration of any given SR method.

The last three methods are important as they are part of a counter-narrative against CNN strategies and one can find SRCNN in their comparisons.

3.2. State of the Art Methods on Image SR

The SISR method [5], called SRCNN[4], is illustrated in Fig. 3. The said work is an extension of an earlier work [4][5] after the introduction of larger filter sizes and additional mapping layers. The authors rely on learning directly an end-to-end mapping, between the input LR image and the corresponding HR output image, that is represented as a deep CNN. The method uses a bicubic interpolation as its pre-processing step followed by the extraction of overlapping patches, via convolution, as high dimensional vectors with as many feature maps as their dimensions. The vectors are then non-linearly mapped to each other and subsequently aggregated in the form of patches to get the reconstructed HR image that is supposed to be as close to the ground truth as possible. The authors boast results comparable to ANR [47] and "somehow to" KK [91]. However, it is reported in [73] that on Urban100 dataset (and even with BSD100), SRCNN is outperformed by methods like Self-Ex [73], A+ [48] and KK, by a fair margin. The authors of [5] view CNN as an extension to sparse-coding-based SR

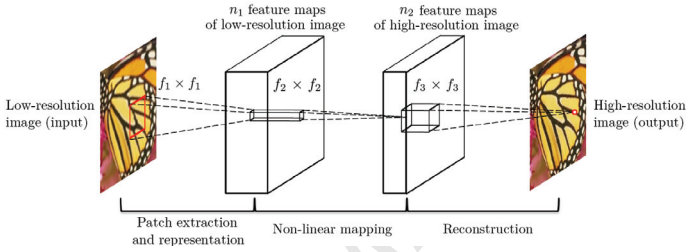


Figure 3: SRCNN [5].

methods [40] but for the fact that the former jointly optimizes all layers unlike the segregation of components in the latter.

SRCNN "has only convolutional layers which has the advantage that the input images can be of any size and the algorithm is not patch-based." [98]. Although SRCNN claims efficiency in view of what the authors call a lightweight structure, it's still a far cry. With that being said, an effort to improve the efficiency is in the offing in the form of fast SRCNN (FSRCNN) [71] but a look at the preliminary version has a lot to answer especially about the quantification of speed improvements. In addition, the proposed FSRCNN[6] improves upon the original in terms of PSNR. For the sake of efficiency, the FSRCNN [71] proposal replaces pre-processing bicubic interpolation step of SRCNN by a post-processing (fifth) step in the form of deconvolution. Other than that, the pipeline has four convolution layers in the form of feature extraction, shrinking, mapping and expanding. That is, the mapping is preceded by the shrinking feature dimensions and followed by expanding back. Moreover, the filter sizes are proposed to be reduced again but with the introduction of additional mapping layers. The author claims 40 times improvement which is debatable in the absence of theoretical analysis. For the sake of efficiency, the work in [99] follows FRCNN in carrying out the upsampling as a post processing step that employs the network in network (NIN) approach proposed in [100]. But only in its compact form, the method produce comparable results to FRCNN. The mechanism to compute the theoretical complexity is also not clearly defined.

[4] http://mmlab.ie.cuhk.edu.hk/projects/SRCNN.html
[5] In literature, SRCNN refers to [5]- the earlier version [4] is referred to as SRCNN-Ex by its authors.
[6] http://mmlab.ie.cuhk.edu.hk/projects/FSRCNN.html

Another deep learning method, the deep network cascade (DNC [101]), gradually upscales the input LR, in a layer-wise manner with each layer contributing by a small scale factor, to eventually get the HR image. The successive refinements are based on searching non-local self-similarities in order to ameliorate the high frequency details of the patches to which the image is partitioned in a way similar to the internal example-based method [38], as pointed out in [102]. The patches are then fed to each layer of a cascading multi-stacked network of collaborative local auto-encoder (CLA) for noise suppression and collaborating the conformance among overlapping patches enhanced by the aforementioned process. The stacked CLA (SCLA) is a concatenation to incorporate multiple models into the cascade for better SR. The reported results demonstrate the superiority of DNC in comparison of various state of the art methods on sparse coding. The reported set of images for experimentation is small, however; very few images[7].

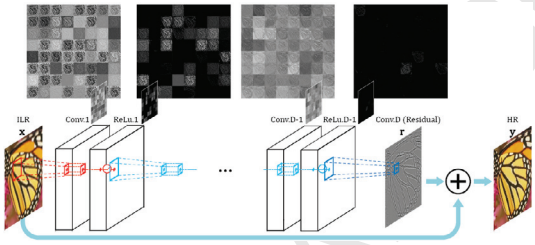


Figure 4: VDSR [52].

For obvious reason the deeper the CNN, more accurate should be results. The VDSR[8] [52] method[9], illustrated in Fig. 4 uses a very deep convolution network inspired by VGG-net used for ImageNet classification [103]. Using a depth of 20 weight layers in a cascaded deep network, involving small filters multiple times, the authors report efficiency in exploiting context information over large image regions. The slow convergence issue, related to high depth, is tackled by using very high learning rates; the downside is gradient exploding which is mitigated by learning residuals only and adaptive gradient clipping. The work is also extended to multiscale SR using a single network. The authors choose the very first figure in their article to demonstrate the superiority of VDSR par rapport the state of the art, both in terms of efficiency and visual quality; especially SRCNN which is shown to be outperformed by 0.87dB [104]. Moreover, further reported results show that VDSR outperforms five state of the art reference methods, with SRCNN finishing second in terms of PSNR and A+ finishing second in terms of time efficiency.

Deeper networks may lead to high accuracies but can bring two issues to the fore, *viz.* overfitting and huge model. With that in view, Kim *et al.* [64] propose what they call deeply-recursive convolutional network (DRCN), to apply the same convolutional layer recursively as many as 16 times. The idea is to avoid introducing additional parameters while having an increased depth. The authors deal with training problems, especially the exploding/vanishing gradients, by adopting the recursive-supervision strategy of [105] and skipping of layers

[7]The authors do refer http://vipl.ict.ac.cn/paperpage/DNC/ for detailed results but the link had been inaccessible, at least at the time of writing this article.

[8]may stand for Very Deep Super-Resolution, as the authors do not mention it

[9]http://cv.snu.ac.kr/research/VDSR/

from [106]. The authors use the same simulation environment as they have used for VDSR [52]. Although they have not included VDSR as a reference, a comparison is still possible as the datasets and the reference methods are the same. The reported results are comparable to VDSR, and even outperforms the latter in case of Set5, as far as final image quality is concerned. The authors, however, ignore the time efficiency which was important in the context of the article in hand.

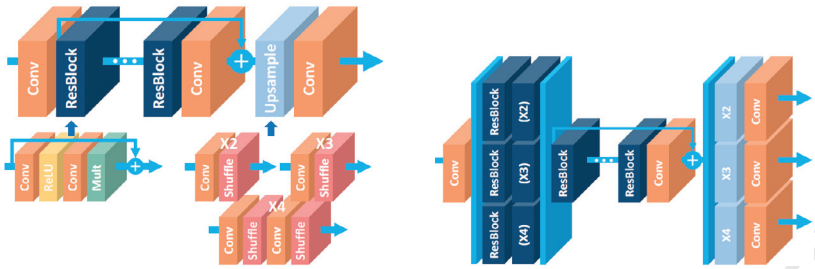
The authors in [107] argue that despite the emerging popularity of data driven approaches, the sparse coding paradigm has not lost its value for its domain knowledge and can be potentially very beneficial if combined with deep learning, especially if embedded in a cascaded structure. This, they say, may not only lead to efficiency and better training but also a reduced model size. Their method, the sparse coding based network (SCN), exploits NN approximation of sparse coding in the form of the learned iterative shrinkage and thresholding algorithm (LISTA) proposed in [108]. SCN - later extended to [109] - is claimed to have been compact, accurate and imperceptible in relation to SRCNN. The authors also propose a cascaded version (CSCN) of their method, that employs multiple SCNs, with better artifact reduction and scaling flexibility. The reported results seem better than those by SRCNN and A+, in terms of PSNR, SSIM, subjective perception and time efficiency. However, the time efficiency results are empirical and no theoretical analysis is carried out.

In another yet to be formally published work [110] , the authors apply a number of SR methods to the LR image, independently, to get various HR estimates. The latter are then combined, on the basis of adaptive weights, to get the final result. The authors refer to their method as MSCN-n, where n is the number of employed inference modules. The reported results demonstrate superiority of MSCN to the state of the art methods - like A+, SCN and SRCNN - as far as image quality is concerned. The method is also shown to be efficient, with a comparable performance to [107].

Another wok that exploits various image priors during the training phase of a deep CNN [51] is called SCRNN-Pr. One aspect of prior information focuses edge/texture restoration and the other concentrates on gradual upscaling via parallel structure recurrence. The authors claim efficient training speed along with better image quality par rapport the state of the art. Strangely enough, despite the claim of SRCNN-Pr, the same first author has not even alluded to SRCNN-Pr as a reference method, in his later related works [53, 111]; all the other state of the art method, he relied on for comparison, are there.

A yet to be formally published work, 'Zero Shot' super-resolution (ZSSR [112]), train its CNN on recurring sub-images within the training/test image at run-time with an idea of auto-adjusting to various image settings; the principal target being poor-resolution old photos with unknown degradations. The article needs to be convincing in its approach in its final form and this whole idea of 'run-time' CNN for each single image seems over the top. The reported results are comparable to the state of the art.

The authors of [114] employed ResNets [115] for SR to propose what they call SRResNet and obtained improved results as far as the time/memory duo is concerned. Following the deblurring method in [116], Lim et al. [113] eliminate batch normalization (BN) from successive layers in SRResNet to propose what they call enhanced deep super-resolution (EDSR) and multiscale deep super-resolution (MDSR) networks, as illustrated in Fig. 5. While expecting about 40% lesser memory usage by undoing BN, the authors try to invest this saving in building larger model in terms of layers. The EDSR has a depth of 32 layers, 256 filters (feature channels),



(a) EDSR (b) MDSR

Figure 5: Methods from [113].

43M parameters, residual scaling factor of 0.1 and does not have any ReLU activation layer outside. The latter is also true for MDSR which is inspired by VDSR's inter-scale correlation. MDSR treats SR at multiple scales and has a depth of 80 with 64 filters and 8M parameters. The work is the winner of NTIRE2017 challenge, for which the authors also supplemented their methods with the self-ensemble of [69] to get what they call EDSR+ and MDSR+. The reported results of the proposed methods outdo the benchmarking methods - like SRCNN, A+, VDSR and SSResNet - by a margin, with EDSR+ being the best.

Another ResNet-inspired SISR work [88], which was later extended in [53] to include some more related reference methods, employs skip connections to counter exploding/vanishing of gradients and a parameter economic CNN (width, depth and skip connections) for faster training. In essence they propose two schemes, namely R-basic and R-deep, with the former having 22 convolutional layers with 0.3M parameters while the latter having 34 convolutional layers with 5M parameters. They have chosen half a dozen reference method but their main focus is on comparison with a 20-layer VDSR, involving 0.7M parameters, which is showing better results than other reference and R-basic methods but is outperformed by R-Deep. Both R-basic and R-deep are demonstrated to converge faster than VDSR, in terms of the number of epochs; without highlighting the per epoch time complexity, however.

With robustness and efficiency in perspective, the Deep Projection CNN (DPN) method in [111] employs what the authors call model adaptation to exploit the repetitive structures in the LR image. The DPN has three parts: feature extraction/representation via stacked CNN/ReLU layers, inference via CNN/ReLU based projections, and CNN only HR reconstruction. The authors claim up to 0.3dB PSNR gain with a 40-layer DPN (and 0.7M parameters) over reference methods, like A+, SRCNN and VDSR (same 20 layers and 0.7M parameters, as in [53]). The reported results are almost the same as reported [53] and the latter, despite being from the same first author, does not appear on the reference list. In addition, no time efficiency results are reported.

The pixel recursive super-resolution network in [78] is composed of a conditioning network and a prior network. Whilst the former is a CNN that converts the input LR image to logits to predict the log-likelihood of each HR pixel, the latter is a PixelCNN [117]. The probability distribution is represented as a softmax operator applied to the sum of the output logits of the two networks. The model thus "synthesizes realistic details into images while enhancing their resolution". The PixelCNN is by itself a CNN variant of Pixel Recurrent Neural Network (PixelRNN) [118]. The authors claim success in the form of subjective quality perception of the resultant images but the method is outperformed by one of their reference methods [115] in terms of quantitative measures, like

PSNR. Strangely enough, the method is not compared with any CNN based method.

In. [119], the authors propose a model named deep joint super resolution (DJSR) in order "to adapt deep model for joint similarities". The authors argue that the CNN "model has a clear analogy to classical sparse coding methods", like [40]. While using the Stacked Denoising Convolutional Auto-Encoder (SDCAE[10]), reported in [120], the method takes randomly corrupted LR images as input to output HR images by combining the auto-encoders and CNNs (even SRCNN may be possible). SDCAE pre-trained on external examples followed by refinement with reliable multi-scale self examples. The reported results demonstrate comparable performance to SRCNN and better than A+, DNC etc.

The authors in [121] term their method very deep Residual Encoder-Decoder Network - "RED-Net" for short - that consists of a series of convolutional and subsequent deconvolutional layers in order to learn the overall mappings from input images to the ground truths. Whilst the convolution being the feature extractor while eliminating the noise, deconvolution aims at recovering the details. The authors propose to introduce "skip connections" between each convolution layer and the corresponding deconvolution layer for the back-propagation of gradients to lower layers and passing image details to upper layers. The authors claim "setting new records" on image denoising and super-resolution, which is somehow exaggerated. The method is claimed to be compared with half a dozen methods, including SRCNN. The results are comparable as only a slight improvement is reported on the test cases in hand; there is no mention of time efficiency. It can be deduced from the results that increasing the number of layers improve the results slightly but the efficiency part is still elusive and does not seem cost effective. Moreover, the method consists of n convolutional layers followed by n deconvolutional layers and there is a skip connection between every kth convolution layer and $(n-k+1)$th deconvolution layer, for $k = 1, 2, \ldots, n$. In other words, the first convolution layer must wait the last deconvolution layer for the sake of correct correspondence. For such a deep network, this may be an additional overhead.

A method presented in [122], to treat partially pixelated images for super-resolution, is named Depixelated Super Resolution Convolutional Neural Network (DSRCNN). It consists of an autoencoder inspired by [123] combined with two depixelate layers (de-noising and encoding) via deconvolution as illustrated in [124]. The autoencoder is composed of a generator and a discriminator. The generator is modeled on SRCNN-Ex [4] and consists of de-noising, encoding and decoding layers. Application of the method, on randomly pixelated images, shows results comparable to SRCNN but the reported sample size is too small. Moreover the work suffers from clarity of presentation, especially on the placement of depixelate layers; are these part of autoencoder or outside?

A yet to be formally published work [125] argues against treating all pixels equally and call for taking into account the salient structures in the form of local and holistic contents. With that in view, a "local structure-preserving sub-network (LSP)", followed by a "holistic structure-preserving sub-network (HSP)", are proposed to be incorporated into the fully-convolutional learning. By using deconvolution, the LSP upsamples the low resolution patches which are thereafter refined through convolution by HSP. The authors claim superior results in comparison to state of the art methods especially the SRCNN. The paper, in its present form however, suffers from the effective presentation of the results. The authors criticize SRCNN for its equal treatment of pixels while

[10] An implementation: https://github.com/ifp-uiuc/anna

ignoring the texture contrast and its use of bicubic interpolation in its pre-processing step for it may have adverse effects on the main structure if not initialized properly.

For single text-image Super-Resolution, the authors in [126] compare two methods after essential tweaking, namely the Multi-Layer Perceptron (MLP) method reported in [127] and one of the pioneering works on CNN or ConvNets [17]. Their example-based strategy attempt to learn a non-linear mapping between pairs of text patches and high-frequency coefficients. According to their results, the CNN based method outperformed the other and a few other contemporary methods they used for comparison. Later in the article, however, the authors extend their experiments to SRCNN, in which case the latter fared to be superior; still their basic conclusion remains unaltered, i.e. CNN based methods fare far better. The work in [128] carries out SR without any prior information on the blurring kernel (hence the word "Blind"in the title) by using CNNs. The authors do some limited simulations on Set5 and Set14 in comparison to SRCNN, A+ and SRF [49]. The claimed results are above par not only in a blind set-up but also with the same reference methods in a non-blind environment; one exception is the superior performance of SRF on Set5.

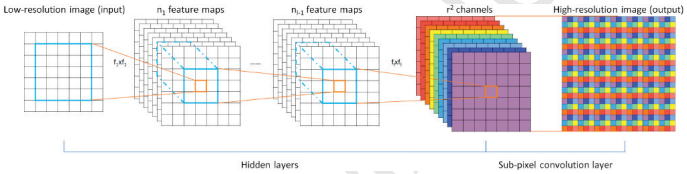


Figure 6: ESPCN [66].

For the sake of economizing on space and execution time, the method in [66][11] first downsamples the HR image, after Gaussian filtering via convolution, to LR and then extracts the feature maps from the latter. Then each layer, excluding the last sub-pixel convolution layer, is characterized by its own upscaling filter for the feature map of the concerned layer. The learning of these filters is, however, the job of the last "sub-pixel convolution layer" which upscales the low resolution image to a super-resolved image. The authors call their method the "efficient sub-pixel convolutional neural network (ESPCN)" which is illustrated in Fig. 6. Note that, for a total of L layers, the preprocessing bicubic filter of SRCNN has been replaced with $L-1$ upscaling filters, each trained for every feature map. The authors report results that are at least better from state of the art methods (read SRCNN) by +0.15dB with images. As of efficiency, the authors report a run-time average performance of 29.5dB over less than 0.2 seconds in a specific environment as against 29.0dB over more than 1 second for SRCNN, while using the set14 dataset.

With ESPCN [66], two issues come to fore. One, why an available HR image should be reduced to its low resolution counterpart? One may agree that for testing and learning purposes, the HR images are kept as references (ground truths) under the assumption that they are unknown to the system. But what the author suggests is on the contrary. If the aim is just to reduce the size, why not partition the HR image to small tiles, feed it, in lieu of low resolution image, to the pipeline to get the super-resolved tile and then stitch or tessellate [129] the output tiles? The second issue is the claim of best real-time performance on the basis of a few

[11] https://github.com/Tetrachrome/subpixel

datasets. While the reported results are commendable, nothing can substitute theoretical complexity analysis, which is missing from the article. In a later explanation [130], the authors do try to address the first issue, while quoting the works in [21, 131][12]. It must be borne in mind that the quoted works are mainly concerned with classification problems, as against image reconstruction. What is hard to understand that deconvolution is an inverse problem and how can it be equivalent with convolution.

A yet to be formally published deep CNN based SISR method, the Gradual Upsampling Network (GUN [132]), utilizes a gradual process to magnify from LR to HR, as against the two extremes in SRCNN and ESPCN. The architecture comprises of an input ReLU based convolution layer, followed by a series of alternating upsampling and convolutional layers, and an output layer. The authors believe that their gradual upsampling strategy of adopting very small magnification factor is cost effective in terms of efficiency but no empirical or theoretical evidence is provided. Their training strategy is also gradual and trains an initial network with edge-like samples, followed by gradual tuning with more complex samples. The authors report results that are superior to about nine state of the art methods with VDSR and SRCNN finishing second and third, respectively.

Ren *et al.* [133] argue that the operators in CNNs and sparse auto-encoders are translation invariant and are therefore not suited for scenarios requiring translation variant interpolation (TVI). By employing the Shepard interpolation framework [134], they propose what they call Shepard Convolutional Neural Networks (ShCNN)[13] in order to introduce and train "end-to-end TVI operators in the network", efficiently, while accentuating the new Shepard layers by a few feature maps. The reported super-resolution results, with reference to a few methods including SRCNN and A+, demonstrated superiority of the method. The authors claim 'competitive' running time but no supporting evidence is given, although they mention casually to prefer introducing a few Shepard layers rather than going for a deeper architecture, for efficiency. Secondly, the PSNR results on Sect14 are under-reported for SRCNN.

By employing the Gibbs model as the conditional probability distribution, the authors of [135] initialize their CNN with filters having good geometric properties - e.g. multiscale complex wavelets based scattering networks [136] or VGG networks [137][14]. They also put forward a time-costly fine-tuning algorithm via gradient estimation of the conditional log-likelihood. The CNN architecture is inspired SRCNN. For training simulations, the authors have randomly selected 64x64 image patches from a subset of the training set of ImageNet [138]. The ensued results seem hardly enviable with respect to both PSNR and time complexity.

The authors of [139] argue against the use of bicubic upsampling as a first step and instead propose to go first for feature extraction, in order to map the LR image into a deep feature space, followed by a learning based upsampling of such features to the desired dimensions with learned filters . For the HR reconstruction, context information is derived from the upsampled features, in a multi-scale way that incorporates both short- and long-range contextual information at the same time. By taking a cue from [115], the authors introduce a shallow network as part of their architecture, in order to facilitate training, for the sake of efficiency and

[12] http://caffe.berkeleyvision.org/doxygen/classcaffe_1_1DeconvolutionLayer.html

[13] http://www.deeplearning.cc/shepardcnn and https://github.com/jimmy-ren/vcnn_double-bladed/tree/master/applications/Shepard_CNN

[14] http://www.robots.ox.ac.uk/∼vgg/research/deeptex/

faster convergence. The method is compared with more than ten other methods - including SRCNN, ARFL [49], NBSRF [50] and CSCN (strangely, not with CSCN-MV from the same work [107]) - and is reported to be superior, with CSN being second. The authors also show gradual increase in quality with increase in kernel size for single scale reconstruction, but still the multiscale version outperforms all. Time complexity is given in the form the number of iterations; the EEDS convergence is of the order of 10^5 iterations, for even a small dataset like set5. It would have been interesting, however, had the EEDS been compared with other methods in this context. The multiple scale upsampling in also relied upon in [140] wherein an adaptive selection of the optimal scale of information is realized via a "competition" among various multi-scale convolutional filters. The reported results are comparable to SRCNN with slightly better PSNR and convergence rate at the expense memory cost. The reported results of benchmark methods are not consistent with the literature; may be due to some pre-processing?

The "shallow and deep convolutional networks for image super-resolution (SDSR)" [141], as the name suggests, consists of one shallow and one deep channel; both containing a deconvolution based upsampling module, like FRSCNN, for efficiency. Whilst the former attempts to restores the general outline, the latter extracts, in a multi-scale way, both the short- and long-scale texture information. The reconstruction is based on a multi-scale approach in the deep channel wherein high and low frequencies are extracted simultaneously. The claimed results show marginal improvements in image quality but CNN based methods are mainly ignored in the benchmarking, especially EDSR/MDSR. In addition, the convergence of the various sized versions of SDSR is almost in the same number of iterations with marginal improvement in PSNR.

Two inter-related works [142, 143], with the former being a specific case of the latter with $k = 2$, employ a $k \times k$ -channel CNN with the output pixels, having the same coordinates, grouped together. The result is a magnified image, with an scaling factor of k. The method does not involve any bicubic interpolation. The authors report average PSNR gain of about 0.2dB par rapport the SRCNN across all the upscaling factors of 2,3 and 4. In [142], they claim an improvement of 0.39 dB (the average PSNR 36.88 dB) over SRCNN for an upscaling factor of 2 but none of SRCNN works [5, 4] have reported any such average PSNR (36.49dB).

In [144], the LR image is input to a scattering convolution network [145, 136] to get scattering maps from which sparse codes are extracted which serve as input to a CNN. The method is referred to by the authors as hybrid wavelet convolution network (HWCN) as it is a mélange of both sparse coding and CNN. Given a tiny dataset, HWCN could train complex deep network with better generalization by regularization from scattering convolutions, and thereby is a competitive alternative to CNNs. Although the reported results outperform sparse-coding methods by a small margin but no comparison is given par rapport the CNN methods. Our own reading of the comparison with SRCNN, suggests that the latter outperforms it. As of time complexity of convergence, only comparison with the bicubic interpolation is given. The convergence time is reported to be in the order 1000 iterations, but the convergence to a PSNR of around 34dB seems a bit dubious.

The multi-channel-input SR convolutional neural network (MC-SRCNN [146]) takes LR and its various enhanced interpolated versions (hence the term multichannel) as inputs to extend the SRCNN. With this the expectation is to extract better features for HR restoration as the interpolated versions of LR may have complementary information to that of SRCNN exclusive architecture. With an upscaling factor of 3 and 18-channel MC-SRCNN, the authors report an improvement of 0.34dB and 0.18dB in PSNR on set5 and set14, respectively.

In [147], both CNN and sparse coding are employed whereby the input LR image is subjected to an efficient convolutional sparse coding module followed by 'perforated' upsampling and convolutional decoding. The 'perforated' upsampling may bring more sparsity and somehow approximates inverse max-pooling operator of deep CNNs. For comparison, the upsampling part is improvised, to give various reference methods, with classical interpolation methods, like bilinear, bicubic, nearest neighbor and linear shift. The authors call such variations as state of the art methods, which is a bit naïve. Anyhow a PSNR gain of about 0.8dB, over simple bicubic interpolation, may not be helpful in establishing the effectiveness of the method. According to [66], such an approach of increasing image resolution, in the middle of the network gradually, may escalate the computational complexity which may be especially problematic with CNNs where the processing speed is a function of the input image resolution.

In [148], the authors improvise some contemporary state of the art super-resolution methods [49, 4, 48, 47] using conditioned regression models in order to exploit supplementary kernel information during training and inference. The idea is to have a single training model, rather than repeating it for every candidate blur kernel, especially if the latter is different for each image. In the proposed "Regression-conditioned" SRCNN, the first convolutional layer is replaced with parametrized convolution in the form of a non-linear function derived via an additional neural network trained jointly with SRCNN. In the ensued results, however, the conditioned SR forests outperform conditioned SRCNN by a small margin.

As a part of their work in [149], the authors train their CNN with semantics in the form of feature reconstruction loss rather than the per pixel loss of SRCNN. In their results, the subjective quality is reported to be far superior to SRCNN, but quantitative metrics do not support the claim. The authors attribute it to "the feature reconstruction loss" that leads "to a slight cross-hatch pattern visible under magnification, which harms its PSNR and SSIM". A SISR method [150] employs CNN in combination with regularization constraints involving both the local and non-local image similarities. The authors claim a state-of-the-art resolution quality. The work in [151] takes into account both the local intensity and local gradient to reduce edge blurring while training the SRCNN model.

The article on SRResNet also proposes the SR generative adversarial network (SRGAN [114][15]). The authors claim photo-realistic natural images with $\times 4$ magnification after proposing a perceptual loss function constituted by both an adversarial loss (uses a discriminator network to differentiate photorealistic images from SR images) and a content loss that depends on perceptual similarity rather than pixel similarity. The reported PSNRs (27.02dB on Set14 which is even less than bicubic interpolation) are not that encouraging but the perceptual quality seems enviable.

The work in [152] deals with the "problem of efficient training of convolutional deep belief networks by learning the weights in the frequency domain" in order to avoid the time-expensive convolutions. The authors claim about

[15] Links to implementations:

https://github.com/leehomyc/Photo-Realistic-Super-Resoluton

https://github.com/junhocho/SRGAN

https://github.com/titu1994/Super-Resolution-using-Generative-Adversarial-Networks

×8 efficiency on 2D images and ×200 on 3D volumes from medical data.

Another deep learning method, the Laplacian Pyramid Super-Resolution Network (LapSRN [153]), attempts to progressively reconstruct the sub-band residuals of HR images. At each pyramid level, the input coarse-resolution feature maps which are used to predict the high-frequency residuals and subsequent transposed convolutions for upsampling to the finer level. The authors claim both efficiency and accuracy par rapport the state of the art methods, especially SRCNN.

The authors in [154] propose to to train multiple CNNs (read SRCNN), each with a different network, and integrate their outputs via additional convolution layers; an approach they call Context-wise Network Fusion (CNF). Obviously, the complexity may escalate by the introduction of multiple networks in parallel plus the later one needed for integration. And the gain is not that much, as revealed by the ensued results.

3.3. Set14-Based Benchmarking

In Table 2, based on the application of a given method to Set14 dataset, we have collected the data from the literature on PSNR of final super-resolved image with an upscaling factor of 3. The reason to select Set14 and upscaling factor of 3 is the popularity of this benchmark with most the presentations. We tried to standardize the empirical time complexity but every work has used its own experimental settings and objective analysis may not be possible in the absence of theoretical complexity analysis. We will not pass any verdict based on a small dataset for a particular scaling factor and leave it to the reader. However the winner [113] of NTIRE2017 challenge [87] seems to have the reasons to claim the superiority of his methods with PSNRs of 30.52dB (EDSR) and 30.44dB (MDSR); the improvement is glaring par rapport the rest of the methods. They have improved it further to 30.66dB (EDSR+) and 30.54dB (MDSR+).

Note that we have not included SRResNet [114] and GUN [132], in the table, for different reasons. While SRResNet has claimed better results than the state of the art, a fact later endorsed by the winner [113] of NTIRE2017 challenge, the article never reported the results for Set14 with ×3; with ×4, it does outdo methods like VDSR, SRCNN and A+. As of GUN, the authors have re-executed all their benchmarking methods by applying these to Y channel and doing simple bicubic interpolation to the CbCr components, as against resorting to RGB[16]. Their PSNR results on Set14 at ×3 are therefore on the high side, i.e. in dBs; 33.35 (GUN), 33.07 (VDSR), 32.93 (SRCNN), 32.39(A+) and even 30.36 (bicubic). The authors of DJSR [119] report 0.3dB higher PSNR for SRCNN which may be a typo.

3.4. Importance of SRCNN

All the comparisons notwithstanding, SRCNN is still a landmark work on SR. Not only being a quintessential benchmarking method, it has also been employed in many application scenarios. In a work on face recognition [155], the authors got improved recognition rates when the input face images were subjected to SRCNN-Ex [4] before the use of HMM and SVD. In [156], SRCNN is successfully employed to improve the quality of

[16] In a background communication, the first author clarified that they have simply excluded the "Bridge" image from Set14; the reason being its 6-bit depth which had been zero-padded to 8 bits. The authors think that it is against reason to include a six bit image in an 8-bit SR method.

Table 2: Benchmarking over Set14 with ×3 upscaling.

S/No.	Name with reference	PSNR (dB)
1	Bicubic	27.54
2	SC [40]	28.31
3	K-SVD [96]	28.67
4	ANR [47]	28.65
5	A+ [48]	29.13
6	NE+LLE [46]	28.60
7	KK [91]	28.94
8	SRCNN-Ex [4]	29.00
9	SRCNN [5]	29.30
10	FSRCNN [71]	29.43
11	DPN [111]	29.80
12	ShCNN [133]	29.39
13	Self-Ex [73]	29.16
14	VDSR [52]	29.77
15	DRCN [64]	29.76
16	HWCN [144]	29.17
17	SCN [107]	29.41
18	CSCN [109]	29.55
19	MSCN-4 [110]	29.65
20	RED30 [121]	29.61
21	DSRCNN [122]	28.60
22	NBSRF [50]	29.25
23	R-basic [53]	29.67
24	R-deep [53]	29.80
25	RFL [49]	29.05
26	ARFL [49]	29.13
27	RFL+ [49]	29.17
28	ARFL+ [49]	29.23
29	ESPCN [66]	29.49
30	IA [69]	29.69
31	ZSSR [112]	29.80
32	DJSR [119]	29.96
33	EDSR [113]	30.52
34	MDSR [113]	30.44

infrared thermography (IRT) images for object recognition. The work in [157] exploits SRCNN to enhance chest CT images from the Cancer Imaging Archive[17]. In [158], the authors compare the performance of SRCNN with another SR method [159] in improving the resolution of lensless blood cell counting images; SRCNN is reported to have 9.5% better results. The method in [160] can process sketch images of any resolution and employs CNNs to auto-clean rough raster sketch drawings. The authors of [161] apply CNN to Light-Field (LF) images - hence the name LFCNN - to upsample both the angular and spatial resolutions.

4. Video Super-resolution

Many of the methods described in the previous section are also applicable to videos but due to the inherent peculiarities, especially those related to the various compression algorithms, a separate section dealing with SR may not be out of place here.

4.1. Video Benchmarks

Table 3 lists some of the popular video datasets that are publicly available. Note that still image datasets are still valid and are mainly employed to check the quality of single frames.

Not only the video specific methods, but also the popular image SR methods described in Section 3 - especially SRCNN, A+, ESPCN and bicubic interpolation - are usually employed for comparison of video SR works. In addition, following are the non-CNN video SR methods preferred for benchmarking:

1. **3DSKR [168]:** adaptive enhancement and spatio-temporal scaling, without explicit motion estimation, by solving a local weighted least-squares problem, where the weights are derived from the space/time comparison of neighboring pixels.
2. **ANN [169]:** employs artificial neural network (ANN) for learning "spatio-temporal" mappings between LR and HR frames.
3. **BayesSR [163]:** a Bayesian strategy to adaptively carry out HR frames reconstruction while at the same time also estimating motion, blur kernel and noise.
4. **Bayesian-MB [170]:** an expectation maximization (EM) strategy that specially focuses motion blur (MB) by optimally searching least blurred pixels for residual blur estimation and HR reconstruction.

A few works have benchmarked against Video Enhancer [171], a commercially available software.

4.2. Video Super-Resolution

As a variation to the iterative Bayesian adaptive multi-frame SR (MFSR) strategy [163] of maximum *a posteriori* (MAP) optical flow, noise level, and blur kernel estimation, the work in [164] proposes a non-iterative framework that generates a draft-ensemble from LR frame sequence, followed by the employment of a CNN to determine the optimal draft. The framework[18] has thus two parts, *viz.* a) the feed-forward SR draft ensemble

[17] http://www.cancerimagingarchive.net/

[18] http://www.cse.cuhk.edu.hk/leojia/projects/DeepSR/

Table 3: Publicly available video datasets.

S/No.	Name with reference	Details
1	Middlebury[a]	Has many datasets classified to five categories, like for optical flow [162].
2	Harmonic Inc.[b]	Arrays of 4K (Ultra HD) demo footage, like Myanmar 60p [98], in H.264 or ProRes 422 HQ format with 50p or 60p samples.
3	Videoset4 or Vid4[c]	"Consists of four test videos - *walk, foliage, city,* and *calendar* - which were also used in [163]" [98].
4	Xiph.org Test Media[d]	A collection of test sequences and video clips, e.g. Derf's collection[e]
5	Ultra Video Group[f]	A variety of 4K 120fps test sequences in raw as well as other state of the art formats, like HEVC, in cooperation with Digiturk[g].
6	YFCC100M [83]	As already described elsewhere, has 0.8 million are videos.
7	VidSet12 [102]	Following the guidelines of [164] constructed a test set of 48 sequences from 12 HR videos, each 31 frames long.
8	VideoSR[h] [164]	160 video sequences from "26 high-quality 1080p HD video clips" of "a variety of scenes and objects".
9	CDVL[i]	Consumer Digital Video Library (CDVL) contains 115 uncompressed full HD videos excluding repeated videos.
10	Sintel. [165, 166]	A computer generated dataset; popular for optical flow development.
11	VSB100 [167]	Berkeley Video Segmentation Benchmark with 40 training and 60 test video sequences.
12	25 YUV format video sequences[j] [163]	The available link works no more.

[a] http://vision.middlebury.edu
[b] https://www.harmonicinc.com/free-4k-demo-footage/
[c] https://twitter.box.com/v/vespcn-vid4
[d] https://media.xiph.org/
[e] https://media.xiph.org/video/derf/
[f] http://ultravideo.cs.tut.fi/
[g] http://www.digiturk.com.tr/
[h] http://www.cse.cuhk.edu.hk/leojia/projects/DeepSR/
[i] http://www.cdvl.org/
[j] http://www.codersvoice.com/a/webbase/video/08/152014/130.html

generation and b) a deep CNN to non-linearly combine the drafts (Fig. 7). Note that a SR draft ensemble is "the set of high-resolution patch candidates before final image deconvolution." The method is reported to have outperformed the reference method [163], as far as video quality is concerned. Time complexity has not been mentioned.

The video SR method (VSRnet) in [98] uses motion compensated consecutive frames as input to a CNN[19]. Following a three layer SRCNN architecture, the method offers three alternatives to combine consecutive video frames, i.e. concatenation a) before layer 1, b) between layer 1 and 2 or c) between layer 2 and 3. For the sake of accuracy and speed, the authors train their system on images (what they call pre-training) and then employ the ensued filter coefficients to initialize the video training. In addition, the authors claim 20% more time efficiency by employing what they call Filter Symmetry (in temporal sense) Enforcement (FSE); by this they mean equal weightage to t-i and t+i filter, if t represent the central frame temporally. Their motion compensation strategy

[19] http://ivpl.eecs.northwestern.edu/software

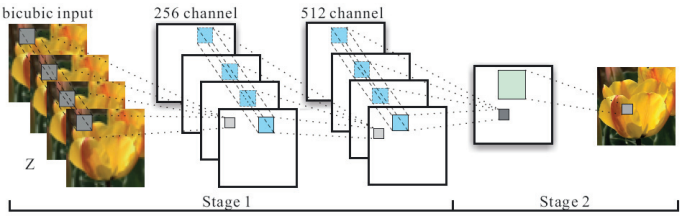


Figure 7: The CNN part of DraftCNN [164].

is flexible enough to deal with motion blur, in the case of fast moving objects. A comparison, with four image based (including A+ and SRCNN) and four video based methods, has been reported and the method is shown to be superior to all at upscale factors of 2 and 3; with upscale factor of 4, it is outperformed by BayesSR [163]. As far as time efficiency is concerned, VSRnet demonstrate marginally better execution time (accompanied with even better PSNR) in comparison to the four video-based methods (ANN, Bayesian-MB, BayesSR and Video Enhancer) but it is still slow for the usual reasons associated with videos. In addition, the video dataset is small (Mayanmar sequence with 59 scenes out of which 6 for training plus 4 from other sources they call Videoset4). They offer image pre-training as an alternate solution to the the creation of a large video database; a strategy also adopted in [172]. Still an image cannot be taken as a replacement for a video.

In [172], VSRnet is tweaked for compressed videos; hence the name CVSRnet. In contrast to a typical video super-resolution method, rather than taking the compression information - like frame type or quantizer step - from the encoder, CVSRnet relies on the compressed LR frames for the reconstruction of HR video. The rest of the strategy is the same as VSRnet; even the same eight methods have been used for comparison with almost similar results.

A strategy, outlined in [173], generates super-resolved video frames from LR videos and periodically realizes HR still frames. The original idea was to have a separate sensor in the video camera for periodically acquiring stills while recording a low resolution video; these stills were to be used in refining the videos. Zeng [174] carries out a CNN based implementation of this strategy, with the input being a sequences of video frames in which the first and last frames are HR, while the rest are LR. The HR frames are warped using the CNN based optical flow scheme, called FlowNet [175], wherein the optical flow, from the peripheral frames to each of the middle frame (LR), is calculated. The warped HR frames are subjected to a Graph-cut composition along with the upsampled low resolution frames and concatenated thereafter to produce a super-resolved video. The implementation [174] suffers from efficiency in comparison to SRCNN. This may be attributed to the expensive nature of FlowNets [175]. The implementer puts the blame on the time spent on data transit. Moreover, with Sintel database they used, the results are not enviable which may be due to the lacking of natural details because of the synthetic nature of the dataset and the author feels that there is less information to discard during downsampling.

The multi-frame (MFCNN) video SR method in [102] improves the resolution of a given frame based on the pixel values in the adjacent frames, within distance d on either side. In other words, information from a total of $2d+1$ frames is concatenated along the channel dimension of the CNN, as shown in Fig. 8. The authors employ a SRCNN inspired architecture, with 9 ReLU based layers and dropout, to train on still images (SICNN), followed by single frames (SFCNN) and ultimately extending to multi-frames (MFCNN). Testing with Set14 databases

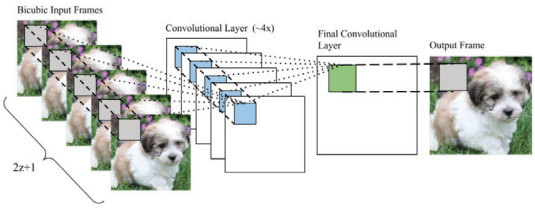


Figure 8: MFCNN [102].

shows that SICNN has comparable results to SRCNN for obvious reason, but both are outperformed by CSCN-MV [107]. The video version (MFCNN) is, however, reported to have outperformed DraftCNN and BayesSR on their own set of videos (VidSet12). The authors claim minimal data pre-processing and computation cost but themselves recognize inconsistencies in results in certain situations, e.g. LR scenes with trees and leaves. They are hopeful about the potential role of RCNN (R for recurrent) in video super-resolution.

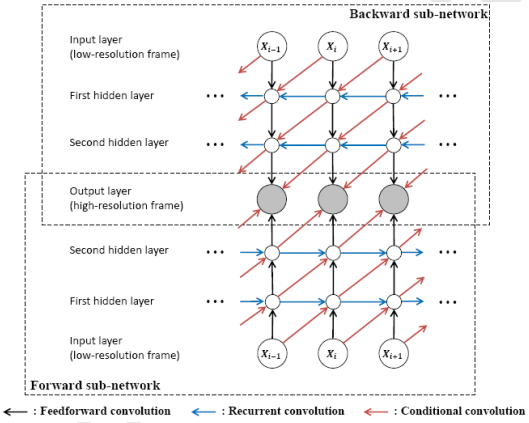


Figure 9: BRCN [176].

In [176], a bidirectional recurrent CNN (BRCN) is proposed for MFSR, under the belief that a recurrent neural network (RNN) is better suited to "model long-term contextual information of temporal sequences". As illustrated in Fig. 9, the method employs three types of convolutions, *viz.* a) feed-forward convolution for LR to HR spatial correspondence, b) recurrent convolution for learning temporal dependence via weightage based linking of hidden layers of adjacent frames, and c) conditional convolution to connect previous inputs to current hidden layer in order "to enhance visual-temporal dependency". Empirically, the method is reported to have been slower than methods, like ANR and SRCNN, but with better visual quality (read PSNR); theoretical complexity analysis is, however, lacking.

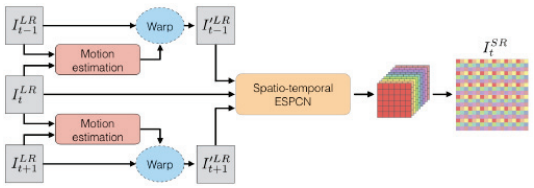
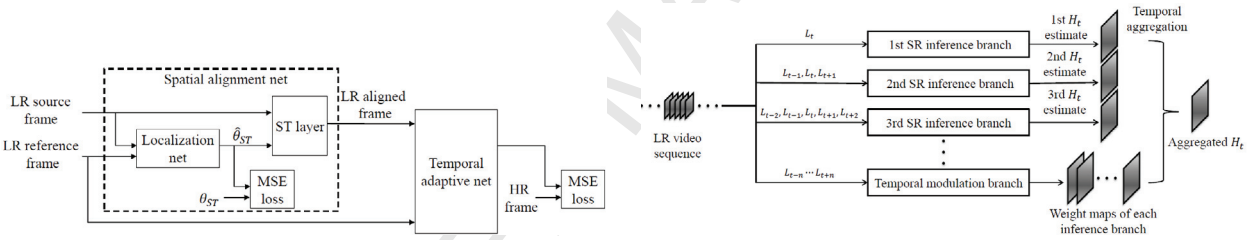


Figure 10: VESPCN [177].

The ESPCN [66], already described elsewhere, had been shown to perform well (+0.39dB) on 1080p videos with 0.03–0.04s per frame execution time as against 0.434 for SRCNN. The video version of ESPCN (VESPCN [177]), illustrated in Fig. 10, relies on spatio-temporal sub-pixel CNNs that take into account the temporal redundancies and a fast multi-resolution spatial transformer motion compensation module. Although the authors analyze a) early fusion (as employed in [98]), b) slow fusion [178] and 3) 3D convolutions [179] for the joint processing of multiple consecutive video frames; the experiments are restricted to early and slow fusions only. The authors boast a computational cost economy of 30% par rapport SFMSR techniques; a 0.2dB PSNR gain if computational cost is kept constant. The benchmarking results on Vid4 dataset establishes the superiority of VESPCN against VSRnet, ESPCN and SRCNN, both in terms of visual quality and time efficiency. An action recognition method [180] is reported to have improved results by enhancing the video quality via SRCNN before subjecting it to the recognition method proposed in [181]. The authors of [182] stress the need of careful frame alignment and motion compensation for video SR. To this end they propose to incorporate a "sub-pixel motion compensation" (SPMC) layer in their CNN while fusing multiple frames. The reported results demonstrate improvements but barring VESPCN, results of other benchmarking methods seem inconsistent with those reported in other sources from the literature.



(a) Overall SAN (b) The TANN module

Figure 11: SAN and its TANN module [183]

One video SR method [183, 184] employs a two-pronged strategy to handle the temporal challenge side by side the spatial challenge[20]. First, their "temporal adaptive neural network" (TANN) introduces temporal aggregation to the MSCN-n [110] pipeline described elsewhere, by replacing its "adaptive weight module" with a "temporal modulation branch" (TMB) in parallel to the already time-scale improvised SR inference branches. The idea is to temporally aggregate output from the SR inference branches based on the weight map outputs from the TMB. Second, the preceding "spatial alignment network" pre-processes each of the neighboring LR frames based on its reference LR frame to produce spatially aligned SR frames that serve as inputs to the TANN, as can bee seen in Fig. 11. The authors see TANN as a module of their method which they call SAN after their spatial alignment network. The method demonstrates improvements over others even in light weight form ('Proposed-S') which seems also time-efficient.

The deep network in [185, 186], learn HR/LR frame difference (spatial residues) and difference between adjacent HR frames (temporal residues) in order to connect the intra- and inter-frame redundancies in the

[20] http://www.ifp.illinois.edu/~dingliu2/videoSR/

video sequence. In [187], the authors take into account spatio-temporal relation between adjacent LR frames, in addition to LR/HR mappings.

Extension of EDSR and MDSR, winners of NITRE2017 SR Challenge as described earlier [113], to video SR is the method called "Deep Recurrent ResNet (DRRNet)" [188]. Rather than relying on motion compensation or optical flow, the authors look up to long-short-term memory (LSTM[21]) and ResNet to deal with adjacent frames. Use of LSTM may boost efficiency of the process for its $\mathcal{O}(1)$ complexity. The method is compared to a few single frame methods and shows comparable results but does not excel them as was done by EDSR/MDSR against their counterparts. Hence there is still room for refinement.

4.3. Videoset4-Based Benchmarking

Table 4 compares various SR methods on the basis of Videoset4 at an upscaling factor of ×3 and ×4. Note that the first five are image SR methods. For the table we have relied on two sources only [98, 177], as most of the works are inconsistent in their approach as far as the datasets are concerned. One can count VSRnet as lossless CVSRnet and obviously lossy CVSRnet will give lower PSNR. ESPCN, which is mainly image based method, is reported to have outperformed VSRnet in the presentation of VESPCN (an extension of ESPCN by almost the same team); even the reported PSNR is lesser (26.64 dB) than shown in the table (26.79 dB). For the effectiveness of the rest of the methods one needs to consult individual works. Strangely enough, in the presentation of MFCNN [102], bicubic interpolation outperforms not only BayesSR but also the single frame version of the method (SFCNN) on the VidSet12 dataset. The results of the work concerning BRCN relied on a dataset which is no more accessible, at least at the time of writing of this article.

Table 4: VSR Benchmarking over Vid4 with ×3 and ×4 upscaling.

S/No.	**Name with reference**	**PSNR (dB)** ×3	**PSNR (dB)** ×4
1	Bicubic	25.28	23.79
2	SC [40]	26.01	–
3	A+ [48]	26.36	24.59
4	SRCNN [5]	26.51	24.69
5	ESPCN [66]	26.97	–
6	ANN [169]	25.94	23.97
7	BayesSR [163]	25.82	25.06
8	Bayesian-MB [170]	26.43	24.14
9	Video Enhancer [171]	26.34	24.55
10	VSRnet [98]	26.79	24.84
11	VESPCN [177]	27.25	25.35
12	Tao *et al* [182]	27.49	25.52
13	SAN [183]	–	25.83

[21] http://www.bioinf.jku.at/publications/older/2604.pdf

5. Depth Maps/3D and higher dimensions

Depth map super-resolutions has recently attracted increased research and even commercial[22] interest. Named variously - such as depth map, range image, DEM, 3rd D, 2.5th D etc. - in many cases, the idea is to create triangulation from height coefficients wherein each coefficient corresponds to a squared area of a related texture image. The texture is then draped onto the triangulated model in order to render a 3D perspective. The underlying idea of depth map SR is to infer/complete it from the LR input under the guidance of the corresponding texture map. Following the pattern of the previous two sections, here we try to look into the literature related to depth/3D images. Note that depth estimation methods [189, 190] are out of scope of this article.

5.1. Benchmarks

Table 5 lists some of the important publicly available datasets. Unlike, the datasets from the previous sections, these datasets mostly address specific situations, as described in the third column of the table.

Table 5: Publicly available 3D sequences and depth maps.

S/No.	Name with reference	Details
1	Middlebury stereo dataset 2001 [191], 2003 [192], 2005/06 [193], and 2014 [194]	contains HR textures and depths with lots of details [195].
2	Laser Scan. [196]	Specially developed for patch based SR.
3	Sintel [165, 166]	a synthesized dataset via physical simulations containing high quality images along with lots of depth details [195].
4	SENTINEL-2[23]	images having 13 channels (as against 3 in RGB) with a ground resolution of up to 10m, and a high radiometric resolution (more than 8 bit per pixel for each channel).
5	ICL-NUIM [197]	For benchmarking RGB-D, Visual Odometry and SLAM algorithms with two different scenes, *viz.* the living room and the office room scene.
6	NYU Depth [198]	Composed of 464 indoor video sequences from a Microsoft Kinect camera; 249 scenes for training and 215 for testing.
7	KITTI [199]	composed of several outdoor scenes captured while driving with car-mounted cameras and depth sensor.
8	ToFMark [200]	Time-of-Flight (ToF) Depth Upsampling Evaluation Dataset having LR ToF depth acquisition together with a HR intensity image.
9	Open Access Series of Imaging Studies (OASIS [201])	A collection of cross-sectional MRI Data from 416 subjects aged 18 to 96 year.

Aside from some of the methods described in Section 3 - like ANR, A+, NE+, K-SVD, SRCNN - many of the following non-CNN methods are usually employed as benchmarks:

1. **Guided Image Sampling [202]:** an edge preserving filter that can be used joint upsampling.

[22] https://patents.google.com/patent/CN107358576A/en

2. **MRF [203]:** Markov Random Fields based enhancement of LR depths with insight from the accompanied HR camera images under the assumption that depth discontinuities are usually in harmony with intensity changes in the associated camera image.
3. **ATGV [200]:** depth image upsampling guided by an anisotropic diffusion tensor that is computed from HR intensity image.
4. **3D-ToF Upsampling [204]:** attempts to super-resolve LR depths from "noisy 3D time-of-flight (3D-ToF) camera coupled with a [HR] RGB camera" by using non-local mean filtering to regularize depth maps and a multi-features based edge weighting scheme based on the HR RGB input.
5. **PatchSDSR [196]:** increases the depth resolution by matching the height field of each LR patch input against "only a generic database of local [HR] patches"; the selection of right HR candidate is done via MRF labeling.
6. **Edge-guided [205]:** super-resolves a single depth image with the help of a HR edge map, obtained by MRF optimization from the edges in its LR counterpart.

Other important non-CNN methods for depth SR benchmarking include [206, 207, 208].

5.2. Literature on CNN based SR of 3D/Depths and Multispectral Data

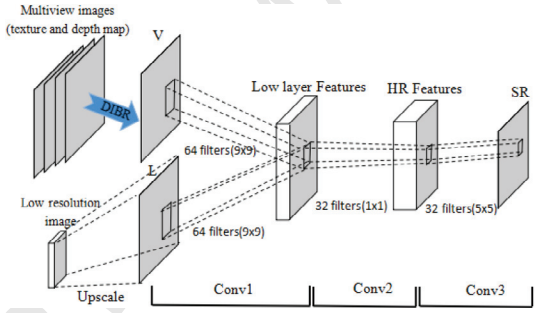


Figure 12: 3DSRCN [209].

In 3DSRCN [209], the LR viewpoints from the input 3D video are upsampled via interpolation which is fed, along with a projected virtual view, to the first layer of a three-layered CNN, as shown in Fig. 12. In the experimental set up, Layer-1 involves 2×64 filters (9×9), layer-2 involves 64x32 filters (1×1) and layer-3 contains and 32×1 filters (5×5), all with the convolution stride of 1. Layer-1 combines the two inputs to get a set of feature maps, layer-2 establishes the mapping from low to high resolution features, while layer-3 uses the ensued features to get the super-resolved image. The simulation results, on samples from Middlebury stereo 2014 [194], report a PSNR gain of about 1dB par rapport the SRCNN and another virtual-view assisted method reported in [210]. The authors also claim empirical time efficiency. The article thus suffers from poor benchmarking coverage and lack of complexity analysis.

The work in [195] proposes what the authors call a progressive deep CNN structure (Fig. 13) that develop HR depth images by a gradual learning of the higher frequencies. Supplementary information in the form of "depth field statistics and color-depth correlation" serves as "two priors provide complement to the CNN" for further refinement of depth maps.The authors argue that the method is equally good in the absence of HR color

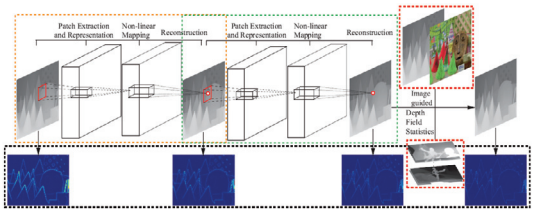


Figure 13: Progressive Deep CNN [195].

images whereby the depth images by themselves are enough to ameliorate the depth images. The reported results compare the method with more than a dozen state of the art methods applied to standard datasets and the results establish the superiority of the method.

The term DEM super-resolution was first coined in [211] wherein the non-local algorithm focused on some learning examples, itself derived from the original DEM by dividing the latter to overlapping patches. The patches are then categorized as test dataset and learning dataset based on the fact whether they pertain to HR measurements or not. By calculating the weighted sums of similar patches, HR DEM are restored. The same team employs a three layers' CNN model [212] to extend this strategy, with special focus on DEM compatibility and robustness, wherein layer 1 detects features from the input, layer 2 integrates the ensued features and layer 3 transforms the integrated features to get the super-resolved DEM. The author claim good results but they are compared with bicubic interpolation only.

In order to get a super-resolved output from a single LR depth map, the method ATGV-Net [72][24] attempts to combine CNN with an energy minimization model, in the form of a powerful variational model. They improvise SRCNN to estimate, in addition to refining the depth map, the locations of discontinuities in the HR output depth maps. The refined depths and the located discontinuities serve as input to a variational model wherein pairwise regularization is carried out via anisotropic Total Generalized Variation (TGV) [213], with weights dependent on the network output. The output is the final HR estimate of the depth map. The authors train their model entirely on their own synthetic depth data. For benchmarking however, they use four different datasets; two derived from Middlebury and one each from Laserscan and ToFMark. The method is demonstrated to show better results in comparison to different state of the art methods, especially [200] which is shown to be second best. The work is extended in [214] by the introduction of a high resolution image to guide the reconstruction. The method relies on two jointly trained end-to-end networks, *viz.* fully-convolutional network (FCN) and the the subsequent primal-dual network (PDN); whilst the former super-resolves the depth map, the latter deals mainly with noise removal.

In [216], the authors employ an already SR trained CNN (transfer learning[25]) to post-process lunar images in order to estimate the DEM. For the latter part, they introduce to their architecture an additional ConvNet, which is trained anew. The authors have compared the method with bicubic interpolation only which is outperformed for obvious reasons.

[24] https://griegler.github.io/

[25] http://cs231n.github.io/transfer-learning/

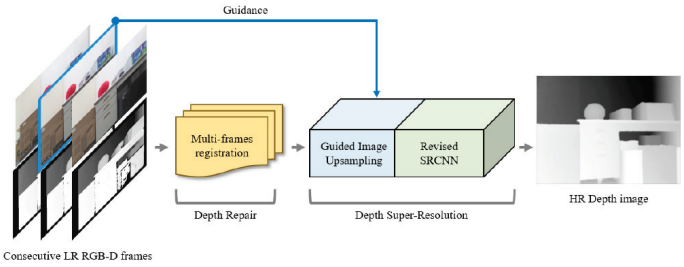


Figure 14: Depth SR by [215].

The method described in [217] tries to update LR depth information with HR intensity information by employing a multi-scale fusion technique; hence the names Multi-Scale Guided convolutional network (MSG-Net). While the introduction of fusion layers has obvious benefits, in the form of better visual quality and better root mean square error (RMSE), complexity may well increase. As a mitigation, the authors claim efficiency due to training in the high-frequency domain.

The Joint Convolution Neural Pyramid (JNCP [218]) method consists of three sub-networks; two convolutional neural pyramids (CNPs [219] concatenated by a normal convolutional neural network". Whilst the former extract informative features from the receptive fields of the depth and guidance maps, the latter completes the depth image from the guidance image, as in [217]. The reported results show improvements in RMSE par rapport the state of the art.

The work in [220] employs CNN to super-resolve the input edge map, obtained from the bicubically upsampled LR depth image, to get the guidance edge map that is used to complete the upsampled LR depth image via the total variation (TV) model. The reported results show some improvements, as far as RMSE is concerned, but not as much as in [195] (e.g. 3.1742 units as against 1.8006 units for *Cones* from Middlebury at $\times 4$; 66.7% inferior). A similar approach [221] is applied in [221] wherein the HR color image is used to complete the LR depth map. The results are not enviable, however, e.g. depth map of *Cones* is super-resolved to an RMSE of 5.8 units.

The authors of [222] call their CNN Depth Enhancement Network (DEN) that learns the end-to-end mapping between LR and HR depths. In parallel, they propose what they call Color-Based Prediction Network (CBPN), an optional deep FCN in order to do color based guidance to complete the depth map. If CBPN is part of the pipeline (they call it DEN+CBPN) then there is a subsequent 'automerging' CNN to further estimate missing HR information, of course at the expense of time escalation while adding little value; the ensued results show that DEN performs better than DEN+CBPN in most cases. The methods itself however seem to under-perform par rapport ATGV-Net [72] over noisy Middlebury data.

The authors of [223] argue that CNN SR approaches are as useful for remote sensing data as for still images. Their extension of SRCNN to multispectral information (hence the names msiSRCNN) is reported to have given superior results to conventional interpolation methods in terms of PSNR. They had applied the method to SENTINEL-2 images with 13-channel (spectral bands), including the 3 RGB channels, encoded with 16 bit per pixel JPEG2000 format. It would have been, however, better to report RMSE for some distance critical channels. A work on 3D cardiac Magnetic Resonance (MR) imaging [224] super-resolves the corresponding 2D planes using an improvised SRCNN that is modified by employing residual learning in LR-HR mapping and using a deeper

network with the upscaling being realized learning a deconvolution layer rather than a fixed kernel. SRCNN has been utilized to enhance spatial features in the hyperspectral image reconstruction model proposed in [225]. Another interesting read on multi-spectral resolution is a three part article on CosmiQ [226] in the context of SpaceNet challenge[26]. As described earlier, the authors of [152] claim time efficiency ×200 on 3D volumes on medical data from OASIS.

5.3. Banchmarking over a standard dataset

Before dwelling on the comparison, it must be noted that whilst for texture quality PSNR and SSIM are suited, the root mean square error (RMSE) is usually the preferred metric for depth maps. RMSE in taken as a length unit, like meter (m) or its subdivisions and even *pixel disparity*. An analysis of RMSE par rapport the view distance is given in [227]. As of comparing various methods, unfortunately, due to their dealings with diverse problems, most of the works have used different approaches of benchmarking, especially in choosing the datasets. Even with the same dataset, the subset chosen is different from one work to another. Then there are inconsistencies in reporting the results. For example, the works in [72, 195] have benchmarked their methods on Middlebury and Laser Scan data, with many common reference methods. Just a look at ×4 upscaling part on RMSE, as listed in Table 6, reveals inconsistencies in their reported results.

Table 6: RMSE results for *Cones* from Middlebury at ×4.

S/No.	**Name with reference**	**Riegler *et al.* [72]**	**Song *et al.* [195]**
1	Nearest neighbor	6.1236	6.0054
2	Bicubic	4.9544	3.8635
3	ANR [47]	3.0256	3.3156
4	K-SVD [96]	3.8468	3.2232
5	ATGV [200]	3.6372	3.9968
6	3PatchSDSR [196]	6.0168	12.6938

6. Difference of Approach Towards Image, Video and Depth Map

We have surveyed SR methods based on the nature of the media, i.e. image, video and depth maps. Strictly speaking, a video can be thought of a collection of frames or images. Same can be true of depth maps, as they can be treated as gray-scale images wherein each pixel may be equivalent to the corresponding depth coefficient; the intensity of the pixel is thus in proportion to the height represented by the coefficient. As such, all image SR methods should work well for videos and depths. While this may be true to some extent, videos and depth maps do have their peculiarities.

With videos, an image SR method may address the spatial aspect if each frame is treated independently as image, but the temporal aspect is left out, altogether. With raw videos, you can treat each frame as image but

[26] http://crowdsourcing.topcoder.com/spacenet

look at the enormity of task in the form of at least 30 frames a second; definitely the time/space overhead is huge. With modern video codecs, one cannot take the risk of being content with still image SR approaches while ignoring facts like motion estimation, optical flow and multi-resolved frames. For a good video SR technique, in addition to spatial challenge common with image SR, the important challenge is to "efficiently and effectively exploit the temporal" dependencies among adjacent LR frames; the underlying complex motion is hard to model and mishandling may eventually prove counter productive. Research on image SR is way ahead and mature as compared to video SR; the focus in the latter case is algorithms incorporating motion estimation and denoising. And here is the opportunity which newer methods usually exploit, i.e. take an established image SR method and modify the pipeline to encompass the temporal aspect; may be in parallel or in sequences.

With depths, the data may be less but the SR task is multiplied by the requirement of higher accuracy, especially in critical applications where a small error in one coefficient can be catastrophic. According to [195], unlike the images, the adoption of CNNs for depth SR has been slow because the ill-posedness of depth SR is worse. This is partly attributed to the differences between the acquisition of depths and still image. In addition, depth images are less textured, exhibit sharp boundaries and are noisier. Hence a different approach is required in dealing with such issues by either modifying existing methods or coming up with a novel method.

7. Conclusion

Despite the claims of accuracy and time efficiency, in almost all the works reviewed in this survey, a lot needs to be done on both fronts. The accuracy improvements are marginal and that too with inconsistencies; even there are instances that work A claims superiority to work B and after a while B comes up with superior results after minor tweaking - and this goes on and on. A minor thing to point out is that none of the works have reported the quality of input LR par rapport the HR during training. When it comes to time efficiency, empirical results are commonplace but little can be found on the theoretical front; in the form of asymptotic analysis etc. Although there are attempts to exploit sparse coding for deep learning, the race between the sparse coding paradigm and deep learning paradigm is in full swing without the realization of the fact that the issue is SR, not the superiority of one paradigm over other.

So far most of the attention has gone to CNNs and the potential of other deep learning approaches is yet to be exhaustively investigated. But the most important thing is the loss of sight from high scaling factors. In [114], the authors rightly pose the question, "how do we recover the finer texture details when we super-resolve at large upscaling factors?" To them the problem is the thrust on choosing the objective function and on minimizing the MSE, which may yield high PSNRs but at the expense of losing high-frequency details, thus leading to poor subjective quality of the output HR. There's also a need to come up with metrics that take into account the peculiarities of the deep learning architecture, for a more justified comparison of the methods; side by side to the attempts like [228].

Deep learning has its constraints. According to Shalev-Shwartz *et al.* [229], it is haunted by failures from at least four origins, viz. a) non-informative gradients, b) inefficiency of a network left to learn itself, c) architecture choice and conditioning and d) flat activations. A researcher must not be oblivious to these while designing a

deep learning network. This discussion is incomplete without an allusion to a news article [230] on the limitations of deep learning. It's worth a read.

References

[1] A. Vedaldi, K. Lenc, Matconvnet – convolutional neural networks for matlab, in: Proc. 25th annual ACM international conference on Multimedia, 2015.

[2] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: Convolutional architecture for fast feature embedding, in: Proc. 22Nd ACM International Conference on Multimedia, MM '14, ACM, New York, NY, USA, 2014, pp. 675–678.

[3] Q. Yin, Z. Cao, E. Zhou, Face hallucination using convolutional neural networks, uS Patent 9,405,960 (Aug. 2 2016). URL `https://www.google.ch/patents/US9405960`

[4] C. Dong, C. C. Loy, K. He, X. Tang, Learning a Deep Convolutional Network for Image Super-Resolution, Springer International Publishing, Cham, 2014, pp. 184–199.

[5] C. Dong, C. C. Loy, K. He, X. Tang, Image super-resolution using deep convolutional networks, IEEE Transactions on Pattern Analysis and Machine Intelligence 38 (2) (2016) 295–307.

[6] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, `http://www.bioinfo.org.cn/~casp/temp/DeepLearning.pdf`, (Accessed on 06/05/2017) (May 2015).

[7] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444. `doi:10.1038/nature14539`.

[8] G. Hinton, IPAM summer school 2012 tutorial on: Deep learning, `https://pdfs.semanticscholar.org/3441/4a6175d9c3c40bcea606b7c457104e973cb3.pdf`, (Accessed on 06/05/2017) (2012).

[9] G. E. Hinton, S. Osindero, Y.-W. Teh, A fast learning algorithm for deep belief nets, Neural Comput. 18 (7) (2006) 1527–1554.

[10] V. Golovko, A. Kroshchanka, D. Treadwell, The nature of unsupervised learning in deep neural networks: A new understanding and novel approach, Optical Memory and Neural Networks 25 (3) (2016) 127–141.

[11] Cs231n convolutional neural networks for visual recognition, `http://cs231n.github.io`, (Accessed on 06/05/2017).

[12] G. Cybenko, Approximation by superpositions of a sigmoidal function, Mathematics of Control, Signals and Systems 2 (4) (1989) 303–314.

[13] M. Nielsen, Neural networks and deep learning, `http://neuralnetworksanddeeplearning.com/chap4.html`, (Accessed on 06/05/2017) (May 2017).

[14] X. Glorot, A. Bordes, Y. Bengio, Deep sparse rectifier neural networks, in: G. Gordon, D. Dunson, M. Dudík (Eds.), Proc. Fourteenth International Conference on Artificial Intelligence and Statistics, Vol. 15 of Proceedings of Machine Learning Research, PMLR, Fort Lauderdale, FL, USA, 2011, pp. 315–323.

[15] Y. L. Cun, B. Boser, J. S. Denker, R. E. Howard, W. Habbard, L. D. Jackel, D. Henderson, Advances in neural information processing systems 2, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1990, Ch. Handwritten Digit Recognition with a Back-propagation Network, pp. 396–404.

[16] J. Schmidhuber, Critique of paper by "deep learning conspiracy" (nature 521 p 436), `http://people.idsia.ch/~juergen/deep-learning-conspiracy.html`, (Accessed on 06/21/2017) (Jun 2015).

[17] Y. LeCun, Y. Bengio, The handbook of brain theory and neural networks, MIT Press, Cambridge, MA, USA, 1998, Ch. Convolutional Networks for Images, Speech, and Time Series, pp. 255–258.

[18] E. Culurciello, Neural network architectures – towards data science – medium, https://medium.com/towards-data-science/neural-network-architectures-156e5bad51ba, (Accessed on 06/06/2017) (Mar 2017).

[19] D. C. Ciresan, U. Meier, L. M. Gambardella, J. Schmidhuber, Deep, big, simple neural nets for handwritten digit recognition, Neural Computation 22 (12) (2010) 3207–3220.

[20] A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, in: F. Pereira, C. J. C. Burges, L. Bottou, K. Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 25, Curran Associates, Inc., 2012, pp. 1097–1105.
URL http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[21] M. D. Zeiler, R. Fergus, Visualizing and Understanding Convolutional Networks, Springer International Publishing, Cham, 2014, pp. 818–833.

[22] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1–9. doi:10.1109/CVPR.2015.7298594.

[23] G. Huang, Z. Liu, K. Q. Weinberger, Densely connected convolutional networks, CoRR abs/1608.06993.
URL http://arxiv.org/abs/1608.06993

[24] A. Canziani, A. Paszke, E. Culurciello, An analysis of deep neural network models for practical applications, CoRR abs/1605.07678.
URL http://arxiv.org/abs/1605.07678

[25] Ujjwalkarn, An intuitive explanation of convolutional neural networks – the data science blog, https://ujjwalkarn.me/2016/08/11/intuitive-explanation-convnets/, (Accessed on 06/05/2017) (Aug 2016).

[26] D. Britz, Understanding convolutional neural networks for nlp – wildml, http://www.wildml.com/2015/11/understanding-convolutional-neural-networks-for-nlp/, (Accessed on 06/06/2017) (Nov 2015).

[27] F. Shaikh, Comparison between deep learning & machine learning, https://www.analyticsvidhya.com/blog/2017/04/comparison-between-deep-learning-machine-learning/, (Accessed on 06/06/2017) (Apr 2017).

[28] C. Olah, Home - colah's blog, http://colah.github.io/, (Accessed on 06/06/2017).

[29] G. Cristobal, E. Gil, F. Sroubek, J. Flusser, C. Miravet, F. B. Rodriguez, Superresolution imaging: a survey of current techniques (2008).

[30] K. Nasrollahi, T. B. Moeslund, Super-resolution: a comprehensive survey, Machine Vision and Applications 25 (6) (2014) 1423–1468.

[31] M. Bevilacqua, Algorithms for super-resolution of images and videos based on learning methods, Theses, Université Rennes 1 (Jun. 2014).
URL https://tel.archives-ouvertes.fr/tel-01064396

[32] T. S. Huang, R. Y. Tsai, Multiframe image restoration and registration, Advances in Computer Vision and Image Processing 1 (7) (1984) 317–339.

[33] M. Irani, S. Peleg, Super resolution from image sequences, in: Proc.10th International Conference on Pattern Recognition, Vol. 2 of ICPR, 1990, pp. 115–120.

[34] H. He, W. C. Siu, Single image super-resolution using gaussian process regression, in: CVPR 2011, 2011, pp. 449–456. doi:10.1109/CVPR.2011.5995713.

[35] K. Zhang, X. Gao, D. Tao, X. Li, Single image super-resolution with non-local means and steering kernel regression, IEEE Transactions on Image Processing 21 (11) (2012) 4544–4556. doi:10.1109/TIP.2012.2208977.

[36] P. Rasti, K. Nasrollahi, O. Orlova, G. Tamberg, T. B. Moeslund, G. Anbarjafari, Reducible dictionaries for single image super-resolution based on patch matching and mean shifting, J. Electronic Imaging 26 (2) (2017) 23024.
URL https://doi.org/10.1117/1.JEI.26.2.023024

[37] W. T. Freeman, T. R. Jones, E. C. Pasztor, Example-based super-resolution, IEEE Computer Graphics and Applications 22 (2) (2002) 56–65. doi:10.1109/38.988747.

[38] D. Glasner, S. Bagon, M. Irani, Super-resolution from a single image, in: 2009 IEEE 12th International Conference on Computer Vision, 2009, pp. 349–356. doi:10.1109/ICCV.2009.5459271.

[39] T.-M. Chan, J. Zhang, J. Pu, H. Huang, Neighbor embedding based super-resolution algorithm through edge detection and feature selection, Pattern Recogn. Lett. 30 (5) (2009) 494–502.

[40] J. Yang, J. Wright, T. S. Huang, Y. Ma, Image super-resolution via sparse representation, Trans. Img. Proc. 19 (11) (2010) 2861–2873. doi:10.1109/TIP.2010.2050625.
URL http://dx.doi.org/10.1109/TIP.2010.2050625

[41] X. Gao, K. Zhang, D. Tao, X. Li, Image super-resolution with sparse neighbor embedding, IEEE Transactions on Image Processing 21 (7) (2012) 3194–3205. doi:10.1109/TIP.2012.2190080.

[42] S. Dai, M. Han, W. Xu, Y. Wu, Y. Gong, Soft edge smoothness prior for alpha channel super resolution, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2007.

[43] H. A. Aly, E. Dubois, Image up-sampling using total-variation regularization with a new observation model, IEEE Transactions on Image Processing 14 (10) (2005) 1647–1659. doi:10.1109/TIP.2005.851684.

[44] Q. Shan, Z. Li, J. Jia, C.-K. Tang, Fast image/video upsampling, ACM Trans. Graph. 27 (5) (2008) 153:1–153:7. doi:10.1145/1409060.1409106.
URL http://doi.acm.org/10.1145/1409060.1409106

[45] W. Dong, L. Zhang, R. Lukac, G. Shi, Sparse representation based image interpolation with nonlocal autoregressive modeling, IEEE Transactions on Image Processing 22 (4) (2013) 1382–1394. doi:10.1109/TIP.2012.2231086.

[46] H. Chang, D.-Y. Yeung, Y. Xiong, Super-resolution through neighbor embedding, in: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004., Vol. 1, 2004, pp. I–I. doi:10.1109/CVPR.2004.1315043.

[47] R. Timofte, V. De Smet, L. Van Gool, Anchored neighborhood regression for fast example-based super-resolution, in: Proc. IEEE International Conference on Computer Vision, 2013, pp. 1920–1927.

[48] R. Timofte, V. De Smet, L. Van Gool, A+: Adjusted Anchored Neighborhood Regression for Fast Super-Resolution, Springer International Publishing, Cham, 2015, pp. 111–126.

[49] S. Schulter, C. Leistner, H. Bischof, Fast and accurate image upscaling with super-resolution forests., in: CVPR, IEEE Computer Society, 2015, pp. 3791–3799.

[50] J. Salvador, E. Perez-Pellitero, Naive bayes super-resolution forest, in: The IEEE International Conference on Computer Vision (ICCV), 2015.

[51] Y. Liang, J. Wang, S. Zhou, Y. Gong, N. Zheng, Incorporating image priors with deep convolutional neural networks for image super-resolution, Neurocomputing 194 (2016) 340 – 347.

[52] J. Kim, J. K. Lee, K. M. Lee, Accurate image super-resolution using very deep convolutional networks, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1646–1654. doi:10.1109/CVPR.2016.182.

[53] Y. Liang, Z. Yang, K. Zhang, Y. He, J. Wang, N. Zheng, Single image super-resolution with a parameter economic residual-like convolutional neural network, CoRR abs/1703.08173.
URL http://arxiv.org/abs/1703.08173

[54] N. K. Bose, N. A. Ahuja, Superresolution and noise filtering using moving least squares, IEEE Transactions on Image Processing 15 (8) (2006) 2239–2248. doi:10.1109/TIP.2006.877406.

[55] A. J. Patti, M. I. Sezan, A. M. Tekalp, Superresolution video reconstruction with arbitrary sampling lattices and nonzero aperture time, IEEE Transactions on Image Processing 6 (8) (1997) 1064–1076. doi:10.1109/83.605404.

[56] H. Ji, C. Fermüller, Robust wavelet-based super-resolution reconstruction: Theory and algorithm, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (4) (2009) 649–660. doi:10.1109/TPAMI.2008.103.

[57] M. B. Chappalli, N. K. Bose, Simultaneous noise filtering and super-resolution with second-generation wavelets, IEEE Signal Processing Letters 12 (11) (2005) 772–775. doi:10.1109/LSP.2005.856875.

[58] J. Tian, K.-K. Ma, Stochastic super-resolution image reconstruction, J. Vis. Comun. Image Represent. 21 (3) (2010) 232–244.

[59] S. P. Belekos, N. P. Galatsanos, A. K. Katsaggelos, Maximum a posteriori video super-resolution using a new multichannel image prior, IEEE Transactions on Image Processing 19 (6) (2010) 1451–1464. doi:10.1109/TIP.2010.2042115.

[60] Q. H. Luong, Advanced image and video resolution enhancement techniques, Ph.D. thesis, (Accessed on 06/09/2017) (Mar 2009).

[61] J. Tian, K.-K. Ma, A survey on super-resolution imaging, Signal, Image and Video Processing 5 (3) (2011) 329–342.

[62] D. Huang, H. Liu, A short survey of image super resolution algorithms, Journal of Computer Science Technology Updates 2 (2) (2015) 19–29.

[63] E. Karimi, K. Kangarloo, S. Javadi, Article: A survey on super-resolution methods for image reconstruction, International Journal of Computer Applications 90 (3) (2014) 32–39, full text available.

[64] J. Kim, J. Kwon Lee, K. Mu Lee, Deeply-recursive convolutional network for image super-resolution, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[65] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei, Imagenet large scale visual recognition challenge, International Journal of Computer Vision 115 (3) (2015) 211–252. doi:10.1007/s11263-015-0816-y.
URL http://dx.doi.org/10.1007/s11263-015-0816-y

[66] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, Zehan, Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network (2016) 1874–1883.

[67] D. Martin, C. Fowlkes, D. Tal, J. Malik, A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics, in: Proc. 8th Int'l Conf. Computer Vision, Vol. 2, 2001, pp. 416–423.

[68] A. Krizhevsky, Learning multiple layers of features from tiny images, Tech. rep., University of Toronto, Canada (2009).
URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf

[69] R. Timofte, R. Rothe, L. V. Gool, Seven ways to improve example-based single image super resolution, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1865–1873. doi:10.1109/CVPR.2016.206.

[70] R. Timofte, V. D. Smet, L. V. Gool, Supplementary material (a+: Adjusted anchored neighborhood regression for fast super-resolution), http://www.vision.ee.ethz.ch/~timofter/ACCV2014_ID820_SUPPLEMENTARY/, (Accessed on 06/05/2017).

[71] C. Dong, C. C. Loy, X. Tang, Accelerating the super-resolution convolutional neural network, CoRR abs/1608.00367.
URL http://arxiv.org/abs/1608.00367

[72] G. Riegler, M. Rüther, H. Bischof, ATGV-Net: Accurate Depth Super-Resolution, Springer International Publishing, Cham, 2016, pp. 268–284.

[73] J. B. Huang, A. Singh, N. Ahuja, Single image super-resolution from transformed self-exemplars, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5197–5206. doi:10.1109/CVPR.2015.7299156.

[74] R. Franzen, True color kodak images, http://r0k.us/graphics/kodak/, (Accessed on 06/05/2017).

[75] D. Dai, R. Timofte, L. Van Gool, Jointly optimized regressors for image super-resolution, Comput. Graph. Forum 34 (2) (2015) 95–104.

[76] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324. doi:10.1109/5.726791.

[77] R. Salakhutdinov, I. Murray, On the quantitative analysis of deep belief networks, in: Proc. 25th International Conference on Machine Learning, ICML '08, ACM, New York, NY, USA, 2008, pp. 872–879.

[78] R. Dahl, M. Norouzi, J. Shlens, Pixel recursive super resolution, CoRR abs/1702.00783.
URL http://arxiv.org/abs/1702.00783

[79] Z. Liu, P. Luo, X. Wang, X. Tang, Deep learning face attributes in the wild, in: Proc. 2015 IEEE International Conference on Computer Vision (ICCV), ICCV '15, IEEE Computer Society, Washington, DC, USA, 2015, pp. 3730–3738.

[80] F. Yu, Y. Zhang, S. Song, A. Seff, J. Xiao, LSUN: construction of a large-scale image dataset using deep learning with humans in the loop, CoRR abs/1506.03365.
URL http://arxiv.org/abs/1506.03365

[81] J. H. van Hateren, A. van der Schaaf, Independent component filters of natural images compared with simple cells in primary visual cortex, Proc. Royal Society of London B: Biological Sciences 265 (1394) (1998) 359–366. arXiv:http://rspb.royalsocietypublishing.org/content/265/1394/359.full.pdf, doi:10.1098/rspb.1998.0303.

[82] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft COCO: common objects in context, CoRR abs/1405.0312.
URL http://arxiv.org/abs/1405.0312

[83] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, L. Li, The new data and new challenges in multimedia research, CoRR abs/1503.01817.
URL http://arxiv.org/abs/1503.01817

[84] H. R. Sheikh, Z. Wang, L. Cormack, A. C. Bovik, Live image quality assessment database release 2, http://live.ece.utexas.edu/research/quality/subjective.htm, (Accessed on 06/12/2017).

[85] A. Fujimoto, T. Ogawa, K. Yamamoto, Y. Matsui, T. Yamasaki, K. Aizawa, Manga109 dataset and creation of metadata, in: Proceedings of the 1st International Workshop on coMics ANalysis, Processing and Understanding, MANPU '16, ACM, New York, NY, USA, 2016, pp. 2:1–2:5.

[86] I. Krasin, T. Duerig, N. Alldrin, V. Ferrari, S. Abu-El-Haija, A. Kuznetsova, H. Rom, J. Uijlings, S. Popov, S. Kamali, M. Malloci, J. Pont-Tuset, A. Veit, S. Belongie, V. Gomes, A. Gupta, C. Sun, G. Chechik, D. Cai, Z. Feng, D. Narayanan, K. Murphy, Openimages: A public dataset for large-scale multi-label and multi-class image classification., Dataset available from https://storage.googleapis.com/openimages/web/index.html.

[87] R. Timofte, E. Agustsson, L. V. Gool, M. H. Yang, L. Zhang, B. Lim, S. Son, H. Kim, S. Nah, K. M. Lee, X. Wang, Y. Tian, K. Yu, Y. Zhang, S. Wu, C. Dong, L. Lin, Y. Qiao, C. C. Loy, W. Bae, J. Yoo, Y. Han, J. C. Ye, J. S. Choi, M. Kim, Y. Fan, J. Yu, W. Han, D. Liu, H. Yu, Z. Wang, H. Shi, X. Wang, T. S. Huang, Y. Chen, K. Zhang, W. Zuo, Z. Tang, L. Luo, S. Li, M. Fu, L. Cao, W. Heng, G. Bui, T. Le, Y. Duan, D. Tao, R. Wang, X. Lin, J. Pang, J. Xu, Y. Zhao, X. Xu, J. Pan, D. Sun, Y. Zhang, X. Song, Y. Dai, X. Qin, X. P. Huynh, T. Guo, H. S. Mousavi, T. H. Vu, V. Monga, C. Cruz, K. Egiazarian, V. Katkovnik, R. Mehta, A. K. Jain, A. Agarwalla, C. V. S. Praveen, R. Zhou, H. Wen, C. Zhu, Z. Xia, Z. Wang, Q. Guo, NTIRE 2017 Challenge on Single Image Super-Resolution: Methods and Results, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017, pp. 1110–1121. doi:10.1109/CVPRW.2017.149.

[88] Z. Yang, K. Zhang, Y. Liang, J. Wang, Single Image Super-Resolution with a Parameter Economic Residual-Like Convolutional Neural Network, Springer International Publishing, 2017, pp. 353–364.

[89] C.-Y. Yang, C. Ma, M.-H. Yang, Single-Image Super-Resolution: A Benchmark, Springer International Publishing, Cham, 2014, pp. 372–386.

[90] M. Irani, S. Peleg, Improving resolution by image registration, CVGIP: Graph. Models Image Process. 53 (3) (1991) 231–239.

[91] K. I. Kim, Y. Kwon, Single-image super-resolution using sparse regression and natural image prior, IEEE Trans. Pattern Anal. Mach. Intell. 32 (6) (2010) 1127–1133.

[92] C.-Y. Yang, M.-H. Yang, Fast direct super-resolution by simple functions, in: Proceedings of IEEE International Conference on Computer Vision, 2013.

[93] J. Sun, Z. Xu, H.-Y. Shum, Image super-resolution using gradient profile prior, in: 2008 IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8. doi:10.1109/CVPR.2008.4587659.

[94] H. R. Sheikh, A. C. Bovik, G. de Veciana, An information fidelity criterion for image quality assessment using natural scene statistics, IEEE Transactions on Image Processing 14 (12) (2005) 2117–2128. doi:10.1109/TIP.2005.859389.

[95] M. Bevilacqua, A. Roumy, C. Guillemot, M. Alberi-Morel, Low-complexity single-image super-resolution based on nonnegative neighbor embedding, in: British Machine Vision Conference, BMVC 2012, Surrey, UK, September 3-7, 2012, 2012, pp. 1–10.

[96] R. Zeyde, M. Elad, M. Protter, On Single Image Scale-Up Using Sparse-Representations, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 711–730.

[97] M. Aharon, M. Elad, A. Bruckstein, K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation, Signal Processing, IEEE Transactions on 54 (11) (2006) 4311–4322.

[98] A. Kappeler, S. Yoo, Q. Dai, A. K. Katsaggelos, Video Super-Resolution with Convolutional Neural Networks, Computational Imaging, IEEE Transactions on 2 (2016) 109–122.

[99] J. Yamanaka, S. Kuwashima, T. Kurita, Fast and accurate image super resolution by deep CNN with skip connection and network in network, CoRR abs/1707.05425.
URL http://arxiv.org/abs/1707.05425

[100] M. Lin, Q. Chen, S. Yan, Network in network, CoRR abs/1312.4400.
URL http://arxiv.org/abs/1312.4400

[101] Z. Cui, H. Chang, S. Shan, B. Zhong, X. Chen, Deep Network Cascade for Image Super-resolution, Springer International Publishing, Cham, 2014, pp. 49–64.

[102] A. Greaves, H. Winter, Multi-frame video super-resolution using convolutional neural networks, http://cs231n.stanford.edu/reports/2016/pdfs/212_Report.pdf, accessed: 2017-04-16 (2016).

[103] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, CoRR abs/1409.1556.
URL http://arxiv.org/abs/1409.1556

[104] M. Knoche, Super-resolution - convolutional neural networks for image and video processing - tum wiki, https://wiki.tum.de/display/lfdv/Super-Resolution, (Accessed on 06/20/2017) (Feb 2017).

[105] C.-Y. Lee, S. Xie, P. W. Gallagher, Z. Zhang, Z. Tu, Deeply-supervised nets., in: G. Lebanon, S. V. N. Vishwanathan (Eds.), AISTATS, Vol. 38, JMLR.org, 2015.

[106] E. Shelhamer, J. Long, T. Darrell, Fully convolutional networks for semantic segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (4) (2017) 640–651.

[107] Z. Wang, D. Liu, J. Yang, W. Han, T. Huang, Deep networks for image super-resolution with sparse prior, in: 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 370–378. doi:10.1109/ICCV.2015.50.

[108] K. Gregor, Y. LeCun, Learning fast approximations of sparse coding, 2010, pp. 399–406.

[109] D. Liu, Z. Wang, B. Wen, J. Yang, W. Han, T. S. Huang, Robust single image super-resolution via deep networks with sparse prior, IEEE Transactions on Image Processing 25 (7) (2016) 3194–3207. doi:10.1109/TIP.2016.2564643.

[110] D. Liu, Z. Wang, N. M. Nasrabadi, T. S. Huang, Learning a mixture of deep networks for single image super-resolution, CoRR abs/1701.00823.

[111] Y. Liang, R. Timofte, J. Wang, Y. Gong, N. Zheng, Single image super resolution - when model adaptation matters, CoRR abs/1703.10889.
URL http://arxiv.org/abs/1703.10889

[112] A. Shocher, N. Cohen, M. Irani, "zero-shot" super-resolution using deep internal learning, CoRR abs/1712.06087. arXiv:1712.06087.
URL http://arxiv.org/abs/1712.06087

[113] B. Lim, S. Son, H. Kim, S. Nah, K. M. Lee, Enhanced deep residual networks for single image super-resolution, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017, pp. 1132–1140. doi:10.1109/CVPRW.2017.151.

[114] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, W. Shi, Photo-realistic single image super-resolution using a generative adversarial network, CoRR abs/1609.04802.

[115] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[116] S. Nah, T. Hyun Kim, K. Mu Lee, Deep multi-scale convolutional neural network for dynamic scene deblurring, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[117] A. van den Oord, N. Kalchbrenner, O. Vinyals, L. Espeholt, A. Graves, K. Kavukcuoglu, Conditional image generation withPixelCNN decoders, CoRR abs/1606.05328.
URL http://arxiv.org/abs/1606.05328

[118] A. van den Oord, N. Kalchbrenner, K. Kavukcuoglu, Pixel recurrent neural networks (2016) 1747–1756.
URL http://dl.acm.org/citation.cfm?id=3045390.3045575

[119] Z. Wang, Y. Yang, Z. Wang, S. Chang, W. Han, J. Yang, T. Huang, Self-tuned deep super resolution, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2015, pp. 1–8. doi:10.1109/CVPRW.2015.7301266.

[120] J. Masci, U. Meier, D. Cireşan, J. Schmidhuber, Stacked Convolutional Auto-Encoders for Hierarchical Feature Extraction, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011, pp. 52–59.

[121] X. Mao, C. Shen, Y. Yang, Image denoising using very deep fully convolutional encoder-decoder networks with symmetric skip connections, CoRR abs/1603.09056.
URL http://arxiv.org/abs/1603.09056

[122] H. Mao, Y. Wu, J. Li, Y. Fu, Super resolution of the partial pixelated images with deep convolutional neural network, in: Proc. 2016 ACM on Multimedia Conference, MM '16, ACM, New York, NY, USA, 2016, pp. 322–326.

[123] A. Makhzani, J. Shlens, N. Jaitly, I. J. Goodfellow, Adversarial autoencoders, CoRR abs/1511.05644.

[124] L. Xu, J. S. J. Ren, C. Liu, J. Jia, Deep convolutional neural network for image deconvolution, in: Proc. 27th International Conference on Neural Information Processing Systems, NIPS'14, MIT Press, Cambridge, MA, USA, 2014, pp. 1790–1798.
URL http://dl.acm.org/citation.cfm?id=2968826.2969026

[125] Y. Shi, K. Wang, L. Xu, L. Lin, Local- and holistic- structure preserving image super resolution via deep joint component learning, CoRR abs/1607.07220.
URL http://arxiv.org/abs/1607.07220

[126] C. Peyrard, F. Mamalet, C. Garcia, A comparison between multi-layer perceptrons and convolutional neural networks for text image super-resolution, in: VISAPP 2015 - Proc. 10th International Conference on Computer Vision Theory and Applications, Volume 1, Berlin, Germany, 11-14 March, 2015., 2015, pp. 84–91.

[127] F. Pan, L. Zhang, New image super-resolution scheme based on residual error restoration by neural networks, Optical Engineering 42 (2003) 3038–3046. doi:10.1117/1.1604397.

[128] C. Peyrard, M. Baccouche, C. Garcia, Blind Super-Resolution with Deep Convolutional Neural Networks, Springer International Publishing, Cham, 2016, pp. 161–169.

[129] K. Hayat, W. Puech, N. Islam, G. Gesquière, Seamless heterogeneous 3d tessellation via dwt domain smoothing and mosaicking, EURASIP Journal on Advances in Signal Processing 2010 (1) (2010) 913681.

[130] W. Shi, J. Caballero, L. Theis, F. Huszar, A. P. Aitken, C. Ledig, Z. Wang, Is the deconvolution layer the same as a convolutional layer?, CoRR abs/1609.07009.
URL http://arxiv.org/abs/1609.07009

[131] M. D. Zeiler, G. W. Taylor, R. Fergus, Adaptive deconvolutional networks for mid and high level feature learning, in: 2011 International Conference on Computer Vision, 2011, pp. 2018–2025. doi:10.1109/ICCV.2011.6126474.

[132] Y. Zhao, R. Wang, W. Dong, W. Jia, J. Yang, X. Liu, W. Gao, Gun: Gradual upsampling network for single image super-resolution, CoRR abs/1703.0424.

[133] J. S. Ren, L. Xu, Q. Yan, W. Sun, Shepard convolutional neural networks, in: C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, R. Garnett (Eds.), Advances in Neural Information Processing Systems 28, Curran Associates, Inc., 2015, pp. 901–909.
URL http://papers.nips.cc/paper/5774-shepard-convolutional-neural-networks.pdf

[134] D. Shepard, A two-dimensional interpolation function for irregularly-spaced data, in: Proc. 1968 23rd ACM National Conference, ACM '68, ACM, New York, NY, USA, 1968, pp. 517–524. doi:10.1145/800186.810616.
URL http://doi.acm.org/10.1145/800186.810616

[135] J. Bruna, P. Sprechmann, Y. LeCun, Super-resolution with deep convolutional sufficient statistics.

[136] J. Bruna, S. Mallat, Invariant scattering convolution networks, IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (8) (2013) 1872–1886. doi:10.1109/TPAMI.2012.230.

[137] M. Cimpoi, S. Maji, A. Vedaldi, Deep filter banks for texture recognition and segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[138] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255. doi:10.1109/CVPR.2009.5206848.

[139] Y. Wang, L. Wang, H. Wang, P. Li, End-to-end image super-resolution via deep and shallow convolutional networks, CoRR abs/1607.07680.
URL http://arxiv.org/abs/1607.07680

[140] X. Du, X. Qu, Y. He, D. Guo, Single image super-resolution based on multi-scale competitive convolutional neural network, Sensors 18 (3).
URL http://www.mdpi.com/1424-8220/18/3/789

[141] S. Li, R. Fan, G. Lei, G. Yue, C. Hou, A two-channel convolutional neural network for image super-resolution, Neurocomputing 275 (2018) 267 – 277.
URL http://www.sciencedirect.com/science/article/pii/S0925231217314625

[142] Y. Kato, S. Ohtani, N. Kuroki, T. Hirose, M. Numa, Image super-resolution with multi-channel convolutional neural networks, in: 2016 14th IEEE International New Circuits and Systems Conference (NEWCAS), 2016, pp. 1–4. doi:10.1109/NEWCAS.2016.7604756.

[143] S. OHTANI, Y. KATO, N. KUROKI, T. HIROSE, M. NUMA, Multi-channel convolutional neural networks for image super-resolution, IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences E100.A (2) (2017) 572–580. doi:10.1587/transfun.E100.A.572.

[144] X. Gao, H. Xiong, A hybrid wavelet convolution network with sparse-coding for image super-resolution, in: 2016 IEEE International Conference on Image Processing (ICIP), 2016, pp. 1439–1443. doi:10.1109/ICIP.2016.7532596.

[145] S. Mallat, Group invariant scattering, Communications on Pure and Applied Mathematics 65 (10) (2012) 1331–1398.

[146] G. Y. Youm, S. H. Bae, M. Kim, Image super-resolution based on convolution neural networks using multi-channel input, in: 2016 IEEE 12th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP), 2016, pp. 1–5. doi:10.1109/IVMSPW.2016.7528224.

[147] C. Osendorfer, H. Soyer, P. van der Smagt, Image Super-Resolution with Fast Approximate Convolutional Sparse Coding, Springer International Publishing, Cham, 2014, pp. 250–257.

[148] G. Riegler, S. Schulter, M. Rüther, H. Bischof, Conditioned regression models for non-blind single image super-resolution, in: 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 522–530. doi:10.1109/ICCV.2015.67.

[149] J. Johnson, A. Alahi, F. Li, Perceptual losses for real-time style transfer and super-resolution, CoRR abs/1603.08155.
URL http://arxiv.org/abs/1603.08155

[150] Y. Qu, C. Shi, J. Liu, L. Peng, X. Du, Single Image Super-Resolution via Convolutional Neural Network and Total Variation Regularization, Springer International Publishing, Cham, 2016, pp. 28–38. doi:10.1007/978-3-319-27674-8_3.
URL http://dx.doi.org/10.1007/978-3-319-27674-8_3

[151] Y. Tang, X. Zhu, M. Cui, Convolutional neural network with gradient information for image super-resolution, in: 2016 IEEE International Conference on Information and Automation (ICIA), 2016, pp. 1714–1719. doi:10.1109/ICInfA.2016.7832094.

[152] T. Brosch, R. Tam, Efficient training of convolutional deep belief networks in the frequency domain for application to high-resolution 2d and 3d images, Neural Comput. 27 (1) (2015) 211–227.

[153] W.-S. Lai, J.-B. Huang, N. Ahuja, M.-H. Yang, Deep laplacian pyramid networks for fast and accurate super-resolution (2017).

[154] H. Ren, M. El-Khamy, J. Lee, Image super resolution based on fusing multiple convolution neural networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017, pp. 1050–1057. doi:10.1109/CVPRW.2017.142.

[155] P. Rasti, T. Uiboupin, S. Escalera, G. Anbarjafari, Convolutional Neural Network Super Resolution for Face Recognition in Surveillance Monitoring, Springer International Publishing, Cham, 2016, pp. 175–184.

[156] H. Zhang, P. Casaseca-de-la Higuera, C. Luo, Q. Wang, M. Kitchin, A. Parmley, J. Monge-Alvarez, Systematic infrared image quality improvement using deep learning based techniques (2016).

[157] K. Umehara, J. Ota, T. Ishida, Application of super-resolution convolutional neural network for enhancing image resolution in chest ct, Journal of Digital Imaging.
URL https://doi.org/10.1007/s10278-017-0033-z

[158] X. Huang, Y. Jiang, X. Liu, H. Xu, Z. Han, H. Rong, H. Yang, M. Yan, H. Yu, Machine learning based single-frame super-resolution processing for lensless blood cell counting, Sensors 16 (11).

[159] G.-B. Huang, Q.-Y. Zhu, C.-K. Siew, Extreme learning machine: Theory and applications, Neurocomputing 70 (1–3) (2006) 489 – 501, neural NetworksSelected Papers from the 7th Brazilian Symposium on Neural Networks (SBRN '04)7th Brazilian Symposium on Neural Networks.

[160] E. Simo-Serra, S. Iizuka, K. Sasaki, H. Ishikawa, Learning to simplify: fully convolutional networks for rough sketch cleanup, ACM Trans. Graph. 35 (4) (2016) 121.

[161] Y. Yoon, H. G. Jeon, D. Yoo, J. Y. Lee, I. S. Kweon, Learning a deep convolutional network for light-field image super-resolution, in: 2015 IEEE International Conference on Computer Vision Workshop (ICCVW), 2015, pp. 57–65. doi:10.1109/ICCVW.2015.17.

[162] S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. J. Black, R. Szeliski, A database and evaluation methodology for optical flow, International Journal of Computer Vision 92 (1) (2011) 1–31. doi:10.1007/s11263-010-0390-2.
URL http://dx.doi.org/10.1007/s11263-010-0390-2

[163] C. Liu, D. Sun, On bayesian adaptive video super resolution, IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (2) (2014) 346–360. doi:10.1109/TPAMI.2013.127.

[164] R. Liao, X. Tao, R. Li, Z. Ma, J. Jia, Video super-resolution via deep draft-ensemble learning, in: Proc. 2015 IEEE International Conference on Computer Vision (ICCV), ICCV '15, IEEE Computer Society, Washington, DC, USA, 2015, pp. 531–539. doi:10.1109/ICCV.2015.68.
URL http://dx.doi.org/10.1109/ICCV.2015.68

[165] D. J. Butler, J. Wulff, G. B. Stanley, M. J. Black, A naturalistic open source movie for optical flow evaluation, in: A. Fitzgibbon et al. (Eds.) (Ed.), European Conf. on Computer Vision (ECCV), Part IV, LNCS 7577, Springer-Verlag, 2012, pp. 611–625.

[166] J. Wulff, D. J. Butler, G. B. Stanley, M. J. Black, Lessons and insights from creating a synthetic optical flow benchmark, in: A. Fusiello et al. (Eds.) (Ed.), ECCV Workshop on Unsolved Problems in Optical Flow and Stereo Estimation, Part II, LNCS 7584, Springer-Verlag, 2012, pp. 168–177.

[167] F. Galasso, N. S. Nagaraja, T. J. Cárdenas, T. Brox, B. Schiele, A unified video segmentation benchmark: Annotation, metrics and analysis, in: 2013 IEEE International Conference on Computer Vision, 2013, pp. 3527–3534. doi:10.1109/ICCV.2013.438.

[168] H. Takeda, P. Milanfar, M. Protter, M. Elad, Super-resolution without explicit subpixel motion estimation, IEEE Transactions on Image Processing 18 (9) (2009) 1958–1975. doi:10.1109/TIP.2009.2023703.

[169] M.-H. Cheng, N.-W. Lin, K.-S. Hwang, J.-H. Jeng, Fast video super-resolution using artificial neural networks, in: 2012 8th International Symposium on Communication Systems, Networks Digital Signal Processing (CSNDSP), 2012, pp. 1–4. doi:10.1109/CSNDSP.2012.6292646.

[170] Z. Ma, R. Liao, X. Tao, L. Xu, J. Jia, E. Wu, Handling motion blur in multi-frame super-resolution, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5224–5232. doi:10.1109/CVPR.2015.7299159.

[171] Infognition, Video enhancer: resize video with super resolution method from sd to hd, http://www.infognition.com/VideoEnhancer/, (Accessed on 06/16/2017) (2010).

[172] A. Kappeler, S. Yoo, Q. Dai, A. K. Katsaggelos, Super-resolution of compressed videos using convolutional neural networks, in: 2016 IEEE International Conference on Image Processing (ICIP), 2016, pp. 1150–1154. doi:10.1109/ICIP.2016.7532538.

[173] A. Gupta, P. Bhat, M. Dontcheva, B. Curless, O. Deussen, M. Cohen, Enhancing and experiencing spacetime resolution with videos and stills, in: International Conference on Computational Photography, IEEE, 2009.
URL http://grail.cs.washington.edu/projects/enhancing-spacetime/

[174] D. Zeng, Convolutional neural network implementation of superresolution video, http://cs231n.stanford.edu/reports/2016/pdfs/203_Report.pdf, accessed: 2017-04-06 (2016).

[175] A. Dosovitskiy, P. Fischer, E. Ilg, P. Häusser, C. Hazirbas, V. Golkov, P. van der Smagt, D. Cremers, T. Brox, Flownet: Learning optical flow with convolutional networks, in: 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015, IEEE Computer Society, 2015, pp. 2758–2766.

[176] Y. Huang, W. Wang, L. Wang, Bidirectional recurrent convolutional networks for multi-frame super-resolution, in: Proc. 28th International Conference on Neural Information Processing Systems, NIPS'15, MIT Press, Cambridge, MA, USA, 2015, pp. 235–243.

[177] J. Caballero, C. Ledig, A. P. Aitken, A. Acosta, J. Totz, Z. Wang, W. Shi, Real-time video super-resolution with spatio-temporal networks and motion compensation, CoRR abs/1611.05250.
URL http://arxiv.org/abs/1611.05250

[178] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei, Large-scale video classification with convolutional neural networks, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1725–1732. doi:10.1109/CVPR.2014.223.

[179] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3d convolutional networks, in: Proc. 2015 IEEE International Conference on Computer Vision (ICCV), ICCV '15, IEEE Computer Society, Washington, DC, USA, 2015, pp. 4489–4497.

[180] K. Nasrollahi, S. Escalera, P. Rasti, G. Anbarjafari, X. Baro, H. J. Escalante, T. B. Moeslund, Deep learning based super-resolution for improved action recognition, in: Image Processing Theory, Tools and Applications (IPTA), 2015 International Conference on, 2015, pp. 67–72. doi:10.1109/IPTA.2015.7367098.

[181] H. Wang, C. Schmid, Action recognition with improved trajectories, in: 2013 IEEE International Conference on Computer Vision, 2013, pp. 3551–3558. doi:10.1109/ICCV.2013.441.

[182] X. Tao, H. Gao, R. Liao, J. Wang, J. Jia, Detail-revealing Deep Video Super-resolution, CoRR abs/1704.02738. arXiv:1704.02738.
URL http://arxiv.org/abs/1704.02738

[183] D. Liu, Z. Wang, Y. Fan, X. Liu, Z. Wang, S. Chang, X. Wang, T. S. Huang, Learning temporal dynamics for video super-resolution: A deep learning approach, IEEE Transactions on Image Processing 27 (7) (2018) 3432–3445. doi:10.1109/TIP.2018.2820807.

[184] D. Liu, Z. Wang, Y. Fan, X. Liu, Z. Wang, S. Chang, T. Huang, Robust video super-resolution with learned temporal dynamics, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2507–2515.

[185] W. Yang, J. Feng, G. Xie, J. Liu, Z. Guo, S. Yan, Video super-resolution based on spatial-temporal recurrent residual networks, Computer Vision and Image Understanding 168 (2018) 79 – 92, special Issue on Vision and Computational Photography and Graphics. doi:https://doi.org/10.1016/j.cviu.2017.09.002.
URL http://www.sciencedirect.com/science/article/pii/S1077314217301583

[186] W. Yang, J. Liu, Z. Guo, Spatial-temporal recurrent residual networks for video super-resolution, in: G. Zhai, J. Zhou, X. Yang (Eds.), Digital TV and Wireless Multimedia Communication, Springer Singapore, Singapore, 2018, pp. 115–127.

[187] L. Linghui, D. Junping, L. Meiyu, R. Nan, F. Dan, Video super-resolution reconstruction based on deep convolutional neural network and spatio-temporal similarity, The Journal of China Universities of Posts and Telecommunications 23 (5) (2016) 68 – 81. doi:https://doi.org/10.1016/S1005-8885(16)60060-2.
URL http://www.sciencedirect.com/science/article/pii/S1005888516600602

[188] B. Lim, K. M. Lee, Deep Recurrent Resnet for Video Super-Resolution, in: Proc. APSIPA Annual Summit and Conference 2017, 2017, p. 4.
URL http://www.apsipa.org/proceedings/2017/CONTENTS/papers2017/15DecFriday/FA-06/FA-06.5.pdf

[189] D. Eigen, C. Puhrsch, R. Fergus, Depth map prediction from a single image using a multi-scale deep network, in: Proc. 27th International Conference on Neural Information Processing Systems, NIPS'14, MIT Press, Cambridge, MA, USA, 2014, pp. 2366–2374.

[190] X. Ma, Z. Geng, Z. Bie, Depth Estimation from Single Image Using CNN-Residual Network, http://cs231n.stanford.edu/reports/2017/pdfs/203.pdf, (Accessed on 04/29/2018) (2017).

[191] D. Scharstein, R. Szeliski, A taxonomy and evaluation of dense two-frame stereo correspondence algorithms, International Journal of Computer Vision 47 (1) (2002) 7–42.

[192] D. Scharstein, R. Szeliski, High-accuracy stereo depth maps using structured light, in: Proc. 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR'03, IEEE Computer Society, Washington, DC, USA, 2003, pp. 195–202.

[193] H. Hirschmuller, D. Scharstein, Evaluation of cost functions for stereo matching, in: 2007 IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8. doi:10.1109/CVPR.2007.383248.

[194] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nesic, X. Wang, P. Westling, High-resolution stereo datasets with subpixel-accurate ground truth., in: X. Jiang, J. Hornegger, R. Koch (Eds.), GCPR, Vol. 8753 of Lecture Notes in Computer Science, Springer, 2014, pp. 31–42.

[195] X. Song, Y. Dai, X. Qin, Deep Depth Super-Resolution: Learning Depth Super-Resolution Using Deep Convolutional Neural Network, Springer International Publishing, Cham, 2017, pp. 360–376. doi:10.1007/978-3-319-54190-7_22.
URL http://dx.doi.org/10.1007/978-3-319-54190-7_22

[196] O. Mac Aodha, N. D. F. Campbell, A. Nair, G. J. Brostow, Patch Based Synthesis for Single Depth Image Super-Resolution, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 71–84.

[197] A. Handa, T. Whelan, J. McDonald, A. Davison, A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM, in: IEEE Intl. Conf. on Robotics and Automation, ICRA, Hong Kong, China, 2014.

[198] N. Silberman, D. Hoiem, P. Kohli, R. Fergus, Indoor segmentation and support inference from rgbd images, in: Proc. 12th European Conference on Computer Vision - Volume Part V, ECCV'12, Springer-Verlag, Berlin, Heidelberg, 2012, pp. 746–760.

[199] A. Geiger, P. Lenz, C. Stiller, R. Urtasun, Vision meets robotics: The KITTI dataset, Int. J. Rob. Res. 32 (11) (2013) 1231–1237.

[200] D. Ferstl, C. Reinbacher, R. Ranftl, M. Ruether, H. Bischof, Image guided depth upsampling using anisotropic total generalized variation, in: Proceedings International Conference on Computer Vision (ICCV), IEEE, 2013.

[201] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, R. L. Buckner, Open access series of imaging studies (oasis): Cross-sectional mri data in young, middle aged, nondemented, and demented older adults, J. Cognitive Neuroscience 19 (9) (2007) 1498–1507. doi:10.1162/jocn.2007.19.9.1498.
URL http://dx.doi.org/10.1162/jocn.2007.19.9.1498

[202] K. He, J. Sun, X. Tang, Guided image filtering, IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (6) (2013) 1397–1409. doi:10.1109/TPAMI.2012.213.

[203] J. Diebel, S. Thrun, An application of markov random fields to range sensing, in: Y. Weiss, B. Schölkopf, J. Platt (Eds.), Advances in Neural Information Processing Systems 18, MIT Press, Cambridge, MA, 2005, pp. 291–298.
URL http://books.nips.cc/papers/files/nips18/NIPS2005_0707.pdf

[204] J. Park, H. Kim, Y.-W. Tai, M. S. Brown, I. Kweon, High quality depth map upsampling for 3d-tof cameras, in: 2011 International Conference on Computer Vision, 2011, pp. 1623–1630. doi:10.1109/ICCV.2011.6126423.

[205] J. Xie, R. S. Feris, M. T. Sun, Edge-guided single depth image super resolution, IEEE Transactions on Image Processing 25 (1) (2016) 428–438. doi:10.1109/TIP.2015.2501749.

[206] M. Hornacek, C. Rhemann, M. Gelautz, C. Rother, Depth super resolution by rigid body self-similarity in 3d, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013, CVPR'13, pp. 1123–1130.

[207] J. Xie, R. S. Feris, M. T. Sun, Edge guided single depth image super resolution, in: 2014 IEEE International Conference on Image Processing (ICIP), 2014, pp. 3773–37777. doi:10.1109/ICIP.2014.7025766.

[208] D. Ferstl, M. Rüther, H. Bischof, Variational depth superresolution using example-based edge representations, in: 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 513–521. doi:10.1109/ICCV.2015.66.

[209] Y. Xie, J. Xiao, T. Tillo, Y. Wei, Y. Zhao, 3d video super-resolution using fully convolutional neural networks, in: 2016 IEEE International Conference on Multimedia and Expo (ICME), 2016, pp. 1–6. doi:10.1109/ICME.2016.7552931.

[210] Z. Jin, T. Tillo, C. Yao, J. Xiao, Y. Zhao, Virtual-view-assisted video super-resolution and enhancement, IEEE Transactions on Circuits and Systems for Video Technology 26 (3) (2016) 467–478. doi:10.1109/TCSVT.2015.2412791.

[211] Z. Xu, X. Wang, Z. Chen, D. Xiong, M. Ding, W. Hou, Nonlocal similarity based {DEM} super resolution, ISPRS Journal of Photogrammetry and Remote Sensing 110 (2015) 48 – 54.

[212] Z. Chen, X. Wang, Z. Xu, W. Hou, Convolutional Neural Network Based DEM Super Resolution, ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (2016) 247–250doi:10.5194/isprs-archives-XLI-B3-247-2016.

[213] K. Bredies, K. Kunisch, T. Pock, Total generalized variation, SIAM Journal on Imaging Sciences 3 (3) (2010) 492–526.

[214] G. Riegler, D. Ferstl, M. Rüther, H. Bischof, A deep primal-dual network for guided depth super-resolution, CoRR abs/1607.08569.
URL http://arxiv.org/abs/1607.08569

[215] C. W. Tseng, H. R. Su, S. H. Lai, J. Liu, Depth image super-resolution via multi-frame registration and deep learning, in: 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2016, pp. 1–8. doi:10.1109/APSIPA.2016.7820834.

[216] S. Moon, H.-L. Choi, Super resolution based on deep learning technique for constructing digital elevation model, in: AIAA SPACE 2016, 2016, p. 5608.

[217] T.-W. Hui, C. C. Loy, X. Tang, Depth map super-resolution by deep multi-scale guidance, in: European Conference on Computer Vision (ECCV), 2016.
URL http://mmlab.ie.cuhk.edu.hk/projects/guidance_SR_depth.html

[218] Y. Xiao, X. Cao, X. Zhu, R. Yang, Y. Zheng, Joint convolutional neural pyramid for depth map super-resolution, CoRR abs/1801.00968.
URL http://arxiv.org/abs/1801.00968

[219] X. Shen, Y. Chen, X. Tao, J. Jia, Convolutional neural pyramid for image processing, CoRR abs/1704.02071. arXiv:1704.02071.
URL http://arxiv.org/abs/1704.02071

[220] B. Chen, C. Jung, Single Depth Image Super-Resolution using Convolutional Neural Networks.
URL http://sigport.org/2577

[221] M. Ni, J. Lei, R. Cong, K. Zheng, B. Peng, X. Fan, Color-guided depth map super resolution using convolutional neural network, IEEE Access 5 (2017) 26666–26672. doi:10.1109/ACCESS.2017.2773141.

[222] W. Zhou, X. Li, D. Reynolds, Guided deep network for depth map super-resolution: How much can color help?, in: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 1457–1461.

[223] L. Liebel, M. Körner, Single-Image Super Resolution for Multispectral Remote Sensing Data Using Convolutional Neural Networks, ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (2016) 883–890doi:10.5194/isprs-archives-XLI-B3-883-2016.

[224] O. Oktay, W. Bai, M. Lee, R. Guerrero, K. Kamnitsas, J. Caballero, A. de Marvao, S. Cook, D. O'Regan, D. Rueckert, Multi-input Cardiac Image Super-Resolution Using Convolutional Neural Networks, Springer International Publishing, Cham, 2016, pp. 246–254.

[225] Y. Li, W. Xie, H. Li, Hyperspectral image reconstruction by deep convolutional neural network for classification, Pattern Recognition 63 (2017) 371 – 383.

[226] P. Hagerty, Super-resolution on satellite imagery using deep learning, part 3, https://medium.com/the-downlinq/super-resolution-on-satellite-imagery-using-deep-learning-part-3-2e2f61eee1d3, (Accessed on 06/19/2017) (Mar 2017).

[227] K. Hayat, W. Puech, G. Gesquière, Adaptively synchronous scalable spread spectrum (a4s) data-hiding strategy for three-dimensional visualization, Journal of Electronic Imaging 19 (2) (2010) 023011–023011–16.

[228] Y. Fang, C. Zhang, Convolutional neural network for blind quality evaluator of image super-resolution, in: 2017 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), 2017, pp. 28–33.

[229] S. Shalev-Shwartz, O. Shamir, S. Shammah, Failures of deep learning, CoRR abs/1703.07950v1, accessed: 2017-03-29.
URL https://arxiv.org/pdf/1703.07950v1.pdf

[230] M. Yao, Understanding the limits of deep learning, https://venturebeat.com/2017/04/02/understanding-the-limits-of-deep-learning/, (Accessed on 06/21/2017) (Apr 2017).

Biographical Notes

Khizar Hayat is currently heading the Department of Mathematical and Physical Sciences (DMPS) at the College of Arts and Sciences, University of Nizwa, Oman. He has also led the Computer Science Department of COMSATS Institute of Information Technology, Abbottabad, Pakistan. He received his PhD degree in 2009 from the University of Montpellier 2, France, while working at LIRMM. His preference areas are image processing and information hiding.