Going Much Wider with Deep Networks for Image Super-Resolution

Vikram Singh Keerthan Ramnath* Subrahmanyam Arunachalam* Anurag Mittal
Computer Vision Lab, Indian Institute of Technology-Madras
{vsingh@cse,keerthan@smail,amittal@cse}.iitm.ac.in, 110116013@nitt.edu

Abstract

Divide and Conquer is a well-established approach in the literature that has efficiently solved a variety of problems. However, it is yet to be explored in full in solving image super-resolution. To predict a sharp up-sampled image, this work proposes a divide and conquer approach based wide and deep network (WDN) that divides the 4× up-sampling problem into 32 disjoint subproblems that can be solved simultaneously and independently of each other. Half of these subproblems deal with predicting the overall features of the high-resolution image, while the remaining are exclusively for predicting the finer details. Additionally, a technique that is found to be more effective in calibrating the pixel intensities has been proposed. Results obtained on multiple datasets demonstrate the improved performance of the proposed wide and deep network over state-of-the-art methods.

1. Introduction

Image Super-Resolution refers to those set of techniques that increase the resolution of an image while maintaining its quality that is commonly measured in terms of Peak Signal to Noise Ratio (PSNR [29]), and Structural Similarity (SSIM [97]) w.r.t. the ground-truth. Recent advances in display device technologies (Full HD and higher resolutions) have brought a surge in the application of super-resolution thus making it a prominent Computer Vision problem that is gaining immense academic and commercial research interests.

Most of the state-of-the-art super-resolution techniques (as discussed in Section 1.1), follow the conventional principle of 'building a deeper network architecture and training it on large datasets', for instance, the deeply recursive architecture proposed by *Kim et al.* [36], residual-in-residual architecture proposed by *Zhang et al.* [91], and multi-scale residual architecture proposed by *Li et al.* [40]. These methods have outperformed classic super-resolution techniques, but they still have a significant scope for improvement.

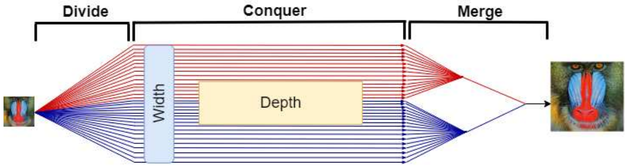


Figure 1: Illustration of the proposed Wide and Deep Network (WDN) for 4× up-sampling. Red lines indicate deep sub-networks for all (low+high) frequency prediction. Blue lines indicate sub-networks for high-frequency prediction. Width: 32, Depth: 270 layers

This work attempts to improve the super-resolution performance by proposing an alternate design of a deep neural network that works on the principle of divide and conquer and is much wider than existing state-of-the-art networks. We name our network as *WDN* due to its wide and deep architectural design. Such a design has the advantage of parallel execution, as wider networks can utilise multiple processing units in parallel. Moreover, the divide and conquer approach also gives an advantage of better learning on wider networks as with this approach a complex problem can be divided into multiple subproblems and then each sub-network along the width of the wide network can be trained to solve a specific subproblem. This eventually leads to increased expertise of the sub-networks in solving subproblems of one particular type, ultimately improving the overall performance.

As illustrated in Figure 1, we divide the 4× up-sampling problem into 32 disjoint subproblems. Half of these subproblems are of predicting the all-frequency (low and high-frequency) details of different disjoint parts of the ground-truth image, while the remaining half are exclusively for predicting the high-frequency details (fine details) of those parts. Half of the subproblems have been created for high-frequency prediction as it is one of the core problems of super-resolution that requires special attention and whose solution brings sharpness in the final result. The expected

*Equal Contribution. Keerthan Ramnath from Chemical Engg., IIT-Madras and Subrahmanyam Arunachalam from I & C Engg., NIT-Tiruchirappalli were interns at Computer Vision Lab, IIT-Madras when this work was performed.

outcome of the model is generated as a weighted function that merges the computed 32 sub-solutions.

Apart from a wider design, this work also introduces a technique (inspired by *Srivastava et al.* [63]) to calibrate the negative pixels of the feature maps that are generated within the network. This technique improves the overall performance of the network by calibrating the pixel intensities based on a self-learned pixel relevance value; here, the relevance indicates the relevance of the pixels in the upsampling task.

1.1. Related work

Super-resolution has garnered much attention in the recent past, especially with the advent of deep neural networks, and this has resulted in numerous works on super-resolution in a very short period.

For instance, [58, 32, 87, 33, 36, 58, 15, 1, 64, 43, 26, 16, 12, 70, 92, 40, 53, 79, 22, 57, 59, 4, 60, 78, 66] are some deep networks for super-resolution. These approaches map the low-resolution input to high-resolution output using a variety of different and novel architectures such as sub-pixel convolutional network, recursive convolutional network and residual networks. Unlike our network WDN, they neither recognise the importance of high-frequency prediction nor take explicit measures to predict it.

There are a few deep networks that do realise the importance of high-frequency prediction such as [54, 39, 67, 83, 48, 86, 6, 7, 88, 80, 11, 44, 75, 76, 94], these techniques use the concepts of generative adversarial networks, perceptual loss, or both. Application of these concepts implicitly work towards the prediction of high-frequency details, subsequently generating realistic-looking results. Perceptual loss attempts to minimise the difference between the model prediction and ground-truth in some feature space that is created by another convolutional network such as *VGG-16* [61]. On the other hand, adversarial networks focus on training, until the network can bluff a trained discriminator in distinguishing between the network's prediction and the actual ground-truth. WDN's approach in generating realistic results is entirely different from these approaches. WDN neither has a discriminator nor perceptual loss. It makes a direct attempt to explicitly predict the fine textures in the upsampled image using wide and deep convolutional neural network architecture.

Apart from the use of adversarial training and perceptual loss, some methods have developed other architectures as well that have helped in predicting the high-frequency details. For instance, [35, 13, 38, 74, 17, 21, 82, 28, 91, 23, 9, 90, 34, 19, 37, 49, 77, 62, 14] have proposed and used residual architectures to preserve the high-frequency details. Such architectures also help in avoiding vanishing gradients, subsequently helping in building and training deeper network. Some other approaches, for instance [84, 81, 65, 41], have gone a step ahead and developed recurrent-residual architectures to improve the results further. Recurrent-residual architectures help in addressing the long-term dependency problem, that causes the initial layers to have little influence on the deeper ones. In addition to the residual and recurrent residual architecture-based approaches, the methods such as [20, 93, 3, 52, 2, 95, 47] have come up with a variety of architectural design such as recurrent, cascaded, encoder-decoder, or ensemble-based; to preserve the high-frequency details. Lastly, methods that have attempted to increase their width to some extent for solving super-resolution are [50, 8, 51, 42, 56, 30, 46, 85, 10, 96, 31, 55, 24].

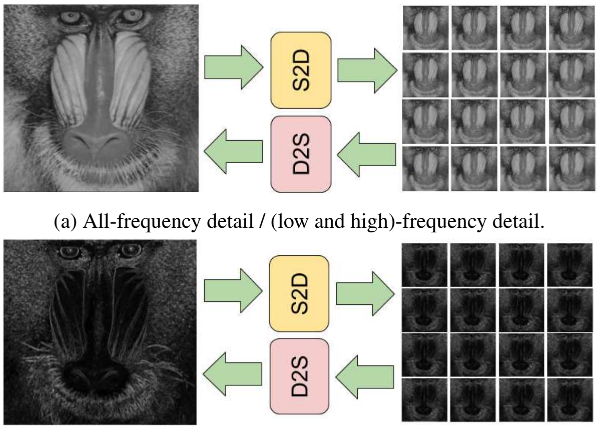
Unlike our network WDN, none of the cited techniques above has progressed in the direction of 'wide networks with divide and conquer principle'. The ones that have attempted to build wide networks have an entirely different working principle as compared to ours. In this work, we divide the single problem of predicting an up-sampled image into multiple subproblems of predicting the disjoint segments of the up-sampled image and their corresponding high-frequency details. We conquer/solve all the subproblems and merge their sub-solutions to generate the final output. Our work takes care of high-frequency details 'explicitly' and has separate networks exclusively for its prediction. *To the best of our knowledge, such ideas have not been used in any of the cited works*, and as we show later in Section 3, these ideas help us in building a network that outperforms the cited state-of-the-art techniques.

2. Wide and Deep Network (WDN)

2.1. Division into 32 subproblems

As stated earlier, we divide the 4× image up-sampling problem into 32 disjoint subproblems. Sixteen of these subproblems are of predicting the 16 disjoint subsets of pixels of the ground-truth image. These subsets are created by dividing the full set of pixels of the ground truth image using space-to-depth [69] as illustrated in Figure 2a. Each such subset is a low-resolution image in itself. Illustrating this with an example: if the ground truth image is of shape 4H×4W×C (H: Height, W: Width, C: Channel), space-to-depth generates 16 images of shape H×W×C, this is the same shape as that of the input low-resolution image (in 4× up-sampling). Hence, the 16 out of 32 subproblems are for predicting the 16 images without performing any up or down-sampling.

However, in an up-sampling problem, one of the most challenging tasks is to predict the high-frequency details, as details of this frequency are lost the maximum in a low-resolution image. Here, the high-frequency details refer to the pixel intensities of those areas of the image where intensity changes rapidly in a small neighbourhood, like, ob-



(a) All-frequency detail / (low and high)-frequency detail.

(b) High-frequency detail.

Figure 2: Illustration of Space-to-Depth (S2D) [69] and Depth-to-Space (D2S) [68] operations on a) all (low and high) frequency image, and b) high-frequency image. Low-resolution images on the right show one set of the ground truths for 32 deep networks.

ject edges. Inaccurate prediction of the high-frequency details results in the generation of a blurred up-sampled image when compared with the ground truth. Hence to give special attention to the high-frequency details, 16 additional subproblems are created that exclusively predict only the high-frequency details of the 16 pixel-subsets generated before. This is illustrated in Figure 2b.

Succinctly speaking, the 16 subproblems of predicting the all-frequency (low and high) details for the 16 subsets of pixels along with 16 subproblems of predicting only the high-frequency details for the same subsets together constitute the 32 subproblems into which the 4× image super-resolution problem is divided in this work.

2.2. Network design

As the sensitivity to the Luminance change is high in human beings, our wide and deep network (WDN) is set up to up-sample (4×) the Luminance (Y) channel of the image. The remaining channels are up-sampled using a simple bi-cubic interpolation. To solve the 32 subproblems, WDN consists of two modules: the prediction module and the output module. These modules are described in the following sections.

2.2.1 Prediction module

Prediction module has 32 (width) deep networks, connected in parallel as illustrated in Figure 3. Each network is trained to predict the solution of one of the 32 subproblems defined in Section 2.1. The 32 deep networks can be grouped into two categories: 1) 16 networks for all-frequency prediction, and 2) 16 networks for high-frequency prediction. The architecture of each deep network is given in Figure 4.

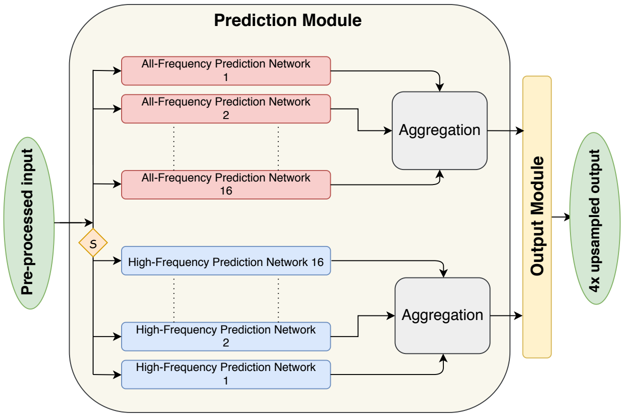
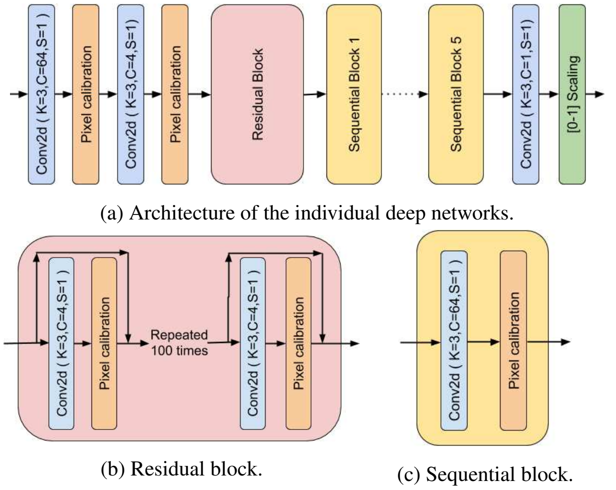

Prediction Module
Pre-processed input
All-Frequency Prediction Network 1
All-Frequency Prediction Network 2
All-Frequency Prediction Network 16
Aggregation
S
High-Frequency Prediction Network 16
High-Frequency Prediction Network 2
High-Frequency Prediction Network 1
Aggregation
Output Module
4x upsampled output

Figure 3: Wide and Deep Network: The pre-processed input is further processed by 16 all-frequency prediction networks and 16 high-frequency prediction networks simultaneously. All the 32 outputs are then aggregated and forwarded to the output module for predicting the final up-sampled image. 'S' within a diamond represents the extraction of high-frequency details using the procedure described in Section 2.3.



(a) Architecture of the individual deep networks.

(b) Residual block.

(c) Sequential block.

Figure 4: The architecture of the deep networks in prediction module. Pixel calibration has been described in Section 2.4. K: Kernel Size, C: No. of output feature maps, S: Stride

To generate the input for all-frequency prediction networks, Luminance channel Y (in YCbCr colour-space) is extracted from the given colour image of shape 4H×4W×3. The extracted channel is scaled in the range $[0, 1]$ and is downsampled to generate the low-resolution image. Instead of up-sampling the generated low-resolution image, WDN works upon refining its bi-cubically interpolated high-resolution version. Hence, the low-resolution image is upsampled using bi-cubic interpolation. Finally, space-to-depth [69] is applied to the upsampled image to generate the input that is of shape H×W×16. The high-frequency maps of these 16 channels are the input to the high-frequency prediction sub-networks of WDN. During inference, the process starts directly from the bi-cubic upsampling step.

The 16 disjoint ground truths for the all-frequency prediction networks are generated by applying space-to-depth [69]] on the Luminance channel of the given ground-truth (scaled in the range $0 - 1$). Mean squared error and dissimilarity measure (using SSIM [97]) as shown in Eq. 1 are minimised to train the individual deep-networks.

$$\begin{aligned} L = (\frac{1}{n}\sum_{i=1}^{n}(y_{gt}^{i} - y_{pred}^{i})^2) \times \lambda \\ + (1.0 - SSIM(y_{gt}, y_{pred})) \times (1.0 - \lambda) \end{aligned} \tag{1}$$

where y_{gt} is the ground-truth, y_{pred} is the prediction, λ is the weighting parameter and is set to 0.16, n is the number of pixels.

The 16 disjoint ground truths for the high-frequency prediction networks are generated by extracting high-frequency map from the Luminance channel of the given ground-truth and then applying space-to-depth over it. Huber loss by *Huber et al.* [27] as shown in Eq. 2 (with $\delta = 0.1$) between the predicted output and the ground truth is minimised to train the individual deep-networks.

$$\begin{aligned} L_{\delta}(a) = \begin{cases} \frac{1}{2}a^2 & \text{for } |a| \le \delta, \\ \delta(|a| - \frac{1}{2}\delta), & \text{otherwise.} \end{cases} \\ a = y_{gt} - y_{pred}; \end{aligned} \tag{2}$$

where y_{gt} is the ground-truth, y_{pred} is the prediction, δ is the point where the Huber loss function changes from quadratic to linear.

The aggregation of all the 32 sub-solutions is also the responsibility of this module. For aggregation, the 16 all-frequency feature maps (each of size H×W×1) are processed by Depth-to-space [68], as shown in Figure 2 to generate a single feature map of size 4H×4W×1. The same operation is performed with the 16 feature maps having the high-frequency details to generate another feature map of size 4H×4W×1. These two up-sampled feature maps are concatenated on the channel dimension. The resultant feature map, *i.e.* of shape 4H×4W×2, is further processed by the output module that fuses and refines the two channels for generating the final output.

2.2.2 Output module

This module performs the fusion of the concatenated data coming from the prediction module. It accepts the high-frequency and all-frequency feature maps generated by the prediction module that is of shape 4H×4W×2 and fuses them together to predict a single sharper up-sampled image of shape 4H×4W×1.

The architecture of this module is similar to the one shown in Figure 4a but with only twenty residual blocks. The loss minimised to train this module is the weighted sum of Mean squared error and Dissimilarity metric as shown in Eq. 1 between the predicted up-sampled image and the corresponding ground truth.

We now describe the high-frequency map extraction procedure and the pixel calibration technique that has been frequently used in the proposed architecture.

2.3. Extraction of high-frequency map

To extract the high-frequency map from the given image (scaled in range 0-1), Sobel filters X and Y are applied on it to generate $\mathbb{D}_x$, $\mathbb{D}_y$, that are the approximations of the derivatives for the horizontal and vertical changes.

$$X = \begin{bmatrix} +1 & 0 & -1 \\ +2 & 0 & -2 \\ +1 & 0 & -1 \end{bmatrix}, \quad Y = \begin{bmatrix} +1 & +2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}$$

$\mathbb{D}_x$, $\mathbb{D}_y$ are used to compute the high-frequency map, using Eq. 3.

$$\begin{aligned} \text{high-frequency map} = \mathbb{S}\left(\sqrt{\mathbb{D}_x^2 + \mathbb{D}_y^2}\right) \\ \mathbb{S} \text{ scales the values in range } [0, 1] \end{aligned} \tag{3}$$

2.4. Pixel calibration technique

The pixel calibration technique calibrates the pixel values by taking into account pixel relevance in the up-sampling task. It is a self-learned, non-linear complex layer (*i.e.* a layer built with multiple layers and activations) and is applied after each convolutional layer, in place of the typical activation function, to scale up or down the pixel/feature intensities of the input feature map. Since the relative importance values of the pixels are learned and computed from the pixels themselves, this layer is considered as self-learned.

To calibrate the pixels of a feature map, we compute the relevance score of each pixel of the feature map in the range [0-1] using a two-dimensional convolution operation with Sigmoid activation. We also computed the irrelevance score of each pixel by subtracting its relevance score from one. Finally, to calibrate the negative pixels, we weigh the positive pixel values with relevance score and all values of the feature map with irrelevance scores, before summing them up to generate the calibrated output, as shown in Eq. 4.

$$\begin{aligned} Calibrate(x) = (relu(x) \times U) + (x \times (1.0 - U)) \\ U = sigmoid(conv2d(x)) \end{aligned} \tag{4}$$

where conv2D has a stride of 1, kernel size of 3 and its number of output feature maps is the same as the number of feature maps in x. U is interpreted as the relevance score,

Table 1: Results obtained on 4× scaling factor for quantitative comparison with the current state-of-the-art methods. All SSIM values have been multiplied by 100.

Dataset	Metric	[64]	[65]	[28]	[3]	[90]	[48]	[23]	[24]	[41]	[92]	[42]	[14]	[91]	WDN
Set5	PSNR	31.68	31.74	31.82	31.92	31.96	32.27	32.47	32.53	32.56	32.61	32.62	32.70	32.73	**32.89**
	SSIM	88.88	88.93	89.03	89.03	89.30	89.38	89.80	89.92	89.92	90.03	89.84	90.13	90.13	**90.54**
Set14	PSNR	28.21	28.26	28.25	28.42	28.35	28.71	28.82	28.86	28.87	28.92	28.94	29.05	28.98	**29.09**
	SSIM	77.20	77.23	77.30	77.62	77.70	78.35	78.60	78.78	78.81	78.93	79.01	79.21	79.10	**79.36**
B100	PSNR	27.38	27.40	27.41	27.44	27.49	27.64	27.72	27.75	27.77	27.80	27.79	27.86	27.85	**27.96**
	SSIM	72.84	72.81	72.97	73.04	73.40	73.78	74.00	74.28	74.19	74.34	74.37	74.57	74.55	**74.89**
Urban100	PSNR	25.44	25.50	25.41	25.63	25.68	-	27.08	26.79	26.73	26.82	26.86	27.23	27.10	**27.35**
	SSIM	76.38	76.30	76.32	76.88	77.30	-	79.50	80.68	80.43	80.69	80.80	81.69	81.42	**82.06**

The design of this technique has its roots in the work of *Srivastava et al.* [63], but our version is conceptually different from it. The cited work had the idea of 'transform and carry'. However, in the domain of super-resolution, relating transform and carry with the relevance and irrelevance scores seemed more logical to us. Thus, we modified *Srivastava et al.* [63] to suit our purpose and 'instead of activating the transformed (Conv2d) version of the input feature map (x) we activated the input feature map directly without transformation'. We refer the reader to *Srivastava et al.* [63], to have a better understanding of the stated difference. Our empirical observations in Table 4 indicates that the proposed variation generates much better results than the original version. The details required to train WDN are described next.

2.5. Training procedure

The prediction module is trained and frozen before the training of the output module begins. All the trainable weights of the model are initialised with *Glorot initialisation* [18] and tuned with Adam Optimiser having $\beta1 = 0.5$ and $\beta2 = 0.9$ at a learning rate of 10^{-4}. The training is performed until no significant improvement is observed in the validation data for three consecutive epochs. We use 800 images from the DIV2K dataset by *Timofte et al.* [73] as the training set. Some of the common augmentation techniques such as random 1) cropping, 2) rotation and 3) horizontal flipping, have been used to increase the size of the training data.

3. Experiments and analysis

3.1. Datasets and evaluation metric

We performed extensive experiments on four publicly available datasets, Set5 by *Bevilacqua et al.* [5, 71], Set14 by *Zeyde et al.* [89], B100 by *Martin et al.* [45, 72], and Urban100 by *Huang et al.* [25] to evaluate the performance and efficiency of our proposed Wide and Deep network (WDN) in terms of the standard PSNR [29] and SSIM [97] metrics. In this section, we present the results that were obtained upon conducting those experiments.

3.2. Comparison with the state-of-the-art

Figure 5 and 6 shows the output of our model, along with the outputs of other state-of-the-art models for visual comparison. It can be observed that the up-sampled image generated by our model is sharper and visually more similar to the ground truth than other images.

To quantitatively compare the performance of our model with other state-of-the-art models, we computed the average of the PSNR/SSIM values between all the predicted images and their corresponding ground-truths on Y (Luminance) channel within a dataset. Table 1 shows the results on Set5, Set14, B100, and Urban100 datasets, obtained by current state-of-the-art methods and WDN. It is evident that WDN outperforms all the cited methods. For this improved performance of our model, we credit to the: 1) Core principle of divide and conquer, 2) Wide and deep network architecture, and 3) Pixel calibration technique.

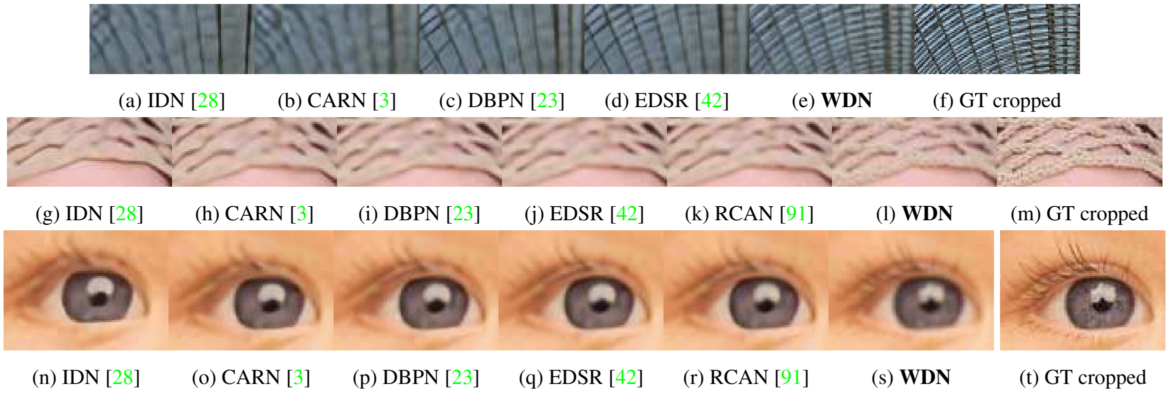
3.3. Performance on 2× and 3× scaling factors

To evaluate the performance of our approach on other scaling factors, WDN was architecturally modified as follows: 1) For 2× scaling, the network width was set to eight, *i.e.* four deep networks for predicting the all-frequency details and the other four for predicting the high-frequency details. 2) For 3× scaling, the network width was set to eighteen, *i.e.* nine deep networks for predicting the all-frequency details and the other nine for predicting the high-frequency details. The overall approach of divide and conquer with the aggregation of sub-solutions, and high-frequency fusion in output module remained the same. The results that were obtained are shown in Table 2 for quantitative comparison. It is evident from the results that WDN outperforms all the cited methods in 2×, and 3× up-sampling, due to the reasons as mentioned in Section 3.2.

3.4. Ablation studies

3.4.1 Effectiveness of the high-frequency fusion

The Output module can be configured in three modes: 1) Disabled mode: In this mode, the up-sampled image generated by the prediction module is considered as the final out-

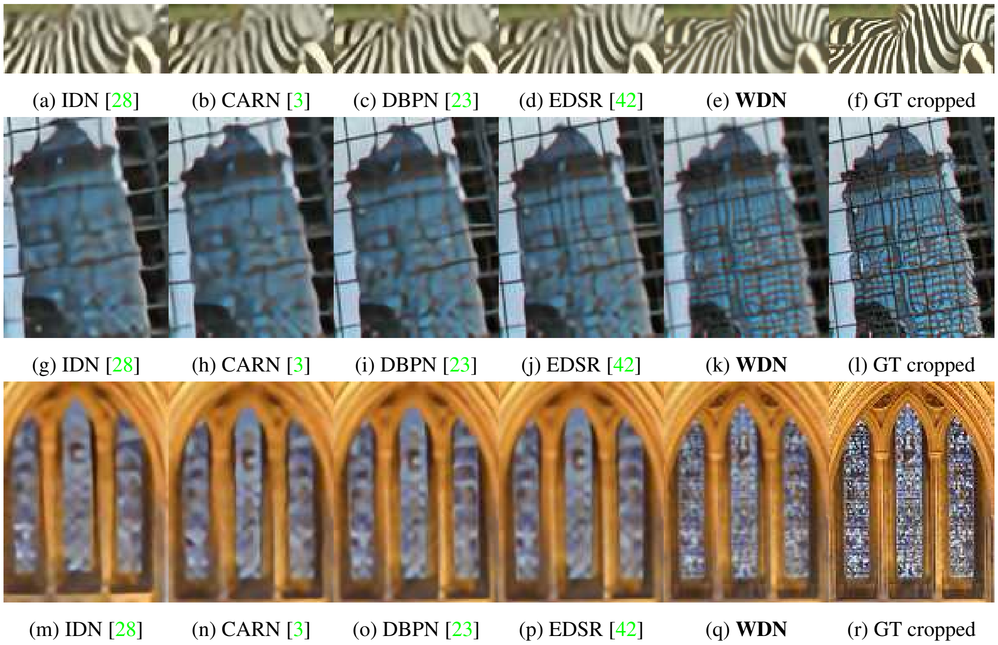


(a) IDN [28] (b) CARN [3] (c) DBPN [23] (d) EDSR [42] (e) **WDN** (f) GT cropped

(g) IDN [28] (h) CARN [3] (i) DBPN [23] (j) EDSR [42] (k) RCAN [91] (l) **WDN** (m) GT cropped

(n) IDN [28] (o) CARN [3] (p) DBPN [23] (q) EDSR [42] (r) RCAN [91] (s) **WDN** (t) GT cropped

Figure 5: Visualisation of the upsampled images for qualitative comparison with the current state-of-the-art methods on 4× scaling factor. **Better viewed on-screen after zooming.**



(a) IDN [28] (b) CARN [3] (c) DBPN [23] (d) EDSR [42] (e) **WDN** (f) GT cropped

(g) IDN [28] (h) CARN [3] (i) DBPN [23] (j) EDSR [42] (k) **WDN** (l) GT cropped

(m) IDN [28] (n) CARN [3] (o) DBPN [23] (p) EDSR [42] (q) **WDN** (r) GT cropped

Figure 6: Visualisation of the upsampled images for qualitative comparison with the current state-of-the-art methods on 4× scaling factor. **Better viewed on-screen after zooming.**

put and Output module is not used. 2) Refinement mode: In this mode, the Output module does not fuse the high-frequency details with the up-sampled image but refines the predicted up-sampled image generated after aggregation of sub-solutions without using any high-frequency details, 3) Fusion mode: In this mode, the output module fuses the high-frequency details with the up-sampled image, as proposed in this paper. The test results obtained under these settings are presented in Table 3.

It can be observed that, when the Output module is disabled, the results are the lowest. This is because all the images generated by different deep networks of the prediction module requires combined processing/refinement. Thus, when the Output module is set to refinement mode, the results are improved. A significant increase is observed when the Output module is used for high-frequency fusion, as proposed in this work. It can be inferred from these results that the explicit prediction of the high-frequency details and its fusion with the up-sampled image is indeed effective and high-frequency prediction plays an important role in the up-sampling task.

3.4.2 Effectiveness of the pixel calibration technique

To analyse the efficacy of pixel calibration, we trained the network after replacing the calibration layer with some of

Table 2: Results obtained on 2× and 3× scaling factors for quantitative comparison with current state-of-the-art methods. All SSIM values have been multiplied by 100.

(a) Comparing on 2× scaling factor.

Dataset	Metric	[42]	[92]	[91]	**WDN**
Set5	PSNR	38.20	38.30	38.33	**38.42**
	SSIM	96.06	96.16	96.17	**96.26**
Set14	PSNR	34.02	34.10	34.23	**34.31**
	SSIM	92.04	92.18	92.25	**92.39**
B100	PSNR	32.37	32.40	32.46	**32.53**
	SSIM	90.18	90.22	90.31	**90.40**
Urban100	PSNR	33.10	33.09	33.54	**33.73**
	SSIM	93.63	93.68	93.99	**94.28**

(b) Comparing on 3× scaling factor.

Dataset	Metric	[42]	[92]	[91]	**WDN**
Set5	PSNR	34.76	34.78	34.85	**34.90**
	SSIM	92.90	93.00	93.05	**93.11**
Set14	PSNR	30.66	30.67	30.76	**30.83**
	SSIM	84.81	84.82	84.94	**85.01**
B100	PSNR	29.32	29.33	29.39	**29.42**
	SSIM	81.04	81.05	81.22	**81.29**
Urban100	PSNR	29.02	29.00	29.31	**29.44**
	SSIM	86.85	86.83	87.36	**87.51**

Table 3: Results obtained upon setting the Output module in different modes. Output module mode: 1-Disabled, 2-Refinement, and 3-Fusion mode. All SSIM values have been multiplied by 100.

Dataset	Mode→	1	2	3
Set5	PSNR	32.71	32.77	**32.89**
	SSIM	90.16	90.32	**90.54**
Set14	PSNR	29.01	29.03	**29.09**
	SSIM	79.15	79.20	**79.36**
B100	PSNR	27.83	27.88	**27.96**
	SSIM	74.50	74.64	**74.89**
Urban100	PSNR	27.12	27.22	**27.35**
	SSIM	81.41	81.70	**82.06**

the possible alternatives that are 1) ReLU layer, 2) ReLU with Batch normalisation, and 3) Configuration proposed by *Srivastava et al.* [63]. The test results that were obtained are shown in Table 4. It can be observed in these results that the model performs the worst with the use of only ReLU layer. These results are improved with the use of ReLU + Batch Normalisation together. Much better results are obtained with the configuration proposed by *Srivastava et al.* [63], and the proposed calibration technique gives the best results among other possible alternatives. The effectiveness of the proposed pixel calibration in identifying the relevance of the pixels for the up-sampling task can be inferred from these results.

Table 4: Results obtained upon replacing the pixel calibration with possible alternatives. Configuration: 1) ReLU, 2) ReLU with Batch normalisation, 3) Config. given by *Srivastava et al.* [63], and 4) Pixel calibration as described in this work. All SSIM values have been multiplied by 100.

Dataset	Config.→	1	2	3	4
Set5	PSNR	31.49	31.84	32.70	**32.89**
	SSIM	88.74	89.09	90.11	**90.54**
Set14	PSNR	28.06	28.23	28.96	**29.09**
	SSIM	76.89	77.25	79.14	**79.36**
B100	PSNR	26.32	27.42	27.86	**27.96**
	SSIM	71.04	72.95	74.59	**74.89**
Urban100	PSNR	25.02	25.60	27.07	**27.35**
	SSIM	76.11	76.80	81.36	**82.06**

Table 5: Results obtained by varying the number of Residual blocks. All SSIM values have been multiplied by 100.

Dataset	Metric	50	75	100	125	150
Set5	PSNR	32.61	32.72	**32.89**	32.69	32.57
	SSIM	89.66	89.95	**90.54**	89.86	89.54
Set14	PSNR	28.85	28.97	**29.09**	28.94	28.85
	SSIM	78.68	78.94	**79.36**	78.87	78.60
B100	PSNR	27.79	27.86	**27.96**	27.83	27.75
	SSIM	74.38	74.58	**74.89**	74.50	74.25
Urban 100	PSNR	27.05	27.15	**27.35**	27.13	27.01
	SSIM	81.14	81.40	**82.06**	81.31	81.02

3.4.3 Effect of the number of residual blocks

To study the effects of the 'number of residual blocks' on PSNR/SSIM values, the network was trained and tested by varying the number of residual blocks. Table 5 presents the results that were obtained. It can be observed that the model with 100-layers performed the best among other options tried upon that might be over/under-fitting the data.

3.4.4 Effect of the number of sequential blocks

Further, to analyse the effects of the 'number of sequential blocks' on PSNR/SSIM values, the network was trained and tested by varying the number of sequential blocks. Table 6 contains the results that were obtained. It can be observed that the effects of the change are not much. Still, it can be said that setting up the sequential blocks to five gives optimal performance.

3.4.5 Analysis of the number of subproblems

For the 4× up-sampling scale, the super-resolution problem was divided into 16 subproblems of predicting the all-frequency details and 16 subproblems of predicting the corresponding high-frequency details. The size of 16 was chosen so that there is no up or down-sampling involved in solving any of the subproblems. Mathematically, if the low-resolution input is of size H×W and the ground truth is of

Table 6: Results obtained by varying the number of Sequential blocks. All SSIM values have been multiplied by 100.

Dataset	Metric	3	4	5	6	7
Set5	PSNR	32.88	**32.89**	**32.89**	32.87	32.85
	SSIM	90.50	90.52	90.54	**90.55**	**90.55**
Set14	PSNR	29.05	29.08	**29.09**	**29.09**	29.07
	SSIM	79.29	79.34	**79.36**	79.34	79.31
B100	PSNR	27.94	27.96	27.96	**27.97**	27.95
	SSIM	74.82	74.84	**74.89**	74.85	74.84
Urban 100	PSNR	27.34	**27.36**	27.35	27.33	27.31
	SSIM	82.01	82.03	**82.06**	**82.06**	82.00

Table 7: Results obtained with the different number of subproblems. All SSIM values have been multiplied by 100.

Dataset	Config.→	4+4	16+16	64+64
Set5	PSNR	32.34	**32.89**	29.60
	SSIM	88.67	**90.54**	81.20
Set14	PSNR	28.57	**29.09**	26.23
	SSIM	77.80	**79.36**	71.43
B100	PSNR	27.57	**27.96**	24.96
	SSIM	73.63	**74.89**	66.74
Urban100	PSNR	26.78	**27.35**	24.48
	SSIM	80.21	**82.06**	73.33

size 4H×4W, then the 16 subproblems are of predicting images, each of size H×W *i.e.* the same size as that of the input. The same reason was for dividing 2× up-sampling problem into four subproblems and 3× up-sampling problem into nine subproblems of predicting the all-frequency details. We also analysed the effect of the other number of subproblems in 4× up-sampling scale. Specifically, the problem was divided into 4 + 4 and 64 + 64 subproblems to predict all and high-frequency details.

To work upon with the four subproblems, space-to-depth was appropriately reconfigured to generate only four subsets of pixels. The last convolutional layer that is shown in Figure 4a was replaced with Convolutional Transpose layer with a stride of two, as now each subproblem required 2× up-sampling. Similarly, for the 64 subproblems, the last convolutional layer was updated with a stride of two, as each subproblem required 2× down-sampling. An equivalent number of subproblems were also created for predicting the high-frequency details, in both the cases. We present the results of changing the number of subproblems to 4 + 4 and 64 + 64 for the 4× up-sampling scale in Table 7.

It can be observed that the best results are obtained on 16 + 16 division of the subproblem. This is the case that has no up-sampling or down-sampling involved. The results also indicate that the 4 + 4 division of subproblems, can be further divided to utilise the divide and conquer approach better. The excessive number of the subproblem in 64 + 64 division is also not good as with this division, a significant amount of relevant information is lost in the down-sampled image. Moreover, the 64 + 64 division requires a very heavy network that might not be feasible always.

4. Summary, limitation, and future work

In this work, we proposed a wide and deep network (WDN) based on divide and conquer approach, to solve the image super-resolution problem. Particularly, we divided the 4× up-sampling problem into 32 disjoint, simultaneously solvable subproblems. Half of these subproblems were for predicting the overall features, while the rest were for predicting the finer details. We also proposed a technique to calibrate pixel/feature intensities. We demonstrated that our approach outperforms current state-of-the-art methods, both qualitatively and quantitatively on four publicly available datasets. We performed extensive ablation studies and empirically verified the efficacy of various components/ideas used in our approach.

The proposed model can be trained faster if the 32 deep networks of the prediction module are trained simultaneously, but this requires multiple graphics processing units or tensor processing unit. The hardware requirements get further increased if the up-sampling scale is 8×, as that requires a network with a width of 128 (*i.e.* 128 deep-nets). However, as each deep network in the prediction module is designed to be independent of the others, hence all the networks can also be trained one at a time on a single GPU when limited hardware is available. The output module can be separately trained after the training of all deep networks of the prediction module is complete. This will increase the training time for obvious reasons but will not have any effect on the accuracy.

The idea of solving a problem using the divide & conquer technique with deep networks connected in parallel can easily be adapted to solve other problems like video super-resolution, and deblurring. Moreover, in the super-resolution problem itself, the architecture of each deep network can be further improved. In the future, we plan to work in these directions.

Acknowledgements

Authors would like to thank TensorFlow Research Cloud (TFRC) program organisers for providing them free access to cloud Tensor Processing Units (TPUs). Authors would also like to thank the WACV-2020 committee (Reviewers, Area and Program chairs) for providing constructive feedback that helped in improving this work.

References

[1] E. Agustsson, R. Timofte, and L. V. Gool. Anchored regression networks applied to age estimation and super resolution. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1652–1661, Oct 2017. 2

[2] N. Ahn, B. Kang, and K. Sohn. Image super-resolution via progressive cascading residual network. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 904–9048, June 2018. 2

[3] N. Ahn, B. Kang, and K.-A. Sohn. Fast, accurate, and lightweight super-resolution with cascading residual network. In V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, editors, *Computer Vision – ECCV 2018*, pages 256–272, Cham, 2018. Springer International Publishing. 2, 5, 6

[4] Y. Bei, A. Damian, S. Hu, S. Menon, N. Ravi, and C. Rudin. New techniques for preserving global structure and denoising with low information loss in single-image super-resolution. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 987–9877, June 2018. 2

[5] M. Bevilacqua, A. Roumy, C. Guillemot, and M. line Alberi Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *Proceedings of the British Machine Vision Conference*, pages 135.1–135.10. BMVA Press, 2012. 5

[6] A. Bulat and G. Tzimiropoulos. Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 109–117, June 2018. 2

[7] A. Bulat, J. Yang, and G. Tzimiropoulos. To learn image super-resolution, use a gan to learn how to do image degradation first. In V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, editors, *Computer Vision – ECCV 2018*, pages 187–202, Cham, 2018. Springer International Publishing. 2

[8] J. Caballero, C. Ledig, A. Aitken, A. Acosta, J. Totz, Z. Wang, and W. Shi. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2848–2857, July 2017. 2

[9] R. Chen, Y. Qu, K. Zeng, J. Guo, C. Li, and Y. Xie. Persistent memory residual network for single image super resolution. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 922–9227, June 2018. 2

[10] Y. Chen, Y. Tai, X. Liu, C. Shen, and J. Yang. Fsrnet: End-to-end learning face super-resolution with facial priors. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2492–2501, June 2018. 2

[11] M. Cheon, J.-H. Kim, J.-H. Choi, and J.-S. Lee. Generative adversarial network-based image super-resolution using perceptual content losses. In L. Leal-Taixé and S. Roth, editors, *Computer Vision – ECCV 2018 Workshops*, pages 51–62, Cham, 2019. Springer International Publishing. 2

[12] J. Choi and M. Kim. A deep convolutional neural network with selection units for super-resolution. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1150–1156, July 2017. 2

[13] R. Dahl, M. Norouzi, and J. Shlens. Pixel recursive super resolution. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5449–5458, Oct 2017. 2

[14] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang. Second-order attention network for single image super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2, 5

[15] R. Dian, L. Fang, and S. Li. Hyperspectral image super-resolution via non-local sparse tensor factorization. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3862–3871, July 2017. 2

[16] S. Donn, L. Meeus, H. Q. Luong, B. Goossens, and W. Philips. Exploiting reflectional and rotational invariance in single image superresolution. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1043–1049, July 2017. 2

[17] Y. Fan, H. Shi, J. Yu, D. Liu, W. Han, H. Yu, Z. Wang, X. Wang, and T. S. Huang. Balanced two-stage residual networks for image super-resolution. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1157–1164, July 2017. 2

[18] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In Y. W. Teh and M. Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR. 5

[19] M. W. Gondal, B. Schölkopf, and M. Hirsch. The unreasonable effectiveness of texture transfer for single image super-resolution. In L. Leal-Taixé and S. Roth, editors, *Computer Vision – ECCV 2018 Workshops*, pages 80–97, Cham, 2019. Springer International Publishing. 2

[20] J. Guo and H. Chao. Building an end-to-end spatial-temporal convolutional network for video super-resolution. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI17, page 40534060. AAAI Press, 2017. 2

[21] T. Guo, H. S. Mousavi, T. H. Vu, and V. Monga. Deep wavelet prediction for image super-resolution. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1100–1109, July 2017. 2

[22] W. Han, S. Chang, D. Liu, M. Yu, M. Witbrock, and T. S. Huang. Image super-resolution via dual-state recurrent networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1654–1663, June 2018. 2

[23] M. Haris, G. Shakhnarovich, and N. Ukita. Deep back-projection networks for super-resolution. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1664–1673, June 2018. 2, 5, 6

[24] X. He, Z. Mo, P. Wang, Y. Liu, M. Yang, and J. Cheng. Ode-inspired network design for single image super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2, 5

[25] J. Huang, A. Singh, and N. Ahuja. Single image super-resolution from transformed self-exemplars. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5197–5206, June 2015. 5

[26] Y. Huang, L. Shao, and A. F. Frangi. Simultaneous super-resolution and cross-modality synthesis of 3d medical images using weakly-supervised joint convolutional sparse

coding. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5787–5796, July 2017. 2

[27] P. J. Huber. Robust estimation of a location parameter. *Ann. Math. Statist.*, 35(1):73–101, 03 1964. 4

[28] Z. Hui, X. Wang, and X. Gao. Fast and accurate single image super-resolution via information distillation network. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 723–731, June 2018. 2, 5, 6

[29] M. Irani and S. Peleg. Motion analysis for image enhancement: Resolution, occlusion, and transparency. *Journal of Visual Communication and Image Representation*, 4(4):324 – 335, 1993. 1, 5

[30] D. S. Jeon, S. Baek, I. Choi, and M. H. Kim. Enhancing the spatial resolution of stereo images using a parallax prior. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1721–1730, June 2018. 2

[31] Y. Jo, S. W. Oh, J. Kang, and S. J. Kim. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3224–3232, June 2018. 2

[32] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision – ECCV 2016*, pages 694–711, Cham, 2016. Springer International Publishing. 2

[33] A. Kappeler, S. Yoo, Q. Dai, and A. K. Katsaggelos. Video super-resolution with convolutional neural networks. *IEEE Transactions on Computational Imaging*, 2(2):109–122, June 2016. 2

[34] J. Kim and J. Lee. Deep residual network with enhanced upscaling module for super-resolution. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 913–9138, June 2018. 2

[35] J. Kim, J. K. Lee, and K. M. Lee. Accurate image super-resolution using very deep convolutional networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1646–1654, June 2016. 2

[36] J. Kim, J. K. Lee, and K. M. Lee. Deeply-recursive convolutional network for image super-resolution. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1637–1645, June 2016. 1, 2

[37] F. Lahoud, R. Zhou, and S. Süsstrunk. Multi-modal spectral image super-resolution. In L. Leal-Taixé and S. Roth, editors, *Computer Vision – ECCV 2018 Workshops*, pages 35–50, Cham, 2019. Springer International Publishing. 2

[38] W. Lai, J. Huang, N. Ahuja, and M. Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5835–5843, July 2017. 2

[39] C. Ledig, L. Theis, F. Huszr, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 105–114, July 2017. 2

[40] J. Li, F. Fang, K. Mei, and G. Zhang. Multi-scale residual network for image super-resolution. In V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, editors, *Computer Vision – ECCV 2018*, pages 527–542, Cham, 2018. Springer International Publishing. 1, 2

[41] Z. Li, J. Yang, Z. Liu, X. Yang, G. Jeon, and W. Wu. Feedback network for image super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2, 5

[42] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee. Enhanced deep residual networks for single image super-resolution. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1132–1140, July 2017. 2, 5, 6, 7

[43] D. Liu, Z. Wang, Y. Fan, X. Liu, Z. Wang, S. Chang, and T. Huang. Robust video super-resolution with learned temporal dynamics. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2526–2534, Oct 2017. 2

[44] X. Luo, R. Chen, Y. Xie, Y. Qu, and C. Li. Bi-gans-st for perceptual image super-resolution. In L. Leal-Taixé and S. Roth, editors, *Computer Vision – ECCV 2018 Workshops*, pages 20–34, Cham, 2019. Springer International Publishing. 2

[45] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pages 416–423 vol.2, July 2001. 5

[46] J. Pan, S. Liu, D. Sun, J. Zhang, Y. Liu, J. Ren, Z. Li, J. Tang, H. Lu, Y. Tai, and M. Yang. Learning dual convolutional neural networks for low-level vision. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3070–3079, June 2018. 2

[47] D. Park, K. Kim, and S. Y. Chun. Efficient module based single image super resolution for multiple problems. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 995–9958, June 2018. 2

[48] S.-J. Park, H. Son, S. Cho, K.-S. Hong, and S. Lee. Srfeat: Single image super-resolution with feature discrimination. In V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, editors, *Computer Vision – ECCV 2018*, pages 455–471, Cham, 2018. Springer International Publishing. 2, 5

[49] K. Purohit, S. Mandal, and A. N. Rajagopalan. Scale-recurrent multi-residual dense network for image super-resolution. In L. Leal-Taixé and S. Roth, editors, *Computer Vision – ECCV 2018 Workshops*, pages 132–149, Cham, 2019. Springer International Publishing. 2

[50] H. Ren, M. El-Khamy, and J. Lee. Image super resolution based on fusing multiple convolution neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1050–1057, July 2017. 2

[51] H. Ren, M. El-Khamy, and J. Lee. Image super resolution based on fusing multiple convolution neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1050–1057, July 2017. 2

[52] H. Ren, M. El-Khamy, and J. Lee. Ct-srcnn: Cascade trained and trimmed deep convolutional neural networks for image super resolution. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1423–1431, March 2018. 2

[53] H. Ren, M. El-Khamy, and J. Lee. Video super resolution based on deep convolution neural network with two-stage motion compensation. In *2018 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, pages 1–6, July 2018. 2

[54] M. S. M. Sajjadi, B. Schlkopf, and M. Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4501–4510, Oct 2017. 2

[55] M. S. M. Sajjadi, R. Vemulapalli, and M. Brown. Frame-recurrent video super-resolution. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6626–6634, June 2018. 2

[56] G. Seif and D. Androutsos. Large receptive field networks for high-scale image super-resolution. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 876–87609, June 2018. 2

[57] M. Sharma, R. Mukhopadhyay, A. Upadhyay, S. Koundinya, A. Shukla, and S. Chaudhury. Irgun : Improved residue based gradual up-scaling network for single image super resolution. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 947–94709, June 2018. 2

[58] W. Shi, J. Caballero, F. Huszr, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1874–1883, June 2016. 2

[59] Z. Shi, C. Chen, Z. Xiong, D. Liu, Z.-J. Zha, and F. Wu. Deep residual attention network for spectral image super-resolution. In L. Leal-Taixé and S. Roth, editors, *Computer Vision – ECCV 2018 Workshops*, pages 214–229, Cham, 2019. Springer International Publishing. 2

[60] A. Shocher, N. Cohen, and M. Irani. Zero-shot super-resolution using deep internal learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3118–3126, June 2018. 2

[61] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 2

[62] J. W. Soh, G. Y. Park, J. Jo, and N. I. Cho. Natural and realistic single image super-resolution with explicit natural manifold discrimination. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[63] R. K. Srivastava, K. Greff, and J. Schmidhuber. Training very deep networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2377–2385. Curran Associates, Inc., 2015. 2, 5, 7

[64] Y. Tai, J. Yang, and X. Liu. Image super-resolution via deep recursive residual network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2790–2798, July 2017. 2, 5

[65] Y. Tai, J. Yang, X. Liu, and C. Xu. Memnet: A persistent memory network for image restoration. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4549–4557, Oct 2017. 2, 5

[66] W. Tan, B. Yan, and B. Bare. Feature super-resolution: Make machine see more clearly. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3994–4002, June 2018. 2

[67] X. Tao, H. Gao, R. Liao, J. Wang, and J. Jia. Detail-revealing deep video super-resolution. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4482–4490, Oct 2017. 2

[68] TensorFlow. Dept-to-space. `https://www.tensorflow.org/api_docs/python/tf/nn/depth_to_space`. Accessed: 2019-02-14. 3, 4

[69] TensorFlow. Space-to-depth. `https://www.tensorflow.org/api_docs/python/tf/nn/space_to_depth`. Accessed: 2019-02-14. 2, 3, 4

[70] Y. Tian, Y. Zhang, Y. Fu, and C. Xu. TDAN: temporally deformable alignment network for video super-resolution. *CoRR*, abs/1812.02898, 2018. 2

[71] R. Timofte, V. De, and L. V. Gool. Anchored neighborhood regression for fast example-based super-resolution. In *2013 IEEE International Conference on Computer Vision*, pages 1920–1927, Dec 2013. 5

[72] R. Timofte, V. DeSmet, and L. VanGool. A+: Adjusted anchored neighborhood regression for fast super-resolution. In D. Cremers, I. Reid, H. Saito, and M.-H. Yang, editors, *Computer Vision – ACCV 2014*, pages 111–126, Cham, 2015. Springer International Publishing. 5

[73] R. Timofte and team. Ntire 2017 challenge on single image super-resolution: Methods and results. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1110–1121, July 2017. 5

[74] T. Tong, G. Li, X. Liu, and Q. Gao. Image super-resolution using dense skip connections. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4809–4817, Oct 2017. 2

[75] S. Vasu, N. Thekke Madam, and A. N. Rajagopalan. Analyzing perception-distortion tradeoff using enhanced perceptual super-resolution network. In L. Leal-Taixé and S. Roth, editors, *Computer Vision – ECCV 2018 Workshops*, pages 114–131, Cham, 2019. Springer International Publishing. 2

[76] T. Vu, T. M. Luu, and C. D. Yoo. Perception-enhanced image super-resolution via relativistic generative adversarial networks. In L. Leal-Taixé and S. Roth, editors, *Computer Vision – ECCV 2018 Workshops*, pages 98–113, Cham, 2019. Springer International Publishing. 2

[77] W. Wang, C. Ren, X. He, H. Chen, and L. Qing. Video super-resolution via residual learning. *IEEE Access*, 6:23767–23777, 2018. 2

[78] X. Wang, K. Yu, C. Dong, and C. Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 606–615, June 2018. 2

[79] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In L. Leal-Taixé and S. Roth, editors, *Computer Vision – ECCV 2018 Workshops*, pages 63–79, Cham, 2019. Springer International Publishing. 2

[80] Y. Wang, F. Perazzi, B. McWilliams, A. Sorkine-Hornung, O. Sorkine-Hornung, and C. Schroers. A fully progressive approach to single-image super-resolution. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 977–97709, June 2018. 2

[81] Z. Wang, P. Yi, K. Jiang, J. Jiang, Z. Han, T. Lu, and J. Ma. Multi-memory convolutional neural network for video super-resolution. *IEEE Transactions on Image Processing*, 28(5):2530–2544, May 2019. 2

[82] J. Xu, Y. Zhao, Y. Dong, and H. Bai. Fast and accurate image super-resolution using a combined loss. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1093–1099, July 2017. 2

[83] X. Xu, D. Sun, J. Pan, Y. Zhang, H. Pfister, and M. Yang. Learning to super-resolve blurry face and text images. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 251–260, Oct 2017. 2

[84] W. Yang, J. Feng, G. Xie, J. Liu, Z. Guo, and S. Yan. Video super-resolution based on spatial-temporal recurrent residual networks. *Computer Vision and Image Understanding*, 168:79 – 92, 2018. 2

[85] X. Yu, B. Fernando, B. Ghanem, F. Porikli, and R. Hartley. Face super-resolution guided by facial component heatmaps. In V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, editors, *Computer Vision – ECCV 2018*, pages 219–235, Cham, 2018. Springer International Publishing. 2

[86] X. Yu, B. Fernando, R. Hartley, and F. Porikli. Super-resolving very low-resolution face images with supplementary attributes. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 908–917, June 2018. 2

[87] X. Yu and F. Porikli. Ultra-resolving face images by discriminative generative networks. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision – ECCV 2016*, pages 318–333, Cham, 2016. Springer International Publishing. 2

[88] Y. Yuan, S. Liu, J. Zhang, Y. Zhang, C. Dong, and L. Lin. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 814–81409, June 2018. 2

[89] R. Zeyde, M. Elad, and M. Protter. On single image scale-up using sparse-representations. In J.-D. Boissonnat, P. Chenin, A. Cohen, C. Gout, T. Lyche, M.-L. Mazure, and L. Schumaker, editors, *Curves and Surfaces*, pages 711–730, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. 5

[90] K. Zhang, W. Zuo, and L. Zhang. Learning a single convolutional super-resolution network for multiple degradations. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3262–3271, June 2018. 2, 5

[91] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu. Image super-resolution using very deep residual channel attention networks. In V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, editors, *Computer Vision – ECCV 2018*, pages 294–310, Cham, 2018. Springer International Publishing. 1, 2, 5, 6, 7

[92] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu. Residual dense network for image super-resolution. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2472–2481, June 2018. 2, 5, 7

[93] Z. Zhang and V. Sze. Fast: A framework to accelerate super-resolution processing on compressed videos. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1015–1024, July 2017. 2

[94] Z. Zhang, Z. Wang, Z. Lin, and H. Qi. Image super-resolution by neural texture transfer. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[95] H. Zheng, M. Ji, H. Wang, Y. Liu, and L. Fang. Crossnet: An end-to-end reference-based super resolution network using cross-scale warping. In V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, editors, *Computer Vision – ECCV 2018*, pages 87–104, Cham, 2018. Springer International Publishing. 2

[96] Z. Zhong, T. Shen, Y. Yang, C. Zhang, and Z. Lin. Joint sub-bands learning with clique structures for wavelet domain super-resolution. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS18, page 165175, Red Hook, NY, USA, 2018. Curran Associates Inc. 2

[97] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, April 2004. 1, 4, 5