Version of Record: https://www.sciencedirect.com/science/article/pii/S0923596518309846
Manuscript_7ee052ccb88a24975afff67a9a0a8982

Multi-Scale Gradient Image Super-Resolution for Preserving Key SIFT Points in Low-Resolution Images

Dewan Fahim Noor[a], Yue Li[a], Zhu Li[a], Shuvra Bhattacharyya[b], George York[c]

[a] *University of Missouri-Kansas City, MO, USA*
[b] *University of Maryland, College Park, MD, USA*
[c] *US Air Force Academy, USA*

Abstract

Low-resolution images present challenges to a variety of object recognition problems in a variety of surveillance and navigation applications. In recent years, deep learning has advanced the state of the art in image super-resolution (SR) in terms of pixel domain peak signal to noise ratio (PSNR)/ mean square error (MSE). Inspired by the recent advances of deep convolutional neural networks in general image SR tasks, we develop a computer vision task-driven image SR solution by learning super-resolved gradient images using multiple convolutional neural networks for different scales. Recovering super-resolved gradient images at multiple scales, enables the system to recover more information useful for high level vision tasks than simply SR in the pixel domain. In particular, we propose a residual learning framework to perform image SR in the Difference of Gaussian (DOG) domain. The trained residual network models are then adapted to drive a widely adopted key point algorithm for image recognition, i.e. the SIFT detection and matching. Experimental results show that the proposed approach can significantly improve the SIFT keypoints repeatability compared to the state of the art in pixel domain image SR solutions.

Keywords: Image Super-resolution, Difference of Gaussian, Gradient Image, SIFT repeatability

Email addresses: dfnrh2@mail.umkc.edu (Dewan Fahim Noor), lytt@mail.ustc.edu.cn (Yue Li), lizhu@umkc.edu (Zhu Li), ssb@umd.edu (Shuvra Bhattacharyya), george.york@usafa.edu (George York)

1. Introduction

In modern days, a key challenge in image recognition lies in dealing with low-resolution images specially in military and surveillance applications. Moreover, the ability to recognize faraway objects, is of great value to many target recognition problems in Department of Defense (DoD) use cases like counter Unmanned Aircraft System (UAS) applications. DDDAS (Dynamic Data Driven Applications Systems) in UAS application is a new paradigm whereby the computation and instrumentation aspects of an application system are dynamically integrated into a control loop having feedback [1]. The data is dynamically incorporated into the executing model of the application, and in reverse the executing model can control the instrumentation. The challenges DDDAS seeks to advance include data modeling, context processing, and content application. The data needs to be collected while being pre-processed to determine whether its inherent information matches the context. Some of the examples include clutter reduction, sensor registration and confuser analysis in vehicle tracking [1]. So, the images being taken need to be recognized accurately. One of the popular solutions in this case would be image super-resolution. Image super-resolution is one of the most important research areas in the field of computer vision and pattern recognition. Super-resolution [2] means finding a mapping from the low-resolution (LR) image to its high-resolution (HR) version. In the case of single frame super- resolution (SISR), for a single image, number of pixels is increased so that the super-resolved image can visually look better as well as can be efficacious while recognition. There are various approaches for super resolution. Bi-cubic and Bi-linear upscaling methods [3] are very popular for super-resolution which have been used to a great extent. Moreover, sparse coding representation based SR methods [4, 5], have improved the resolution a lot.

In modern era, deep learning based super-resolution methods have left quite a good impression in research. The Deep Learning based methods have shown more accuracy than the conventional methods. Recently, numerous deep learn-

ing based super-resolution methods have been introduced. In [6], SRCNN method is established. The algorithm is an end to end mapping between the input low-resolution images and its interpolated high-resolution images. The method also shows the jointly optimization of all layers. The results exhibit quite a good gain over the other methods. In [7], VDSR method is introduced which uses a very deep convolutional neural network by simply adding many stages of small filters. The algorithm results in faster convergence and shows excellent gain over the other methods. In [8], an enhanced deep learning based super-resolution (EDSR) method is introduced. The method is actually an enhanced version of residual network which is further replicated in stages to finally produce the deep layers of super-resolution network. However, in addition to super-resolving the image, the key concern is to preserve the features so that it can be recognized accurately.

In general, in typical super-resolution methods, the images which are produced have better visual quality with higher resolution in terms of PSNR as all of them have the loss function of mean square error (MSE). But, in real world producing better visual quality image might result in losing important features. Because the loss function based on MSE in pixel domain only tries to increase the PSNR and makes it visually better. But while identifying those images, we need to preserve the important local and global features. So, in practical world, we need to design a network which emphasizes more on preserving the features which contribute towards better recognition and detection of the robust objects. For example, captured images from surveillance cameras have very low-resolution. These low-resolution images have less number of pixels which actually mean that they have less information while being identified. So, these images should be super-resolved as well as be enriched with more quality pixels. While super-resolving those images, we also need to be very efficient in preserving the features. Otherwise the identification will be corrupted. In Air Force, while detecting any aircraft , the accuracy of detection should be very high. So the quality of image should be enhanced by super-resolving it as well as not losing the features. In short, the application of super-resolving the image as well as

preserving the key features/points is of paramount importance. There is quite a few work on low-resolution image recognition. In [9], very low-resolution recognition (VLRR) problem has been dealt with. Here, deep learning model has been developed for demonstrating the task with face recognition, font recognition, digit recognition criteria. In [10], a generative adversarial network (GAN) also known as SRGAN for image super-resolution is proposed. The method not only super-resolves images but also recovers photo-realistic textures from heavily downsampled images. In [11], another deep convolutional network based method is proposed to deal with face and other object with low quality. In [12], a multi-frame SR (MFSR) method is introduced for bio-metric purpose which reduces the equal error rate in person identification. On the constraint, we proposed a super-resolving method which aims at preserving the features by super-resolving the images in gradient image domain.

In many recognition tasks, gradient images are important information derived from pixel images. To define, gradient image generally refers to a change in the direction of the intensity or color of an image. Numerous works regarding image recognition have been done using gradient of images. In [13], Harris Detector is used to find out the edges and extract corners of the image as well as discovering the infer features of the image. In [14], Laplacian of Gaussian is used for blob detection. In [15], SIFT feature detection is used which discovers local features after computing maxima and minima from the DoG image set. In recognition, key points from an objects are extracted to provide a description of the features which are used for recognizing the object. So, it should be important to keep in mind that extracted features should be able to be used in case of scale, noise and illumination changes. SIFT can handle these changes which makes SIFT an ideal method for feature extraction. There is few research regarding the preservation of features. In [16], a visual query compression for preserving local features is introduced. Here, they go through a new method in visual key points compression which uses subspaces for optimization of preserving key point feature matching properties than the reconstruction performance. Moreover, SIFT features preservation plays important role in image recogni-

tion. There numerous research on the role of SIFT features to increase the accuracy in image recognition. In [17], [18], the application of SIFT features in image recognition are explained. In our proposed method, we are motivated to preserve these key SIFT points so that it can be fruitful in recognizing low-resolution images. Our proposed method in this paper is not an end to end system. Rather, it is a super-resolving network which generates SIFT repeatability. So, the goal of our proposed method is to produce SIFT repeatability and to show how these SIFT points contribute towards better recognition . In order to fulfill our goal to produce SIFT repeatability to preserve more features, we will do the super-resolution in gradient domain. In [19], the SIFT repeatability is tested on a small scale. In this paper, we tested our method on a larger scale with diversified datasets.

To be accurate, our main concern is not how much accuracy we are gaining for super-resolution in pixel domain. The main idea is to preserve the fine SIFT features which are the contributors of low-resolution image recognition. To preserve SIFT features, we aim at super-resolving images in gradient domain. Our SR network is built upon the concept of generating gradient images. The network actually consists of many stages SR networks. For each of the SR network , we establish deep learning method inspired from EDSR and Squeeze and Excitation Network [20] but instead of producing the super-resolved image of original input, we produce the Difference of Gaussian Images (DoG). In SIFT, DoG images [21] are produced from the input image with different scale and different standard deviations. In our method, the network produces the DoG images and integrate with SIFT method to find out the key points which are used for matching. Overall, our proposed method intends to generate super-resolved gradient images which preserves the SIFT features to produce SIFT repeatability.

2. Proposed method

Our proposed method contains a deep learning pipeline for image super-resolution. The original purpose of our network is quite different than the other deep learning based super-resolution methods. We aim to produce SIFT repeatability. So, instead of generating the upsampled image from the low-resolution input image, we target to generate gradient images hence DoG images in our case. The idea is to first generate DoG images and then finally integrate with SIFT to preserve SIFT matching points.The network's target is not to just create high-resolution but also preserve features, hence preserve the key points for SIFT.

Our proposed super-resolving infrastructure is constructed on the basis of generating super-resolved gradient image. Gradient images are generally constructed from the original image being convolved with a filter. In a gradient image, in a certain direction, each pixel finds out the the change in intensity of that same point in the original image. Our image gradient method is based on the SIFT method. In SIFT method, from an input image, different Gaussian blurred images are first produced with different standard deviation. Then difference of Gaussian[12] is computed for different scales which are called octaves. From DoG images, maxima and minima are computed to find key points. In SIFT method, from the key points, the edges and low contrast points are eliminated considering them as bad points. With rotation and scale invariance being considered, the key points are detected. Let, *I(x,y)* is the original image; *G(x,y,σ)* is the Gaussian Kernel. Equation (1) and (2) [22],[23] show the formulation of Gaussian blurred images.

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{\frac{-(x^2+y^2)}{2\sigma^2}} \tag{1}$$

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \tag{2}$$

Where, *L(x,y,σ)* is the Gaussian blurred image with specific σ which is the standard deviation, x is the distance from the origin in the horizontal axis, y is the distance from the origin in the vertical axis

So, the DoG will be as followed in equation (3) and (4)[22],[23]:

$$D(x, y, \sigma_1, \sigma_2) = (G_1(x, y, \sigma_1) - G_2(x, y, \sigma_2)) * I(x, y)) \tag{3}$$

$$D(x, y, \sigma_1, \sigma_2) = L_1(x, y, \sigma_1) - L_2(x, y, \sigma_2) \tag{4}$$

Where, $D(x,y,\sigma_1,\sigma_2)$ is the of DoG image, $\sigma_1$is the standard deviation of the first blurred image and $\sigma_2$is the standard deviation of the second blurred image. G_1,G_2 are Gaussian filters. L_1,L_2 are Gaussian blurred images.

The loss function E is the MSE loss between the DoG of the super-resolved blurred generated image and the DoG from convolution with original image which can be shown in (5)[24]:

$$E(\hat{D}, D_{original}) = \sum_{i=1}^{n}\sum_{j=1}^{m}(\hat{D}^{ij} - D^{ij}_{original}\ )^2 \tag{5}$$

Where $\hat{D}$ is the predicted DoG image which is upscaled and $D_{original}$ is the DoG image computed from of the original one convolved with Gaussian filter. n and m are the numbers of pixels in x and y direction.

The gradient descent of the loss function is the differentiation with respect to $\hat{D}$ as followed in equation (6),(7) and (8):

$$\frac{\delta E}{\delta \hat{D}} = \frac{\delta(\sum_{i=1}^{n}\sum_{j=1}^{m}(\hat{D}^{ij} - D^{ij}_{original}\ )^2)}{\delta \hat{D}} \tag{6}$$

$$\frac{\delta E}{\delta \hat{D}} = 2\sum_{i=1}^{n}\sum_{j=1}^{m}(\hat{D}^{ij} - (\frac{1}{2\pi{\sigma_1}^2}P - \frac{1}{2\pi{\sigma_2}^2}Q))(1 - (\frac{1}{2\pi{\sigma_1}^2}\frac{\delta P}{\delta \hat{D}} - \frac{1}{2\pi{\sigma_2}^2}\frac{\delta Q}{\delta \hat{D}})) \tag{7}$$

$$P = e^{\frac{-(x_i^2+y_j^2)}{2{\sigma_1}^2}} * I(x_i, y_j), Q = e^{\frac{-(x_i^2+y_j^2)}{2{\sigma_2}^2}} * I(x_i, y_j) \tag{8}$$

Here, equation 7 is derived from equation 6 after differentiating it with respect to $\hat{D}$. In equation 7, due to the complexity of the equation we introduce two terms P and Q [shown in equation 8]which are the exponential terms for the Gaussian filter in each image convolved with the original image $I(x_i, y_j)$ where x_i is the distance from the origin in the horizontal axis, y_j is the distance from the origin in the vertical axis.

As the loss function and its gradient descent seem to be very complex, it can be simplified if we use the MSE loss between Gaussian blurred images as our loss

function and then we compute the DoG images from the Gaussian blurred image. The following equation (10) is the simplified loss function. But in this case, the output will be Gaussian blurred image instead of DoG images.

$$E(\hat{L}, L_{original}) = \sum_{i=1}^{n}\sum_{j=1}^{m}(\hat{L}^{ij} - L_{original}^{ij})^2 \tag{9}$$

Where $\hat{L}$ is the predicted blurred image which is upscaled and $L_{original}$ is the Gaussian blurred image of the original image with same standard deviation.

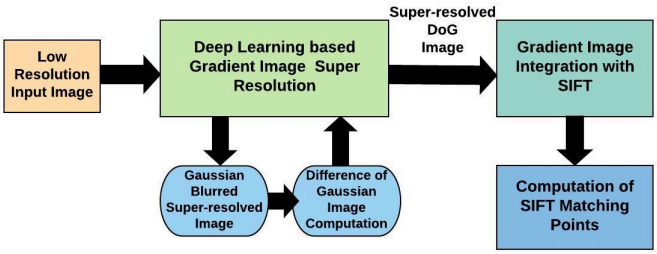


Figure 1: The diagram of the proposed method

There are different stages in our proposed method. From the low-resolution input images, the deep learning based Gradient image super-resolution stage generates DoG images. The SIFT integration stage integrates the DoG images for show-casting SIFT repeatability. Finally, a SIFT points matching comparison is done to evaluate the performance. Figure 1 shows the full network architecture of our proposed methods. We basically, compute the DoG images in two different ways. In first method, we directly learn the DoG images from the network. In second method, we learn the Gaussian blurred images first and then compute the DoG images from the Gaussian blurred images.

Method-1 is shown in Figure 2. For the super-resolution network design, the residual blocks(ResBlocks) concept is taken from EDSR. Residual learning [25] is very instrumental for faster convergence. So, in our network, we construct residual blocks. The network is supposed to build four super-resolved DoG images. So, it has four deep learning based SR networks. Each of the four networks contains several ResBlocks followed by deconvolution layers. Each ResBlock contains a residual block which is followed by a Squeeze and Excitation network. Residual blocks have a convolutional

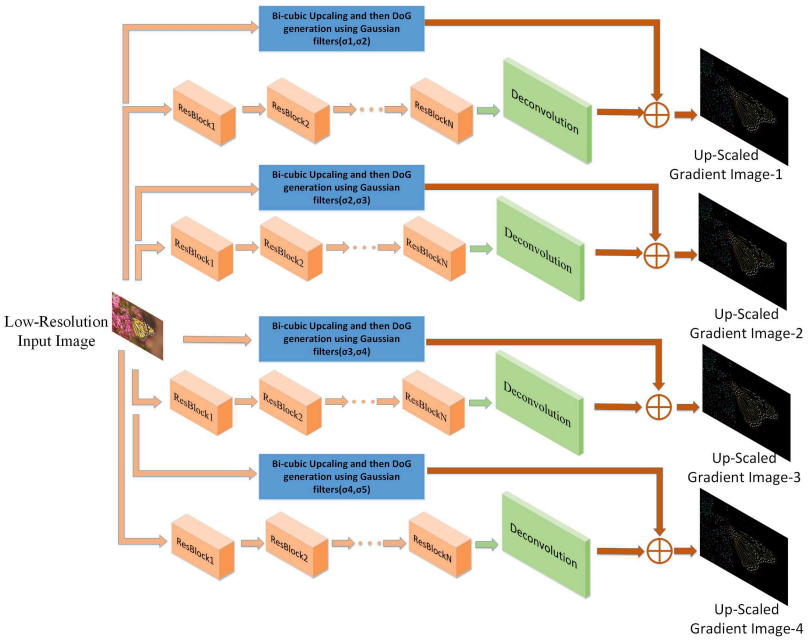


Figure 2: Deep Learning based Gradient Image Super-resolution for method-1

layer followed by rectified linear unit(ReLU)[26] and again a convolutional layer. Each convolutional layer has filter kernel size of 3X3 with 64 number of channels. In the Squeeze and Excitation network, the output from residual block is followed by a global pooling layer, fully connected layer, ReLU, a fully connected layer again and a sigmoid function followed by the scaling. The input to the residual block is added to the output of Squeeze and Excitation network for the residual learning. The Squeeze and Excitation network improves channel-wise feature responses by modelling the relationships between channels [20] as shown in Figure 3 which works as a boosting factor in our method. We combined the residual learning concept with squeeze and excitation channel to enhance the feature to a certain level by developing the response created by scaling in squeeze and excitation network. Next, the deconvolutional layer [27] does the upscaling of the image. Here, stride value 2 or 4 is used for either 2X or 4X upscaling. A predictor is also added to the output. The predictor is the upsampled version of the input convolved with two Gaussian filters to compute the DoG image. So, the network is learning the residue. The ground truth for the method-1 is the DoG images computed from the original images convolved with Gaussian filters with

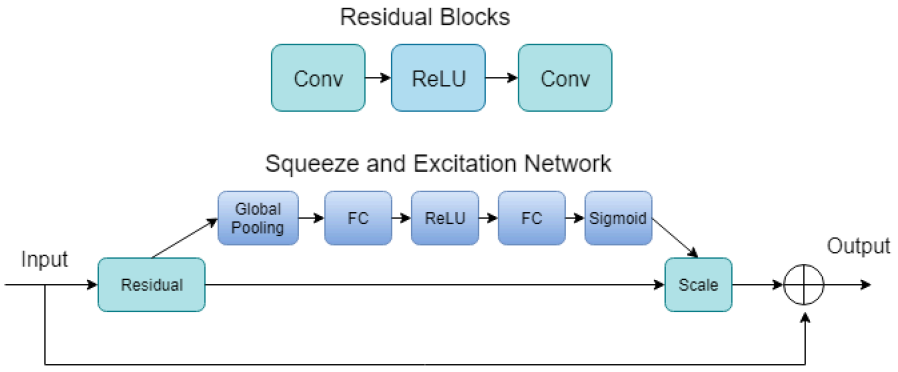


Figure 3: Residual Blocks with Squeeze and Excitation Network. The combination of residual blocks with Conv-ReLU-Conv similar to EDSR and the squeeze and excitation channel shows its novelty for the purpose of both storing features by improving channel-wise feature response

different standard deviations.

Method-2 is shown in Figure 4. Here, instead of learning the DoG images , Gaussian blurred images are predicted first. So, it has five deep learning based SR networks which compute Gaussian blurred images and then DoG images are computed simply by subtraction between the images. However, the residual learning is not added here. As including the residual learning did not change the performance much in method-2, rather increase the complexity of the network, we concluded not to use the residual learning here. Rest of the network structure is same as method-1. The ground truth for the method-2 is the Gaussian blurred images computed from the original images convolved with Gaussian filters with different standard deviations.

However, the number of ResBlocks is not fixed. We design an adaptive solution to the number of Resblocks. As we have 4 separate SR networks for method-1 and 5 networks for method-2 to generate DoG images from Gaussian filters different standard deviation value(σ), we adapt the number of blocks according to the sigma value. For higher σ value the number of ResBlocks is reduced. In equation 10 and 11, DoG images in different scales are derived.

$$L_k(x, y, \sigma_k) = G_k(x, y, \sigma_k) * I(x, y) \tag{10}$$

$$D_i(x, y, \sigma_k, \sigma_{k+1}) = L_k(x, y, \sigma_k) - L_{k+1}(x, y, \sigma_{k+1}) \tag{11}$$

Here, G_k is the Gaussian filter with σ_k, L_k is the Gaussian blurred image at σ_k

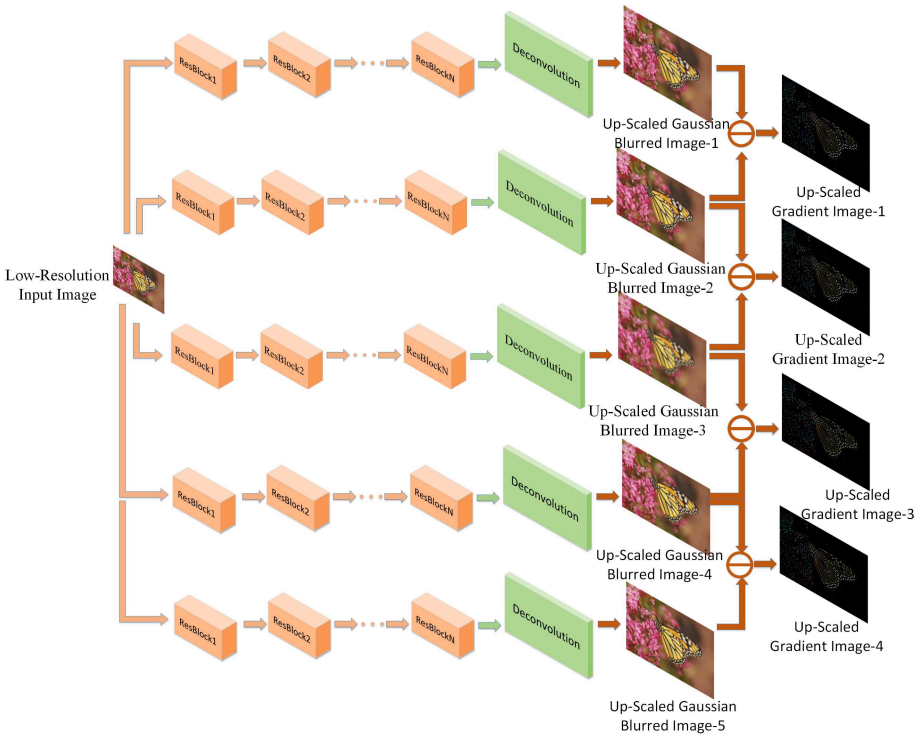


Figure 4: Deep Learning based Gradient Image Super-resolution for method-2

where k=1,2,3,4,5 and D_i is the DoG image at σ_k and σ_{k+1} scale where i=1,2,3,4. So, the four DoG images are labelled as D_1,D_2,D_3,D_4.

Table 1: Average of power spectrum density of images from each dataset.

σ_1	σ_2	σ_3	σ_4	σ_5
1.249	1.545	1.96588	2.452527	3.090016

We choose the σ values of 1.249, 1.545, 1.946588, 2.452527 and 3.090016 in accordance with the design of SIFT which is shown Table 1 . We analyzed the power spectrum density of the original image and the DoG images D_1,D_2,D_3,D_4 in the CDVS dataset[28], Oxford building dataset[29] and Paris dataset[30]. Table 2 shows the average power spectrum of 100 images from each dataset. It is viewed that the original image has more power spectrum density than the DoG images. As we increase the values of σ, the value of power spectrum density decreases. It means it cuts a lot

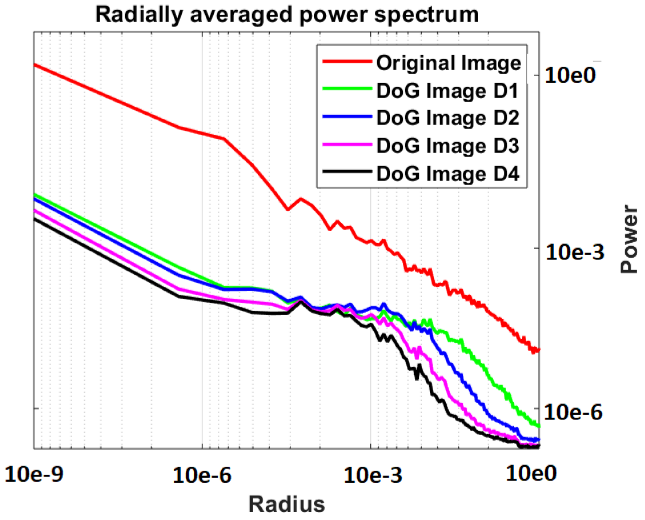


Figure 5: Radially averaged Power Spectrum of 100 images

of information compared to the original pixel domain image. That's why it super-resolved DoG images are easily learnable compared to the image in pixel domain. As power spectrum value decreases in accordance with the increasing σ value, networks with higher σ value can be learnt with more ease. That's why we reduce the number of ResBlocks as we increase σ. Figure 5 also shows a normalized log plot of the comparison of power spectrum density analysis. It shows the radially averaged power spectrum of the 100 samples of the original image and the DoG images D_1,D_2,D_3,D_4. From the figure, we can see that with the increase of σ, the averaged power spectrum value decreases. So, for easier and faster learning purpose, we opt to drecrease the ResBlocks size with increasing σ value. After trial and error, we optimized the number of Resblocks as 16,12,10,8 respectively for lower to higher σ values for method-1 and 16,12,10,8,6 for method-2 . The depth of layers has been reduced as we increase the value of σ.

Table 2: Average of power spectrum density of images from each dataset.

Dataset	**Original Image**	**DoG Image(σ_1,σ_2)**	**DoG Image(σ_1,σ_2)**	**DoG Image(σ_1,σ_2)**	**DoG Image(σ_1,σ_2)**
CDVS	64.47	47.52	41.39	38.66	37.10
Oxford	63.35	48.10	42.47	39.32	37.81
Paris	65.32	49.90	43.42	40.18	38.55

It is to be noted that in both methods, MSE based loss function is used. However, the target was different. The first method directly learns the DoG images whereas in method-2, we need to compute the DoG images manually once the 5 Gaussian blurred images are predicted from the network.

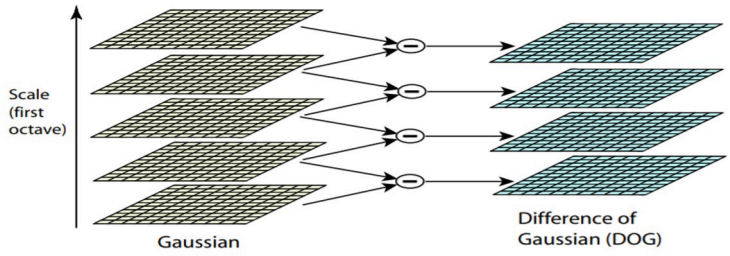


Figure 6: Difference of Gaussian[31]

Integration with SIFT: Once we generate the four DoG images computed , we integrate it to the SIFT network [32]. In SIFT, the DoG images are computed from the Gaussian blurred images with different sigma values in different scaling octaves as shown in Figure 6. Our network produces the DoG images. So, in our case, instead of calculating DoG images by SIFT itself, we directly load our DoG images into the SIFT network. So, the SIFT network will find key points from our produced DoG images. The purpose of integrating with SIFT is that SIFT itself computes DoGs in different scale to find out the maxima and minima in DoG images for identifying key points. As our network already produces super-resolved DoGs, the super-resolution process does not let the images losing their features which will be needed for SIFT while computing the maxima and minma of DoG images. Thus the integration of our DoG images with SIFT actually helps in preserving key features.

3. Experimental Setup and Dataset

A. Training Dataset:

For training, we used the CVPR DIV 2K dataset [33] with 800 training images. We first downscaled the images by both 2 and 4 times. The input images are then cropped to 32X32 patch size. The training process is conducted in a computer equipped with Intel I-7 at 3.2 GHz with 32 GB memory with GPU. The coding platform we used here

is Python with PyTorch [34] deep learning tool. We implemented the architecture and processed in PyTorch.

B. Testing Dataset:

For testing, we used the MPEG Compact Descriptors for Visual Search (CDVS) dataset [28], Oxford building dataset [29] and Paris dataset [30]. CDVS is a comprehensive collection of images of various objects which consists of 186k labeled images of CDs and book covers, paintings, video frames, buildings and common objects as shown in Figure 7(a). We experimented on all the categories of the dataset separately and chose 200 matching image pairs from each one. Oxford building dataset has 5062 images with 55 queries as shown in Figure 7(b) and Paris dataset has 6412 images with 12 queries as shown in Figure 7(c). From Oxford and Paris dataset, we also chose 200 matching image pairs for evaluation. We used our trained networks for generating the upscaled DoG images and then integrated to SIFT. For performance evaluation, we show the number of SIFT matching points.

4. Results

For the evaluation of the performance, we basically compare our result with bicubic interpolation and EDSR that generate upscaled image. We categorize the CDVS dataset into buildings,graphics(books, cards, CDs, DVDs, print), objects, videos and paintings. We collected 200 matching image pairs from each category and evaluated the performance. We also tested the method against Oxford and Paris Dataset with 200 matching image pairs from each one. We compared the PSNR of our predicted DoG images with the DoG images produced from EDSR, SRCNN, SRGAN and bicubic interpolated images for both 2X and 4X upscaling.

Table 3: PSNR(in dB) comparison of DoG Images for 2X and 4X upscaling for CDVS full dataset.

DoG (σ_k,σ_{k+1})	Upscaling Factor	Proposed Method-1	Proposed Method-2	EDSR	SRCNN	SRGAN	Bi-cubic
D_1 (σ_1,σ_2)	2X	32.60	33.30	31.24	30.6	30.92	30.2
D_1 (σ_1,σ_2)	4X	30.68	31.20	29.15	28.92	29.1	28.65
D_2 (σ_2,σ_3)	2X	36.95	37.60	35.58	34.98	35.3	34.75
D_2 (σ_2,σ_3)	4X	34.73	35.68	33.53	33.22	33.40	33.05
D_3 (σ_3,σ_4)	2X	43.80	44.75	42.48	42.05	42.45	41.5
D_3 (σ_3,σ_4)	4X	39.48	40.6	38.15	37.90	38.02	37.68
D_4 (σ_4,σ_5)	2X	47.65	48.38	46.12	45.81	45.95	45.55
D_4 (σ_4,σ_5)	4X	45.12	45.9	43.60	43.31	43.47	43.15

(a) CDVS Dataset[28] .

(b) Oxford Building Dataset[29] .

(c) Paris Dataset[30] .

Figure 7: Experimental Datasets

Table 4: PSNR(in dB) comparison of DoG Images for 2X and 4X upscaling for Paris dataset.

DoG (σ_k,σ_{k+1})	Upscaling Factor	Proposed Method-1	Proposed Method-2	EDSR	SRCNN	SRGAN	Bi-cubic
D_1 (σ_1,σ_2)	2X	33.67	34.15	32.24	31.72	31.95	31.38
D_1 (σ_1,σ_2)	4X	30.97	31.6	29.65	29.41	29.54	29.19
D_2 (σ_2,σ_3)	2X	37.58	38.24	36.32	35.68	35.96	35.45
D_2 (σ_2,σ_3)	4X	35.18	36.05	33.98	33.62	33.79	33.45
D_3 (σ_3,σ_4)	2X	44.55	45.10	43.15	42.91	43.06	42.65
D_3 (σ_3,σ_4)	4X	39.91	41.10	38.90	38.47	38.85	38.31
D_4 (σ_4,σ_5)	2X	47.90	48.76	46.70	46.13	46.46	45.80
D_4 (σ_4,σ_5)	4X	45.55	46.36	44.2	43.83	43.98	43.61

Table 5: PSNR(in dB) comparison of DoG Images for 2X and 4X upscaling for Oxford dataset.

DoG (σ_k,σ_{k+1})	Upscaling Factor	Proposed Method-1	Proposed Method-2	EDSR	SRCNN	SRGAN	Bi-cubic
D_1 (σ_1,σ_2)	2X	31.9	32.65	30.56	30.12	30.08	29.47
D_1 (σ_1,σ_2)	4X	30.03	30.63	28.70	28.36	28.42	28.25
D_2 (σ_2,σ_3)	2X	36.61	37.15	35.32	34.79	34.72	34.30
D_2 (σ_2,σ_3)	4X	33.85	34.90	32.76	32.65	32.64	32.55
D_3 (σ_3,σ_4)	2X	43.05	44.02	41.55	40.98	41.10	40.62
D_3 (σ_3,σ_4)	4X	39.12	40.19	37.80	37.54	37.62	37.31
D_4 (σ_4,σ_5)	2X	46.93	47.55	45.32	45.03	45.19	44.80
D_4 (σ_4,σ_5)	4X	44.71	45.38	43.22	42.91	42.97	42.61

Table 3 shows the result for PSNR in dB for four DoG images generated blurred at σ_k and σ_{k+1} (k=1,2,3,4,5) using our proposed methods, DoG images generated from EDSR images convolved with Gaussian filters,DoG images generated from SRCNN images convolved with Gaussian filters,DoG images generated from SRGAN images convolved with Gaussian filters and DoG images generated from bi-cubic interpolated images convolved with Gaussian filters for 2X and 4X upscaling for the CDVS full dataset. It is crystal clear that DoG images from our proposed method-1 have acquired around 1.7 -2.3 dB gain for 2X and 1.6-1.9 dB gain for 4X upscaling over the DoG images generated from original EDSR convolved with Gaussian filter,2.1 -2.7 dB gain for 2X and 1.8-2.2 dB gain for 4X upscaling over the DoG images generated from original SRCNN convolved with Gaussian filter,1.9 -2.6 dB gain for 2X and 1.7-2.1 dB gain for 4X upscaling over the DoG images generated from original SRGAN convolved with Gaussian filter and 2.1-2.8 dB gain for 2X and 2.0-2.3 dB gain for 4X upscaling over bi-cubic interpolation. We can also see that DoG images from our proposed method-2 has acquired around 2 -2.3 dB gain for 2X and 2-2.4 dB gain for 4X upscaling over the DoG images generated from original EDSR convolved with Gaussian filter, 2.4 -2.7 dB gain for 2X and 2.3-2.6 dB gain for 4X upscaling over the DoG images

generated from original SRCNN convolved with Gaussian filter, 2.2 -2.5 dB gain for 2X and 2.1-2.5 dB gain for 4X upscaling over the DoG images generated from original SRGAN convolved with Gaussian filter and 2.7-3.5 dB gain for 2X and 2.5-2.9 dB gain for 4X upscaling over bi-cubic interpolation.

Table 4 shows the result for PSNR in dB for four DoG images blurred at σ_k and σ_{k+1} (k=1,2,3,4,5) using our proposed methods, DoG images generated from EDSR images convolved with Gaussian filters and DoG images generated from bi-cubic interpolated images convolved with Gaussian filters for 2X and 4X upscaling for the Paris dataset. It is viewed that DoG images from our proposed method-1 have acquired around 1.2 -1.5 dB gain for 2X and 1.0-1.4 dB gain for 4X upscaling over the DoG images generated from original EDSR convolved with Gaussian filter, 1.5 -1.9 dB gain for 2X and 1.2-1.7 dB gain for 4X upscaling over the DoG images generated from original SRCNN convolved with Gaussian filter, 1.4 -1.8 dB gain for 2X and 1.1-1.5 dB gain for 4X upscaling over the DoG images generated from original SRGAN convolved with Gaussian filter and 1.8-2.3 dB gain for 2X and 1.7-2.3 dB gain for 4X upscaling over bi-cubic interpolation. It is also seen that DoG images from our proposed method-2 has acquired around 1.9 -2.1 dB gain for 2X and 1.9-2.2 dB gain for 4X upscaling over the DoG images generated from original EDSR convolved with Gaussian filter, 2.3 -2.5 dB gain for 2X and 2.1-2.4 dB gain for 4X upscaling over the DoG images generated from original SRCNN convolved with Gaussian filter, 2.1 -2.4 dB gain for 2X and 2.0-2.3 dB gain for 4X upscaling over the DoG images generated from original SRGAN convolved with Gaussian filter and 2.5-2.9 dB gain for 2X and 2.4-2.8 dB gain for 4X upscaling over bi-cubic interpolation.

Table 5 shows the result for PSNR in dB for four DoG images blurred at σ_k and σ_{k+1} (k=1,2,3,4,5) using our proposed methods, DoG images generated from EDSR images convolved with Gaussian filters and DoG images generated from bi-cubic interpolated images convolved with Gaussian filters for 2X and 4X upscaling for the Oxford dataset. We can see that DoG images from our proposed method-1 have acquired around 1.3 -1.6 dB gain for 2X and 1.1-1.4 dB gain for 4X upscaling over the DoG images generated from original EDSR convolved with Gaussian filter, 1.7 -2.1 dB gain for 2X and 1.3-1.6 dB gain for 4X upscaling over the DoG images generated from original SRCNN convolved with Gaussian filter, 1.6 -2.1 dB gain for 2X and 1.2-1.6 dB gain for 4X upscaling over the DoG images generated from original

SRGAN convolved with Gaussian filter and 2.4-3 dB gain for 2X and 1.8-2.4 dB gain for 4X upscaling over bi-cubic interpolation. It is also viewed that DoG images from our proposed method-2 has acquired around 1.8 -2.4 dB gain for 2X and 1.9-2.2 dB gain for 4X upscaling over the DoG images generated from original EDSR convolved with Gaussian filter, 2.1 -2.8 dB gain for 2X and 2.1-2.5 dB gain for 4X upscaling over the DoG images generated from original SRCNN convolved with Gaussian filter, 2.0 -2.7 dB gain for 2X and 2.1-2.4 dB gain for 4X upscaling over the DoG images generated from original SRGAN convolved with Gaussian filter and 2.5-3.0 dB gain for 2X and 2.4-2.8 dB gain for 4X upscaling over bi-cubic interpolation.

Figure 8 shows the comparison of DoG images using proposed method-2, EDSR, SRGAN,SRCNN and Bi-cubic interpolation. It is viewed that DoG image using proposed method-2 has the best PSNR which is 50.17 dB.

Table 6: Average number of SIFT matching points for 200 matching image pairs from each category of the CDVS full dataset.

Category	Factor	Original	Proposed Method-1	Proposed Method-2	EDSR	SRCNN	SRGAN	Bi-cubic
Buildings	2X	125.8	124.5	130.4	116.3	114.5	115.8	112.4
Buildings	4X	125.8	110.8	115.4	105.6	104.2	104.3	100.4
Graphics	2X	101.6	99.8	102.8	94.5	93.8	94.2	92.8
Graphics	4X	101.6	87.2	90.4	86.7	86.1	85.8	85.4
Objects	2X	115.3	113.9	118.5	106.9	103.9	104.8	102.6
Objects	4X	115.3	105.1	108.8	99.1	98.2	98.5	96.2
Paintings	2X	114.4	114.7	120.5	105.9	104.4	104.9	100.7
Paintings	4X	114.4	106.1	109.8	101.5	100.1	100.2	96.1
Video	2X	94.3	90.3	94.4	87.2	86.2	85.8	85.2
Video	4X	94.3	82.2	85.5	80.1	79.4	79.6	79.2

Table 7: Average number of SIFT matching points for 200 matching image pairs of the Paris dataset.

Factor	Original	Proposed Method-1	Proposed Method-2	EDSR	SRCNN	SRGAN	Bi-cubic
2X	110.5	107.9	113.2	101.4	99.2	99.8	99.1
4X	110.5	99.8	102.4	97.1	95.4	95.9	95.3

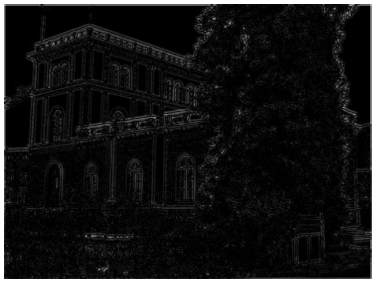
Table 8: Average number of SIFT matching points for 200 matching image pairs of the Oxford dataset.

Factor	Original	Proposed Method-1	Proposed Method-2	EDSR	SRCNN	SRGAN	Bi-cubic
2X	105.4	101.4	107.3	97.1	96.2	96.4	94.2
4X	105.4	93.2	97.8	91.1	90.4	90.3	89.9

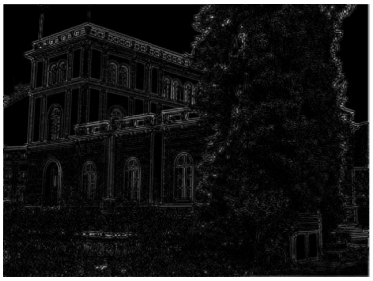
As we generate DoG images, we integrate them into SIFT to generate SIFT repeatability. In Table 6, the result is shown for five different categories for the average number of SIFT matching points from generated super-resolved gradient images using our proposed methods, EDSR method and bi-cubic interpolation method and also the original images which are already high-resolution images for 200 matching image pairs. For 2X upscaling, proposed method 1 is having a gain of around 3-8 points over EDSR, 5-10 points over SRCNN, 4-9 points over EDSR and 5-12 points over bi-cubic interpolation. For 4X upscaling, the gain is around 2-5 points over EDSR, 3-7 points over SRCNN, 3-7 points over SRGAN and 3-10 points over bi-cubic interpolation. For 2X upscaling, proposed method 2 is having a gain of around 7-14 points over EDSR, 9-16 points over SRCNN, 8-15 points over SRGAN and 9-18 points over bi-cubic interpolation. For 4X upscaling, the gain is around 4-10 points over EDSR, 5-12 points over SRCNN, 5-11 points over SRGAN and 5-15 points over bi-cubic interpolation. The best result is achieved with the proposed method-2 in the buildings category with 10-14 points, 12-16 points, 11-15 points and 15-18 points gain over EDSR, SRCNN, SRGAN and bi-cubic respectively. The worst result is achieved with the proposed method 1in the graphics and video categories with a gain of around 2-4 points gain. The goal of our proposed method is to produce SIFT repeatability rather than constructing an end to end system for full recognition. The SIFT repeatability bears the testimony that the produced images have more matching feature points which contribute for recognition.

Table 7 shows the result for the average number of SIFT matching points from generated super-resolved gradient images using our proposed methods, EDSR method and bi-cubic interpolation method and also the original images which are already high-resolution images for 200 matching image pairs of the Paris dataset. For 2X upscaling, proposed method 1 has a gain of around 6 points over EDSR, 8 points over SRCNN, 7 points over SRGAN and 8 points over bi-cubic interpolation. For 4X upscaling, the gain is 2 points over EDSR, 4 points over SRCNN, 3 points over SRGAN and 4 points over bi-cubic interpolation. Proposed method 2 has a gain of around 12 points over EDSR, 12 points over SRCNN, 11 points over SRGAN and 14 points over bi-cubic interpolation for 2X upscaling. For 4X upscaling, the gain is 5 points over EDSR, 7 points over SRCNN, 6 points over SRGAN and 7 points over bi-cubic interpolation.

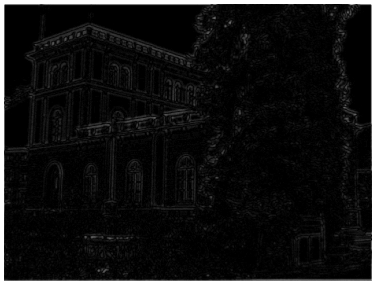
Table 8 shows the result for the average number of SIFT matching points from



(a) DoG image generated from original image



(b) DoG Image using proposed method-2 (50.17 dB PSNR)



(c) DoG Image using EDSR (47.58 dB PSNR)

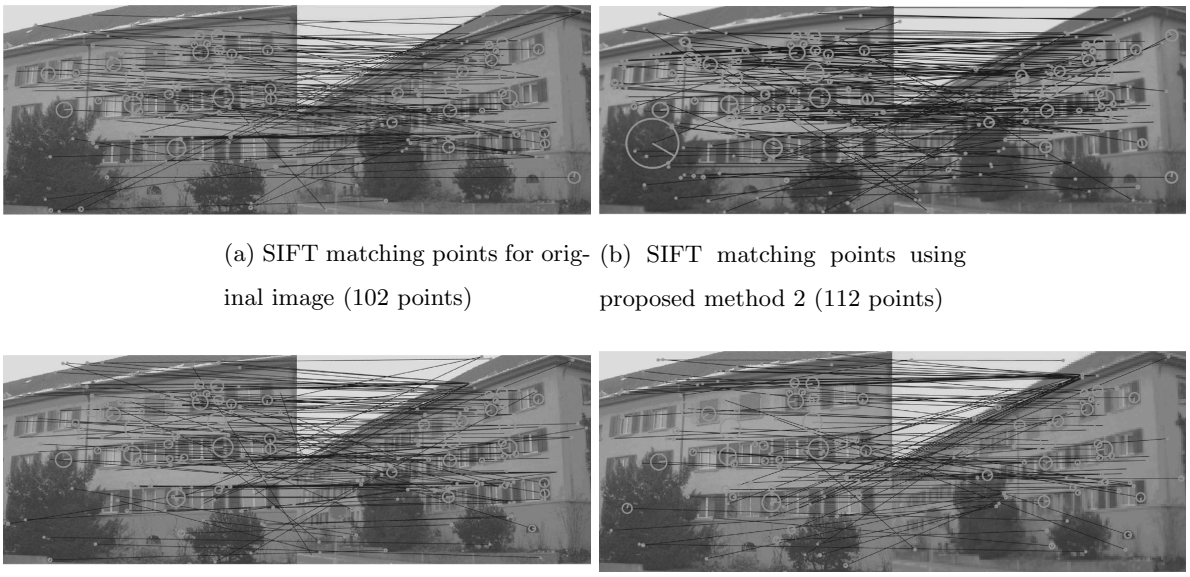
(d) DoG Image using SRGAN (46.97 dB PSNR)

(e) DoG Image using SRCNN (46.69 dB PSNR)

(f) DoG Image using Bi-cubic interpolation (44.91 dB PSNR)

Figure 8: DoG PSNR Comparison

generated super-resolved gradient images using our proposed methods, EDSR method



(a) SIFT matching points for original image (102 points)

(b) SIFT matching points using proposed method 2 (112 points)

(c) SIFT matching points using EDSR (100 points)

(d) SIFT matching points using bi-cubic interpolation(96 points)

Figure 9: SIFT Matching Points Comparison

and bi-cubic interpolation method and also the original images which are already high-resolution images for 200 matching image pairs of the Oxford dataset. For 2X upscaling, proposed method 1 has a gain of around 4 points over EDSR, 4 points over SRCNN, 5 points over SRGAN and 7 points over bi-cubic interpolation. For 4X upscaling, the gain is 2 points over EDSR, 4 points over EDSR, 3 points over SRGAN and 4 points over bi-cubic interpolation. Proposed method 2 has a gain of 10 points over EDSR, 12 points over SRCNN, 11 points over SRGAN and 13 points over bi-cubic interpolation for 2X upscaling. For 4X upscaling, the gain is 6 points over EDSR, 8 points over SRCNN, 7 points over SRGAN and 8 points over bi-cubic interpolation.

In Table 6,Table 7,Table 8,in comparison with the original image, our proposed method 2 achieves approximately 0.1-4 more matching points than the original image for 2X upscaling factor. The reason is that while super-resolving from lower resolution image, the Gaussian blurred image stored the information of the features more rigorously. So after computing the DoG, SIFT feature extraction method finds more

maxima and minima while discovering key points. That's why the super-resolved gradient image version can achieve more SIFT matching points than the original high-resolution image for 2X upscaling. The goal of our proposed method is to produce SIFT repeatability rather than constructing an end to end system for full recognition. The SIFT repeatability bears the testimony that the produced images have more matching feature points which contribute for recognition.

Figure 9 shows the images of SIFT matching points for original image, image generated using proposed method-2 (which ahs better gain than method-1), EDSR and bi-cubic interpolation for 2X upscaling. We can see that our method is producing more points than EDSR for 4X upscaling with 11 points gain over EDSR and 15 points gain over bi-cubic for the sample image. As our input to the SIFT method is the DoG images, one of the template from the generated blurred image is used for showing the matching points.

In comparison, our proposed method-2 performs better than proposed method-1 in all the aspects. The reason is that proposed method-2 is a simplified version of method-1. Method-1 directly computes the upscaled DoG images. So, the networks needs to learn the difference of the Gaussian blurred images in terms of MSE loss function. But method-2 constructs the upsclaed Gaussian blurred image first and then computes the DoG images from simple subtraction. The prediction of upsclaed Gaussian blurred image is easier than the prediction of upsclaed DoG image. Hence, method-2 learns the output more accurately than method-1. Although, both of proposed methods shows significant gain over the state of method in terms of producing SIFT repeatability, proposed method-1 is preferable for its less complexity and accuracy.

5. Conclusion and Future Work

Improving low-resolution and quality image recognition performances has a lot of values in the real world vision, navigation and surveillance applications. In this work, we developed a deep learning framework for gradient image super resolutions at multiple scales. This improved the super-resolving network Degree of Freedom (DoF) by allowing gradient images at different scales to be super-resolved by different networks, with good performance gains in low-resolution key points detection and repeatability, compared with the state of the art pixel domain super-resolving solutions. Next, we pan to optimize the network structure, including new architectures like U- Net, and

also investigate gradient image enhancement with the presence of noises and low light conditions, to have a full suite solution for the low-resolution/quality image recognition.

In the future we will further extend the framework to combat quantization and communication losses in image communication, for the subsequent vision tasks with a task-integrated deep learning solution.

6. Acknowledgement

This research was supported in part by the U.S. Air Force Office of Scientific Research under the Dynamic Data Driven Applications Systems (DDDAS) Program.

References

[1] DDDAS, Dynamic Data Driven Application Systems. Online Available:http://www.1dddas.org/, 2000.

[2] W. Siu, K. Hung, Review of image interpolation and super-resolution, in: Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference, 2012, pp. 1–10.

[3] R. Matsuoka, M. Sone, N. Sudo, H. Yokotsuka, Comparison of image interpolation methods applied to least squares matching, in: 2008 International Conference on Computational Intelligence for Modelling Control Automation, 2008, pp. 1017–1022. doi:10.1109/CIMCA.2008.107.

[4] J. Yang, J. Wright, T. S. Huang, Y. Ma, Image super-resolution via sparse representation, IEEE Transactions on Image Processing 19 (11) (2010) 2861–2873. doi:10.1109/TIP.2010.2050625.

[5] E. M. P. Zeyde, R., single image scale-up using sparse-representations, in: International Conference on Curves and Surfaces, 2012, pp. 711–730.

[6] 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. doi:10.1109/TPAMI.2015.2439281.

[7] 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.

[8] 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.

[9] Z. Wang, S. Chang, Y. Yang, D. Liu, T. S. Huang, Studying very low resolution recognition using deep networks, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4792–4800. doi:10.1109/CVPR.2016.518.

[10] C. Ledig, L. Theis, F. Huszr, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, W. Shi, Photo-realistic single image super-resolution using a generative adversarial network, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 105–114. doi:10.1109/CVPR.2017.19.

[11] S. Prasad Mudunuri, S. Sanyal, S. Biswas, Genlr-net: Deep framework for very low resolution face and object recognition with generalization to unseen categories, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2018.

[12] N. Reddy, D. F. Noor, Z. Li, R. Derakhshani, Multi-frame super resolution for ocular biometrics, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2018, pp. 566–5668. doi:10.1109/CVPRW.2018.00086.

[13] Z. Ye, Y. Pei, J. Shi, An adaptive algorithm for harris corner detection, in: 2009 International Conference on Computational Intelligence and Software Engineering, 2009, pp. 1–4. doi:10.1109/CISE.2009.5366231.

[14] H. Kong, H. C. Akakin, S. E. Sarma, A generalized laplacian of gaussian filter for blob detection and its applications, IEEE Transactions on Cybernetics 43 (6) (2013) 1719–1733. doi:10.1109/TSMCB.2012.2228639.

[15] C. Geng, X. Jiang, Sift features for face recognition, in: 2009 2nd IEEE International Conference on Computer Science and Information Technology, 2009, pp. 598–602. doi:10.1109/ICCSIT.2009.5234877.

[16] Z. Zhang, L. Li, Z. Li, H. Li, Visual query compression with locality preserving projection on grassmann manifold, in: 2017 IEEE International Conference on Image Processing (ICIP), 2017, pp. 3026–3030. doi:10.1109/ICIP.2017.8296838.

[17] J. Luo, Y. Ma, E. Takikawa, S. Lao, M. Kawade, B. Lu, Person-specific sift features for face recognition, in: 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07, Vol. 2, 2007, pp. II–593–II–596. doi:10.1109/ICASSP.2007.366305.

[18] A. Majumdar, R. K. Ward, Discriminative sift features for face recognition, in: 2009 Canadian Conference on Electrical and Computer Engineering, 2009, pp. 27–30. doi:10.1109/CCECE.2009.5090085.

[19] D. F. Noor, Y. Li, Z. Li, S. Bhattacharyya, G. York, Gradient image super-resolution for low-resolution image recognition, in: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 2332–2336. doi:10.1109/ICASSP.2019.8682436.

[20] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, 2018.

[21] Y. Lv, G. Jiang, M. Yu, H. Xu, F. Shao, S. Liu, Difference of gaussian statistical features based blind image quality assessment: A deep learning approach, in: 2015 IEEE International Conference on Image Processing (ICIP), 2015, pp. 2344–2348. doi:10.1109/ICIP.2015.7351221.

[22] AIShack, Convolution. Online Available:http://aishack.in/tutorials/image-convolution-examples/, 2016.

[23] Fourier.Eng, Difference of Gaussian. Online Available:http://fourier.eng.hmc.edu/e161/lectures/gradient/node9.html, 2018.

[24] Z. Wang, A. C. Bovik, Mean squared error: Love it or leave it? a new look at signal fidelity measures, IEEE Signal Processing Magazine 26 (1) (2009) 98–117. doi:10.1109/MSP.2008.930649.

[25] Y. Fan, H. Shi, J. Yu, D. Liu, W. Han, H. Yu, Z. Wang, X. Wang, T. S. Huang, Balanced two-stage residual networks for image super-resolution, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017, pp. 1157–1164. doi:10.1109/CVPRW.2017.154.

[26] K. Hara, D. Saito, H. Shouno, Analysis of function of rectified linear unit used in deep learning, in: 2015 International Joint Conference on Neural Networks (IJCNN), 2015, pp. 1–8. doi:10.1109/IJCNN.2015.7280578.

[27] H. Noh, S. Hong, B. Han, Learning deconvolution network for semantic segmentation, in: 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1520–1528. doi:10.1109/ICCV.2015.178.

[28] S. S. T. N. M. C. H. C. G. T. Y. R. R. V. R. G. J. B. V. Chandrasekhar, D. Chen, B. Girod, The stanford mobile visual search dataset, in: Proceedings of ACM Multimedia Systems Conference (MMSys), San Jose, California, February 2011.

[29] J. Philbin, O. Chum, M. Isard, J. Sivic, A. Zisserman, Object retrieval with large vocabularies and fast spatial matching, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2007.

[30] J. Philbin, O. Chum, M. Isard, J. Sivic, A. Zisserman, Lost in quantization: Improving particular object retrieval in large scale image databases, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2008.

[31] AIShack, SIFT: Theory and Practice. Online Available:http://aishack.in/tutorials/sift-scale-invariant-feature-transform-log-approximation/, 2010.

[32] X. Di, SIFT Feature Extraction. Online Available:https://www.mathworks.com/matlabcentral/ fileexchange/50319-sift-feature-extreaction, 2015.

[33] E. Agustsson, R. Timofte, Ntire 2017 challenge on single image super-resolution: Dataset and study, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2017.

[34] PyTorch, PyTorch Documentation. Online available: https://pytorch.org/docs/stable/index.html, 2016.