# High Frequency Detail Accentuation in CNN Image Restoration

Seyed Mehdi Ayyoubzadeh and Xiaolin Wu, *Fellow, IEEE*

***Abstract*— Given its nature of statistical inference, machine learning methods incline to downplay relatively rare events. But in many applications statistical outliers carry disproportional significance; they can, if being left without special treatment as of now, cause CNNs to perform unsatisfactorily on instances of interests. This is the reason why existing CNN image restoration methods all suffer from the problem of blurred details. To overcome this weakness, we advocate a new training methodology to sensitize the CNNs to desired events even they are atypical. Specifically for image restoration, we propose a so-called high frequency feature accentuation space that promotes image sharpness and clarity by maximally discriminating the ground truth image and the CNN-restored image in atypical but semantically important features. Then we force the restored image to agree with the ground truth image in the feature accentuation space by including an auxiliary loss term in the training process. This aims at a high degree of agreement of the two images on high frequency constructs such as sharp edges and fine textures, i.e., penalizes image blurs. The new CNN design method is implemented and tested for tasks of image super-resolution and denoising. Experimental results demonstrate the achievement of our design objective.**

***Index Terms*— CNN, super resolution, convex optimization, image restoration, semi-definite relaxation, denoising.**

## I. Introduction

THANKS to rapid advances of deep learning research, convolutional neural networks (CNN) have become a ubiquitous method for image restoration and enhancement tasks, including super resolution, denoising, deblurring, etc [25], [27], [34], [38]. However, the existing CNN image restoration methods all have a common weakness: the restored images have blurred details or low contrast compared with the latent pristine images.

There are two reasons for the lack of fidelity in high frequency features of CNN restored images. Foremost, deep learning is an approach of statistical inference; hence CNNs, by nature, favor statistically dominant features. As low-frequency patterns have much higher probabilities of occurrence in natural images, they set a bias of smoothness. The second reason is the use of differentiable norms of error vectors in the objective functions in the CNN training. Minimizing error norms tends to average out similar image waveforms and hence smooth sharp details.

But the occurrence probability is not necessarily proportional to the level of significance in terms of semantics or subjective perception. Neuroscience studies indicate that human vision is built upon fundamental components of the scene encoded by edges (region boundaries), similar to a quick sketch drawn by an artist as an impression [1], [14]. In other words, high frequency features, although having lower probability of occurrence, are nonproportionally important to perception and cognition; therefore, they should be emphasized in image restoration.

The standard counter measure to mitigate the over-smoothing artifacts of the CNN restoration methods is to argument the error norm by a probabilistic divergency loss term. The latter is computed by a generative adversary neural network (GAN) [10] to penalize deviations of the signal distribution of the reconstructed image from that of the ground truth image. But GAN introduces two problems of its own. First, it makes the training process difficult to converge [24]; second, it tends to fabricate unnatural image structures [39].

Troubled by the above weaknesses of the existing GAN methods for image restoration tasks, we set out to find a more effective technique to boost high frequency features in the CNN-restored images without introducing objectionable artifacts. We share the core idea of GANs and search for a space in which the discrimination of the output image of the CNN and the ground truth image is maximized. But unlike GANs, we do not discriminate the two images in the probability distribution space. Instead, we want to find a space in which the CNN restored image and the ground truth image exhibit the maximum discrimination with respect to desired features (e.g., high frequency spatial structures) in the pixel domain. Therefore, successfully passing the discrimination test in this space means a high degree of agreement of the two images in targeted features such as sharp edges and fine textures.

The above idea leads to the main innovation of this paper: the use of a so-called feature accentuation space (FAS), which is spanned by a set of spatially adaptive filters, to promote image sharpness and clarity. The member filters of FAS are designed to maximally discriminate the ground truth image and the CNN-restored image in atypical but semantically important features. These filters are optimized by sample data of the desired features, instead of being manually crafted. Also, the FAS is made to have certain properties so that it

Manuscript received March 6, 2021; revised July 19, 2021, September 4, 2021, and October 5, 2021; accepted October 5, 2021. Date of publication October 21, 2021; date of current version October 28, 2021. This work was supported by the Natural Sciences and Engineering Research Council of Canada. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Dong Xu. *(Corresponding author: Xiaolin Wu.)*

The authors are with the Department of Electrical and Computer Engineering, McMaster University, Hamilton, ON L8S 4L8, Canada (e-mail: xwu@ece.mcmaster.ca).

Digital Object Identifier 10.1109/TIP.2021.3120678

is suited for an auxiliary loss function to be combined with the main CNN objective for whatever the restoration task. The novel FAS-guided CNN restoration system is called feature accentuation network (FANet).

The FAS construction is formulated as an optimization problem. This optimization problem needs to be solved multiple times during the training of the image restoration FANet. Thus, we have to have a fast solution of the underlying optimization problem to facilitate the FAS construction. One of technical contributions of this paper is to convert the original optimization problem to an equivalent one that can be solved efficiently by semi-definite relaxation [5].

In the design of CNNs for image restoration tasks, adding to the objective function an auxiliary loss term defined in the proposed FAS has the following advantages:

- More faithful recovery of sharp edges and fine textures in the restored images without fabricated features.
- A flexible mechanism of incorporating explicit constraints (prior knowledge) into the CNN design, by designing filters to emphasize on the high-frequency structures that are important for given tasks.
- As opposed to GANs, the training process is stable and does not depend on the architecture of the restoration CNN.

More significantly, the way we use an auxiliary loss term in a carefully chosen accentuation space suggests a new training methodology to sensitize the CNN methods to desired events even they have low probability. As all machine learning methods perform statistical inferences using large data, they tend to devote modeling resources mostly to dominant trends in the data at the expense of atypical events. For example, in natural image statistics, smooth transitions are much more common than abrupt discontinuities in the 2D image signal waveform. But in many applications of image processing and computer vision, statistical outliers in the form of rare and unique discontinuous pixel patterns often carry disproportionally important information; they warrant special attention. This research introduces a mechanism to force the CNN methods not to overlook atypical cases that are nevertheless crucial to the intended tasks. In this initial study in the above line of investigation, we focus on image super resolution and denoising tasks; however, our FANet methodology can be easily applied to other image restoration tasks such as deblurring and demoireing. The proposed strategy of sensitizing CNNs for targeted features is general and it may be explored further to boost CNN performances in solving other problems of much biased statistics. Our feature accentuation method is independent of the network architecture and can be coupled with any architecture of CNN.

## II. Related Work

Training CNNs with imbalanced datasets (skewed data distributions) is a well-known issue for classification tasks. In [16] Mako and Henseman proposed over-sampling of the under-represented classes to mitigate this issue. In [26], Lin et. al altered cross-entropy loss to derive a cost function called Focal loss to sensitize the CNN for hard examples. They have used Focal loss for object detection and shown it enforces the CNN to focus more on objects rather than the background while the background is the majority class. However, for image restoration tasks the subject of skewed datasets is quite underdeveloped. Most of the existing works add a fixed term to the loss function that does not depend on the statistical characteristics of the training data. In [19] Johnson et. al suggested an auxiliary loss term based on MSE in the high level feature representation space of the images derived by pretrained VGG network [13]. They called it perceptual loss. There are three concerns about perceptual loss for image restoration tasks: (i) if the distribution of the training images are different from the distribution of the pretrained VGG, then employing this loss is illogical. (ii) One of the incentives about using CNNs is that they are shift invariant to some degree. Therefore, the alignment in the high level feature representation space does not guarantee the alignment in the high frequency components of the signals. (iii) This high level feature representation is fixed and it does not depend on the training data. In [41], Liu *et al.* show that the perceptual loss function can be computed using the networks without any training. However, their so-called Generic Perceptual Loss still suffers from the same weakness as perceptual loss for image restoration tasks.

In [32] Krishna *et al.* have proposed an auxiliary loss function in the edge space for single image super resolution task. They have applied Canny operator to derive the edge map of high-resolution images and ground truths, then computed MSE on these maps. This loss function can recover more details in comparison with MSE. However, this high frequency domain is not optimized based on the outputs of the CNN. Particularly, different frequencies and textures are not considered in designing this loss.

Some researchers have used the loss function of Generative Adversarial Networks (GANs) (GAN loss) [10] for super-resolution task [21] as the auxiliary loss function. Besides the disadvantages of GAN loss for image restoration tasks previously mentioned in I, the GAN discriminator typically has a complex architecture. The training time increases considerably since the only practical approaches to train the CNNs are the first-order optimization methods. In [37] Nazeri *et al.* used two stages of adversarial networks for single image super resolution task. One adversarial stage is used for edge enhancement and the other for image completion. The edge enhancement stage tries to match the distribution of the outputs edges to the distribution of training data edges. In [42] Yang *et al.* pretrained a GAN network and then tried to embed it into another CNN design for face restoration task. In their proposed loss function, one goal is to minimize the difference between the outputs of the discriminator for authentic and restored images ($L_F$). In fact, $L_F$ is similar to the perceptual loss, but the distance is measured in a discriminator space rather than by a pre-trained network. Both above methods also suffer from the same issues of GANs including unintended artifacts and training complexity. In [17] Zhao *et al.* used structural similarity index (SSIM) and multi-scale structural similarity index (MS-SSIM) as the loss function of image restoration CNNs. SSIM and MS-SSIM

are designed to be more aligned with the sensitivity of Human Visual System (HVS). Using SSIM and MS-SSIM as the loss function of the image restoration CNN can lead to more pleasing results for HVS, but still the loss function fails to stress high-frequency details.

Before the deep learning counterpart, there were traditional methods that tried to preserve sharpness of high-frequency features. Banham *et al.* [2] proposed filtering of the images in the 2D wavelet domain. They adjusted the parameters of the filters based on the local information to restore sharp edges. In [9], Naik and Patel proposed a method to promote image sharpness for single image super-resolution and denoising. Their algorithm iteratively used wavelet and spatial domains to minimize the reconstruction error of the back-projected image. In [12], Li *et al.* proposed a method to enhance an image based on a dictionary learning. Their method learnt a dictionary for each block of the image separately. Finally, they try to reconstruct the enhanced image by using adjusted dictionaries for each block. These methods were typically tailored for some niche applications. Very recently, Liu *et al.* developed a hybrid method that combines traditional and CNN approaches. The idea was to use a Weiner-type filter to produce a cartoon-like clean-and-sharp version of the latent image to replace the ground truth in CNN training [40].

Some authors studied the effects of different error norms on perceptual image quality, which is related to the sharpness of details [29], [33]. For $\ell_p$ error norm, a larger value of $p$ exerts a heavier penalty on large errors, regardless of the structure of the image. Which $p$ is most suited for visual quality depends on image structures and on the space in which the error is calculated. This is one of the reasons for us to advocate FAS. Other papers were published to discuss the high-frequency representation learning for image restoration, such as MWCNN [31] and ENet [28]. In terms of basic mechanism, the proposed FAS method, which will be detailed in the next section, is similar to Liu *et al.*'s Orthogonal Network [36]; however, their focus is on network pruning/acceleration in image classification.

## III. FAS Properties

The FAS is the central piece of the proposed CNN image restoration system, because minimizing a loss function in this space promotes the sharpness and clarity of high-frequency details in the restored images. To this end we need to construct the FAS in a way such that it manifests even small differences between the CNN recovered image and the ground truth image in high frequency domain. The FAS is represented by a set of basis filters ($\mathcal{F} = \{\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_M\}$). The filters in this filter bank have the following three key properties:

- Each filter should have band pass or high pass frequency characteristic. This is a necessary feature of the filters in order to extract or emphasize fine details and textures of the images. In absence of the DC component, these filters should satisfy
$$\sum_j \mathbf{f}_{m,j} = 0 \ \ \forall m \tag{1}$$
where $\mathbf{f}_{m,j}$ is the $j$th element of $\mathbf{f}_m$.
- In order to minimize the redundancy between the member filters, or require each member filter to carry new information, we would like to make the set of filters $\{\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_M\}$ a basis that is as orthogonal as possible, namely,
$$|\mathbf{f}_i^T \mathbf{f}_j| \leq \epsilon \ \ \forall i, j, i \neq j \tag{2}$$
With this property, vectors $f_1, f_2, \cdots, f_M$ will span the high frequency domain efficiently, and thus they can represent a rich set of sharp spatial patterns that the existing CNN methods are somewhat inept.
- The filters are learnt from the training data rather than predetermined by an artificial design.
the goal is to learn/discover high frequency structures in natural images in general, or a targeted class of images in particular.
- Finally, we want to make FAS unitary so it is invariant to energy level of member filters. In other words, all basis filters need to be of unit norm:
$$\|\mathbf{f}_m\|_2^2 = 1 \ \ \forall m \tag{3}$$
In this way the FAS filters preserve the energy of the signals.

## IV. Construction of FAS

In this section, we formulate the construction of the FAS as an optimal filter bank design problem. The design objective is to make the estimated and ground truth images maximally differ from each other in the high frequency domain. Let $\mathbf{y}$ and $\hat{\mathbf{y}}$ be the ground truth and the output of the CNN in pixel patch of size $W \times H$, for a fixed filter bank size ($|\mathcal{F}| = M$), the filter bank $\mathcal{F}^*$ constituting the FAS is determined by

$$\begin{aligned} \mathcal{F}^* = \underset{\mathcal{F}}{\text{argmax}} \ & \sum_{m=1}^{M} \sum_{n=1}^{N_s} \|\mathbf{f}_m * (\mathbf{y}_n - \hat{\mathbf{y}}_n(\mathbf{w}))\|_2^2 \\ \text{subject to} \ & \\ & \sum_j \mathbf{f}_{m,j} = 0 \ \ \forall m \\ & |\mathbf{f}_i^T \mathbf{f}_j| \leq \epsilon \ \ \forall i, j, i \neq j \\ & \|\mathbf{f}_m\|_2^2 = 1 \ \ \forall m \end{aligned} \tag{4}$$

where $\mathbf{w}$ is the parameters of the CNN and $N_s$ is a small fraction of the total number of the training data ($N$). The size of each member of the filter bank is $k^2$. As the FAS is much smaller than CNN in terms of the number of parameters, the optimization of the constraining FAS is less prone to overfitting than the optimization of the network itself; a much smaller amount of training data is sufficient to design the FAS. A standard way of solving the non-convex optimization problem Eq(4) is the interior-point (IP) method. However, the IP method for non-convex problems is inefficient and time consuming.

One of our main contributions in this paper is to convert the optimization problem Eq(4) to a form that can be solved more efficiently using mathematical manipulation. This step is necessary since the optimization problem for determining the

FAS is required to be solved repeatedly. We transform and simplify Eq(4) in a way so that the FAS construction problem can be solved efficiently by the Semi-Definite Relaxation (SDR) method [5].

In the following, we outline the required steps for this transformation. We start by simplifying the objective function. To write the objective function of Eq(4) in a more compact form, let $Y_n$ denotes $\mathbf{y}_n - \hat{\mathbf{y}}_n$. Therefore, the objective function is:

$$\sum_{m=1}^{M}\sum_{n=1}^{N_s} \|\mathbf{f}_m * Y_n\|_2^2, \mathbf{f}_m \in \mathbb{R}^{k^2}, Y_n \in \mathbb{R}^{W\times H} \tag{5}$$

For further simplification of Eq(5), it is necessary to write the convolution in the matrix multiplication form. Let $D_{Y_n}$ denotes the doubly block circulant matrix of $Y_n$, the objective function can be rewritten as:

$$\sum_{m=1}^{M}\sum_{n=1}^{N_s} \|D_{Y_n}\mathbf{f}_m\|_2^2, \mathbf{f}_m \in \mathbb{R}^{k^2}, D_{Y_n} \in \mathbb{R}^{l\times k^2} \quad l = (W+k-1)\times(H+k-1) \tag{6}$$

We can simply write Eq(6) in the quadratic form as follows:

$$\sum_{m=1}^{M}\sum_{n=1}^{N_s} \mathbf{f}_m^T D_{Y_n}^T D_{Y_n}\mathbf{f}_m \tag{7}$$

Since $D_{Y_n}^T D_{Y_n}$ is a Positive Semi Definite (PSD) matrix, the objective function in Eq(7) is convex with respects to the parameters of the filters. Next, we need to simplify the orthogonality constraint in order to be able to convert the problem to the standard form. To handle this constraint, we design the FAS filters one by one and then add the orthogonality constraint. In other words, at the time when we want to design $\mathbf{f}_m$, $\{\mathbf{f}_1, \cdots, \mathbf{f}_{m-1}\}$ are determined. In this case, the optimization problem is a nonconvex quadratically constrained quadratic programming (QCQP). To develop $m$th filter, we use the method described in [5] to convert the inhomogeneous QCQP to the homogeneous form.

$$\begin{aligned}
\underset{\mathbf{f}_m}{\text{minimize}} \quad & -\sum_{n=1}^{N_s} \mathbf{f}_m^T D_{Y_n}^T D_{Y_n}\mathbf{f}_m \\
\text{subject to} \quad & \begin{pmatrix}\mathbf{f}_m^T & t_m\end{pmatrix}\begin{pmatrix}\mathbf{f}_m \\ t_m\end{pmatrix} = 2 \ \ \forall m \\
& t_m^2 = 1 \\
& \begin{pmatrix}\mathbf{f}_m^T & t_m\end{pmatrix}\begin{pmatrix}0 & \frac{\mathbf{1}}{2} \\ \frac{\mathbf{1}}{2} & 0\end{pmatrix}\begin{pmatrix}\mathbf{f}_m \\ t_m\end{pmatrix} = 0 \ \ \forall m \\
& \begin{pmatrix}\mathbf{f}_m^T & t_m\end{pmatrix}\begin{pmatrix}0 & \frac{\mathbf{f}_i}{2} \\ \frac{\mathbf{f}_i^T}{2} & 0\end{pmatrix}\begin{pmatrix}\mathbf{f}_m \\ t_m\end{pmatrix} \le \epsilon \ \ , i < m \\
& \begin{pmatrix}\mathbf{f}_m^T & t_m\end{pmatrix}\begin{pmatrix}0 & \frac{\mathbf{f}_i}{2} \\ \frac{\mathbf{f}_i^T}{2} & 0\end{pmatrix}\begin{pmatrix}\mathbf{f}_m \\ t_m\end{pmatrix} \ge -\epsilon \ \ , i < m
\end{aligned} \tag{8}$$

Let $C_n = -D_n^T D_n$, $\mathbf{x}_m = \begin{pmatrix}\mathbf{f}_m \\ t_m\end{pmatrix}$ and $X_m = \mathbf{x}_m\mathbf{x}_m^T$, we can rewrite the problem as follows:

$$\begin{aligned}
\underset{X_m}{\text{minimize}} \quad & \sum_{n=1}^{N_s} Tr(C_n X_m) \\
\text{subject to} \quad & Tr(X_m) = 2 \\
& Rank(X_m) = 1 \\
& X_m \succeq 0 \\
& Tr\left(\begin{pmatrix}\mathbf{0} & \frac{\mathbf{1}}{2} \\ \frac{\mathbf{1}}{2} & 0\end{pmatrix} X_m\right) = 0 \\
& Tr\left(\begin{pmatrix}\mathbf{0} & \frac{\mathbf{f}_i}{2} \\ \frac{\mathbf{f}_i^T}{2} & 0\end{pmatrix} X_m\right) \le \epsilon \ \ , i < m \\
& Tr\left(\begin{pmatrix}\mathbf{0} & \frac{\mathbf{f}_i}{2} \\ \frac{\mathbf{f}_i^T}{2} & 0\end{pmatrix} X_m\right) \ge -\epsilon \ \ , i < m
\end{aligned} \tag{9}$$

where $Tr()$ represents the trace operator. The problem Eq(9) can be solved efficiently using well-known convex optimization techniques SDR or Convex Concave Programming [22]. In the worst case scenario, SDR complexity is $\mathcal{O}(\max\{m,n\}^4 n^{\frac{1}{2}} log(\frac{1}{\epsilon}))$, where $m$ is the number of constraints, $n$ is the dimension of the problem and $\epsilon$ is the given solution accuracy [5].

## V. FANET CONSTRUCTION

Having the FAS basis filters, we are now ready to describe how to train the feature accentuation network FANet for image restoration. In order to train FANet to learn a restoration mapping that avoids blurred high-frequency details, we add an accentuation penalty term to the objective function that is the discrepancy between the output and ground truth in FAS. As the disagreement level in FAS drops in the iterative training process, the reconstruction fidelity of the desired high-frequency patterns increases. The above FAS loss function for the CNN restoration task is:

$$\mathcal{L}_{\text{FAS}}(\mathbf{w}) = \frac{1}{MN}\sum_{n=1}^{N}\sum_{m=1}^{M} \|\mathbf{f}_m * \mathbf{y}_n - \mathbf{f}_m * \hat{\mathbf{y}}_n(\mathbf{w})\|^2 \tag{10}$$

where $F_m$ is the $m$th filter of the FA filter bank. The final loss function of the CNN is a convex combination of the main loss $\mathcal{L}_{\text{MSE}}$ (e.g., the ubiquitous Euclidean norm) and the FA auxiliary loss term $\mathcal{L}_{\text{FAS}}$:

$$L(\mathbf{w}) = (1-\alpha)\mathcal{L}_{\text{MSE}} + \alpha\mathcal{L}_{\text{FAS}} \tag{11}$$

The CNN is trained by minimizing Eq(11) concerning its parameters via backpropagation.

The entire procedure to design the FAS and train the FANet is summarized in Figure 1. As shown, it is a two-stage iterative training process. In the first stage, we design the accentuation control module by solving Eq(9) and adjust the loss function $\mathcal{L}_{\text{FAS}}$ of the FANet; afterwards, we sensitize the FANet to

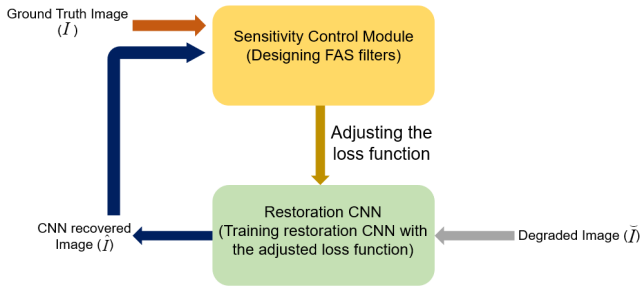
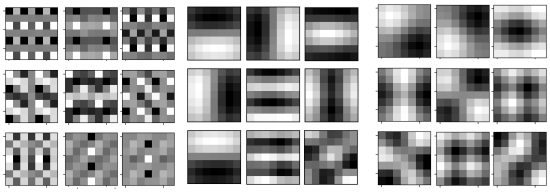



Fig. 1. Schematic description of the FANet construction process.



(a) epoch 0 (b) epoch 50 (c) epoch 200

Fig. 2. The changes of FAS member filters during the FANet training process.

chosen textures and patterns by training it with the adjusted loss function described in Eq(11). We repeat this procedure to continue the training process. Note that the architecture of the restoration CNN can be any type of neural networks, as the application sees fit. Since EDSR [25] is one of the best CNN architectures for image super-resolution, we adopt it in our experiments.

It is helpful to appreciate the advantage of feature accentuation by observing the changes of the FAS filters during the training process, as shown in Figure 2. In the initial stages of the training, outputs of the CNN lacks the capability to recover complex types of textures and details; accordingly the beginning states of the FAS filters are random bandpass and highpass. As the FANet learns to restore high-frequency details with increasing sharpness and clarity, the FAS member filters gradually adjust themselves to fit target textures of certain frequencies and orientations, which the existing CNN methods fail to recover properly.

## VI. Experiments and Evaluations

In this section, we present empirical evidences to establish the validity of our feature accentuation method and the practical value of FANet. The proposed FANet is tested and evaluated on two of the most investigated image restoration tasks: super resolution and denoising. For both tasks, we use the DIV2K dataset [23], [23] to train the FANet. In addition to the common PSNR and SSIM image quality metrics, we introduce two other high-pass metrics to quantify the clarity or the detail sharpness of the restored images. The first metric is the so-called high frequency error $E_h$ that is defined as below:

$$E_h(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{W \times H} \|((1 - G_\sigma) * \mathbf{y} - (1 - G_\sigma) * \hat{\mathbf{y}})\|^2 \tag{12}$$

where $\mathbf{y}$ and $\hat{\mathbf{y}}$ are the ground truth and output images of the CNN respectively, $G_\sigma$ is the Gaussian low-pass filter of standard deviation $\sigma$ and $W$ and $H$ are the width and height of the image. $E_h$ is a measure for the fidelity of restored high frequency features, such as very sharp and ultra fine details and textures.

TABLE I

Super-Resolution Performance Results on Various Datasets for Different Accentuation Level $\alpha$'s ($\alpha = 0$ Corresponds to the MSE Loss Function)

| Dataset | $\alpha$ | PSNR($db$) | SSIM | $E_h$ | $\Delta(db)$ |
|---|---|---|---|---|---|
| DIV2K (x4) | 0.0 | 27.21 | 0.79 | 88.10 | 39.71 |
| | 0.01 | **27.26** | **0.79** | **86.96** | 39.91 |
| | 0.1 | 27.22 | 0.79 | 87.44 | 40.12 |
| | 0.5 | 27.11 | 0.78 | 87.20 | **40.22** |
| Urban100 (x4) | 0.0 | 22.82 | 0.72 | 226.03 | 38.81 |
| | 0.01 | **22.83** | **0.73** | **223.14** | 39.11 |
| | 0.1 | 22.79 | 0.72 | 225.46 | **39.31** |
| | 0.5 | 22.68 | 0.71 | 227.95 | 39.18 |
| BSD100 (x4) | 0.0 | 24.87 | 0.68 | 142.07 | 29.72 |
| | 0.01 | 24.85 | **0.69** | 141.36 | 29.99 |
| | 0.1 | **24.87** | 0.69 | **140.92** | 30.09 |
| | 0.5 | 24.82 | 0.68 | 141.09 | **30.11** |
| Set14 (x4) | 0.0 | 24.48 | 0.71 | 130.22 | 31.72 |
| | 0.01 | 24.46 | 0.71 | 128.85 | 32.23 |
| | 0.1 | **24.51** | **0.71** | **128.61** | 32.14 |
| | 0.5 | 24.33 | 0.70 | 129.04 | **32.63** |
| Set5 (x4) | 0.0 | **27.54** | 0.83 | 59.28 | 29.19 |
| | 0.01 | 27.53 | 0.83 | 58.94 | 29.51 |
| | 0.1 | 27.49 | **0.83** | **58.12** | **29.95** |
| | 0.5 | 27.36 | 0.82 | 59.34 | 29.57 |

By varying the parameter $\sigma$, we can choose the width of the high frequency subband to emphasize. For instance, increasing $\sigma$ will force the resorted image to match the ground truth image on higher frequency features. The second quality metric is for the overall sharpness of restored images. It is defined to be the absolute energy level of the restored high frequency components:

$$\Delta = 20\, log_{10}(\|(1 - G_\sigma) * \hat{\mathbf{y}}\|) \tag{13}$$

### A. Super Resolution

*1) Experiment Setting:* For the superresolution task, we accentuate the EDSR model in FAS. Adam [11] optimizer is used to train the FANet of 16 residual blocks, with learning rate $10^{-4}$. Specifically in our experiments, FANet for supersolution of scaling factor 4 is implemented; the paired training data are generated by bicubic downsampling process. To focus on local textures, we train FANet with relatively small square patches of width 48. This has the side benefit of faster convergence. Only a subset (100 samples) of the training set are used to design the FAS ($N_s = 100$). The training process is carried out for 200 epochs, with batch size 8. Orthogonality error ($\epsilon$ in Eq(4)) is set to 0.1. For solving the optimization probm 9, we use the CVXPY framework [18], [30]. The CNN tool Keras [15] is used in our implementation. The filter support for the FAS bases ($k$) is set to $7 \times 7$ and the number of filters in FAS ($M$) is 9.

In the interest of statistical significance, we have tested the proposed FANet on as many as five different datasets: DIV2K, Set5 [6], Set14 [7], Urban100 [3] and BSD100 [3]. The last four datasets are unseen by the FANet at the training stage at all. The test results are tabulated for different datasets and different levels of accentuation in Table I. The level of

TABLE II

COMPARISON OF VARIOUS LOSS FUNCTIONS AND METHODS (PERCEPTUAL LOSS ($\mathcal{L}_{\text{VGG}}$ [20], SSIM LOSS ($\mathcal{L}_{\text{SSIM}}$) [29], MS-SSIM LOSS ($\mathcal{L}_{\text{MS-SSIM}}$) [29], ADVERSARIAL LOSS ($\mathcal{L}_{\text{ADV}}$) [28], TEXTURE LOSS ($\mathcal{L}_{\text{TEXTURE}}$) [28]))

| Network | EDSR | | | | | ENet-PAT | MWCNN | SRGAN |
|---|---|---|---|---|---|---|---|---|
| Metric | $\mathcal{L}_{\text{MSE}}$ | $\mathcal{L}_{\text{MSE}}+\mathcal{L}_{\text{SSIM}}$ | $\mathcal{L}_{\text{MSE}}+\mathcal{L}_{\text{MS-SSIM}}$ | $\mathcal{L}_{\text{MSE}}+\mathcal{L}_{\text{VGG}}$ | $\mathcal{L}_{\text{MSE}}+\mathcal{L}_{\text{FAS}}$ | $\mathcal{L}_{\text{VGG}}+\mathcal{L}_{\text{adv}}+\mathcal{L}_{\text{texture}}$ | $\mathcal{L}_{\text{MSE}}$ | $\mathcal{L}_{\text{VGG}}+\mathcal{L}_{\text{adv}}$ |
| **DIV2K** | | | | | | | | |
| PSNR | **27.19** | 26.82 | 26.98 | 24.67 | **27.14** | 27.13 | 24.37 | 18.70 |
| SSIM | 0.81 | **0.83** | 0.82 | 0.75 | **0.82** | 0.82 | 0.78 | 0.69 |
| MS-SSIM | 0.93 | **0.94** | 0.93 | 0.91 | 0.93 | **0.94** | 0.91 | 0.86 |
| $\Delta$ | 39.85 | 40.25 | 39.69 | **41.45** | 40.51 | 39.79 | **41.08** | 39.23 |
| UQI | 0.96 | 0.96 | 0.96 | 0.93 | **0.96** | **0.97** | 0.95 | 0.84 |
| LPIPS (Lower is better) | 0.26 | 0.26 | 0.24 | 0.19 | 0.26 | **0.12** | **0.13** | 0.23 |
| NIQE (Lower is better) | 4.38 | 4.66 | 4.66 | 7.64 | **4.13** | 4.47 | 4.88 | **3.41** |
| **Urban100** | | | | | | | | |
| PSNR | **22.74** | 22.55 | 22.56 | 21.22 | **22.73** | 22.33 | 19.37 | 17.52 |
| SSIM | 0.77 | **0.78** | 0.77 | 0.71 | **0.77** | 0.73 | 0.69 | 0.64 |
| MS-SSIM | 0.91 | 0.91 | 0.91 | 0.89 | **0.91** | **0.91** | 0.85 | 0.83 |
| $\Delta$ | 38.85 | 39.25 | 38.73 | 39.30 | **39.59** | 37.89 | **40.06** | 37.50 |
| UQI | 0.95 | **0.96** | 0.95 | 0.92 | **0.95** | 0.95 | 0.93 | 0.86 |
| LPIPS (Lower is better) | 0.24 | 0.25 | 0.24 | 0.20 | 0.25 | **0.20** | 0.22 | **0.20** |
| NIQE (Lower is better) | **4.18** | 4.41 | 4.45 | 8.33 | 4.22 | 4.63 | 4.98 | **3.57** |
| **BSD100** | | | | | | | | |
| PSNR | **24.84** | 24.53 | 24.72 | 23.10 | **24.84** | 24.78 | 22.43 | 19.72 |
| SSIM | 0.73 | **0.75** | **0.74** | 0.67 | 0.73 | 0.71 | 0.69 | 0.64 |
| MS-SSIM | 0.90 | 0.90 | 0.90 | 0.87 | **0.90** | **0.90** | 0.86 | 0.86 |
| $\Delta$ | 29.94 | 30.48 | 29.81 | **31.89** | 30.56 | 30.41 | **31.29** | 30.65 |
| UQI | 0.97 | 0.97 | 0.97 | 0.95 | **0.98** | **0.97** | 0.96 | 0.90 |
| LPIPS (Lower is better) | 0.34 | 0.33 | 0.32 | 0.28 | 0.35 | **0.17** | **0.17** | 0.23 |
| NIQE (Lower is better) | 6.54 | 6.96 | 7.06 | 10.23 | 5.91 | **5.21** | 6.47 | **4.42** |
| **Set14** | | | | | | | | |
| PSNR | 24.34 | 24.25 | 24.33 | 22.73 | **24.39** | **24.40** | 22.00 | 19.30 |
| SSIM | 0.76 | 0.78 | **0.78** | 0.70 | **0.77** | 0.75 | 0.73 | 0.68 |
| MS-SSIM | 0.91 | **0.92** | **0.92** | 0.89 | 0.91 | 0.91 | 0.88 | 0.87 |
| $\Delta$ | 32.16 | 32.33 | 32.17 | **33.48** | 32.89 | 32.25 | **33.14** | 31.71 |
| UQI | 0.96 | 0.97 | 0.97 | 0.95 | **0.97** | **0.97** | 0.95 | 0.89 |
| LPIPS (Lower is better) | 0.28 | 0.27 | 0.25 | 0.22 | 0.28 | **0.17** | **0.16** | 0.19 |
| NIQE (Lower is better) | 6.02 | 6.47 | 6.57 | 9.92 | 5.46 | **5.25** | 6.66 | **3.88** |
| **Set5** | | | | | | | | |
| PSNR | **27.52** | 27.13 | 27.39 | 24.77 | **27.59** | 26.57 | 24.59 | 21.41 |
| SSIM | 0.88 | **0.89** | 0.88 | 0.81 | **0.88** | 0.85 | 0.83 | 0.77 |
| MS-SSIM | 0.96 | 0.96 | **0.96** | 0.95 | **0.96** | 0.95 | 0.93 | 0.92 |
| $\Delta$ | 29.26 | 28.88 | 28.55 | 30.03 | **30.18** | **30.42** | 30.04 | 27.32 |
| UQI | 0.97 | 0.97 | 0.97 | 0.90 | **0.97** | **0.97** | 0.96 | 0.84 |
| LPIPS (Lower is better) | 0.18 | 0.18 | 0.16 | 0.12 | 0.18 | 0.13 | **0.11** | **0.12** |
| NIQE (Lower is better) | 6.61 | 6.85 | 7.11 | 11.47 | **6.44** | 6.87 | 7.51 | **3.88** |

high-frequency accentuation is quantified by parameter value $\alpha$ in Eq 11. For each dataset the FANet is tested at four levels of accentuation, $\alpha = 0, 0.01, 0.1, 0.5$; for $\alpha = 0$ the FANet reduces to the original EDSR model of no accentuation. Four image quality metrics are used to evaluate the test results, the two ubiquitous metrics PSNR and SSIM, plus the two just introduced high-frequency focused metrics $E_h$ and $\Delta$.

*2) Effect of Accentuation Coefficient ($\alpha$):* As shown in the Table I, the energy of the high frequency components in the restored images is higher by 0.61dB on average if feature accentuation is applied when training the restoration network than if only the MSE loss function is used. We can see that for some values of $\alpha$, the FAS loss function not only improves the HFA, but also it increases PSNR value. In fact, the accentuation term acts as the regularizer of the CNN in such cases.

*3) Evaluation of Other Networks and by Other Quality Metrics:* We further compare FANet with three other networks for high-frequency representation learning, ENet-PAT [28], MWCNN [31] and SRGAN [21]. We also add the EDSR network trained under various perceptual loss functions into the comparison group. In addition to PSNR and SSIM, we include the following image quality metrics: NIQE [8], Multi-scale Structural Similarity Index (MS-SSIM) [43], LPIPS [35], and Universal Quality Image Index (UQI) [4]. Note that, except for the ENet-PAT [28] and MWCNN [31] in which we adopt the original architectures proposed by the authors, the architectures for all EDSR variants are the same as FANet. The results are shown in Table II for different datasets. As we can see, FANet outperforms other methods in most of the performance metrics. NIQE and our sharpness metric $\Delta$ are two non-reference image quality metrics, and they can be used for assessing subjective image quality. On the other hand, PSNR and SSIM are widely used objective image quality metrics. Therefore, (NIQE,PSNR) and ($\Delta$,SSIM) can be used as subjective vs. objective quality metric pairs to evaluate different restoration methods. Figure 3 compares the performances of the evaluated methods in the subjective-objective quality plane (averaged over all datasets).

As illustrated, FANet strikes a better balance between the subjective and objective image quality than other methods, which typically sacrifice one to improve the other. In Figure 4, one can see clear advantage of using FAS loss over other loss functions; FANet can recover sharp details with a negligible amount of artifacts compared to other methods.

*4) Comparison of Perceptual Quality:* The provided performance metrics are not always the best indicator of the visual quality of the images. Therefore, in addition to the quantitative results, let us visually compare the results of the restoration CNN coupled with and without high-frequency

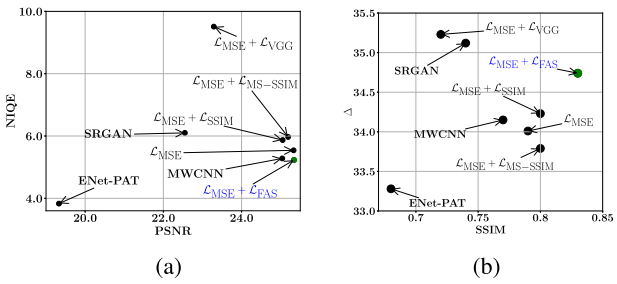



Fig. 3. Objective vs. subjective performance of different methods (lower NIQE values indicate more natural hence better perceptual quality).

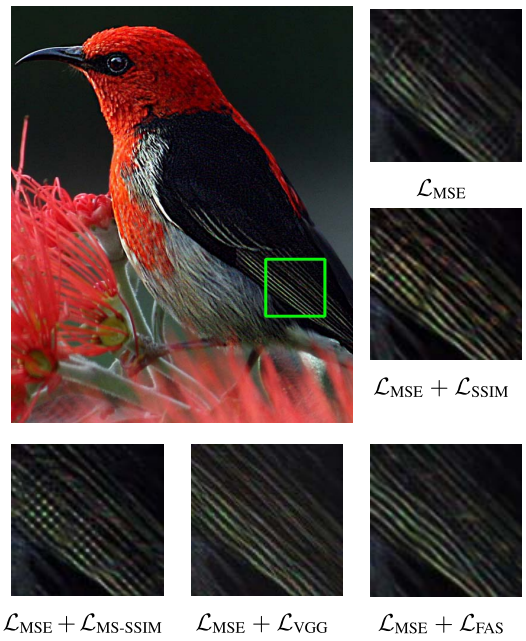



Fig. 4. Comparison between EDSR networks trained with different metrics.

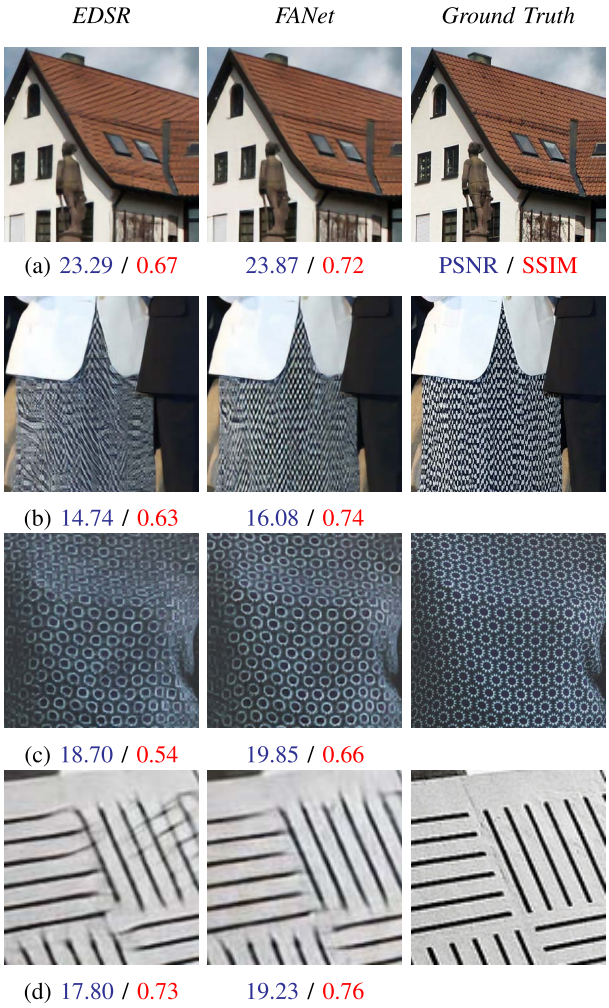



Fig. 5. Visual comparison of EDSR vs. FANet for $\times 4$ super resolution.

feature accentuation. We present some sample output images in Figure 5. As can be seen, when the images contain rich and sharp edges and textures, the CNN trained without accentuation fails to recover them, whereas FANet restores such details successfully. In Figure 5 (a), the EDSR trained with the MSE loss fails to recover the true slope of the lines on the roof as opposed to FANet. In fact, the plain EDSR generates false structures that do not exist at all in the original scene. In Figures 5 (b) and 5 (c), one can see that for more complex textures that are not simple lines, the plain EDSR produces blurry and alias patterns, whereas the corresponding reconstruction of FANet is far superior. Also in Figure 5 (d), the plain EDSR has produced much more artifacts in comparison with FANet. In all of these examples, the FAS accentuation forces the network to recover high-frequency details in order to minimize the FAS loss.

*5) Removal of GAN Artifacts by FAS:* When motivating this research in the introduction, we criticised the common practice of using GAN to generate high-frequency features in image restoration CNNs. Although GAN can alleviate the problem of oversmoothing in CNN-superresolved images by implanting some details, it tends to fabricate false non-existing structures. The FANet method is proposed to fix the above problem, as a new way of restoring sharp details faithfully without the artifacts of GAN. To verify the advantage of FANet over GAN, we need to compare the super-resolution results of FANet and GANs in terms of perceptual quality. To this end, we train a GAN network, in which the generator architecture is the same as the FANet of the previous section, and the discriminator architecture is borrowed from SRGAN [21]. In terms of network size FANet is far more compact than GAN, because the GAN discriminator is an extra part that FANet does not need. Also, the number of parameters in FAS filter bank is far fewer than the number of GAN discriminator parameters.

The superresolution output images of GAN and FANet are presented in Figure 6. As evidently in these figures, the FANet results are visually superior to those of GAN; in particular, FANet is free of the false, objectionable structures that are fabricated by GAN. No users will accept semantically erroneous features in the output image solely for the illusion of more details.

*6) Performance for Lighter Network Architecture:* To demonstrate that the successful learning of high-frequency details is primarily credited to the adoption of FAS loss function rather than a large network size, we test a lighter

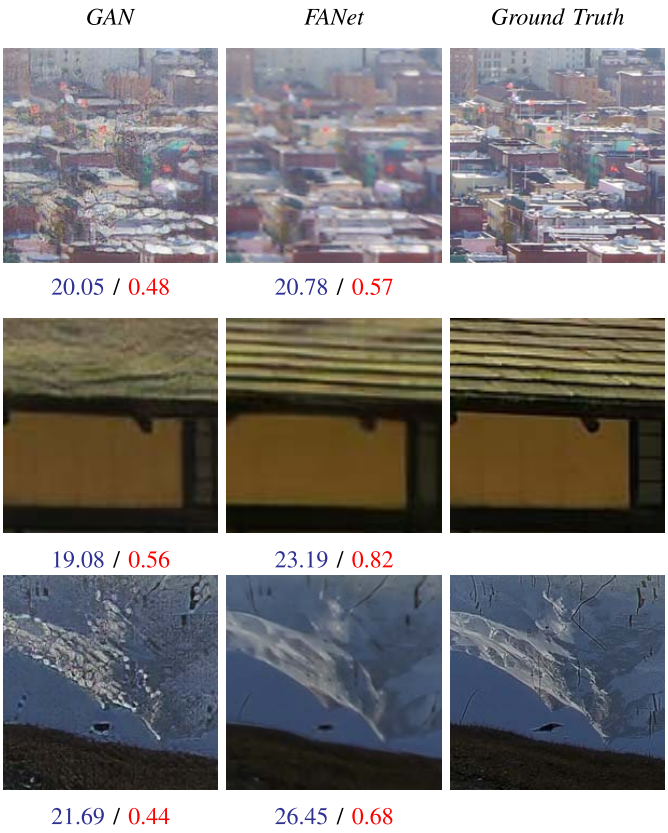



Fig. 6. Visual comparison of GAN vs. FANet for $\times 4$ super resolution.

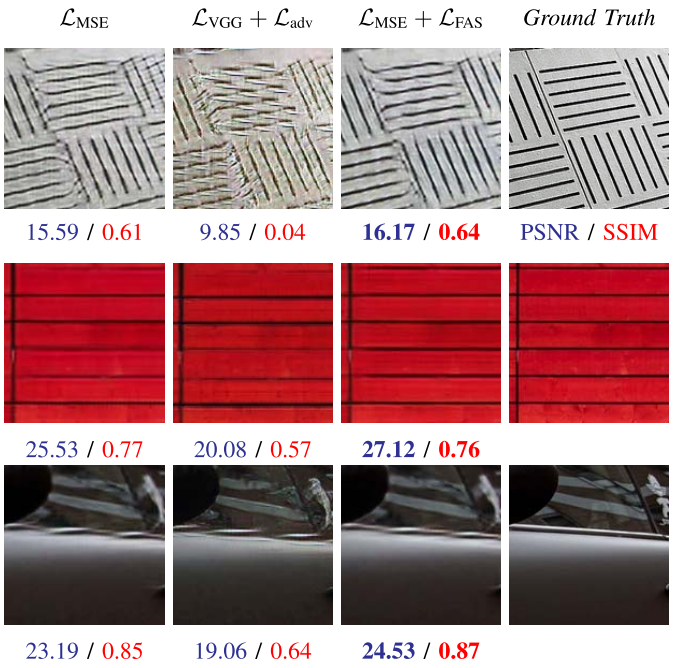



Fig. 7. Comparison of different methods for reduced networks for $\times 4$ super resolution.

TABLE III

PERFORMANCE NUMBERS FOR REDUCED NETWORKS FOR $\times 4$ SUPER-RESOLUTION. THE NUMBERS IN BRACKETS ARE CHANGES DUE TO NETWORK REDUCTION

| Network | EDSR | | GAN |
|---|---|---|---|
| Metric | $\mathcal{L}_{\text{MSE}}$ | $\mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{FAS}}$ | $\mathcal{L}_{\text{VGG}} + \mathcal{L}_{\text{adv}}$ |
| **DIV2K** | | | |
| PSNR | **26.36** [-0.83] | 26.35 [-0.79] | 22.05 |
| SSIM | **0.77** [-0.04] | 0.76 [-0.06] | 0.60 |
| $E_h$ | 92.23 [4.13] | **91.76** [4.8] | 98.74 |
| $\Delta$ | 37.65 [-2.2] | 37.88 [-2.53] | **39.77** |
| **Urban100** | | | |
| PSNR | 21.56 [-1.18] | **21.57** [-1.16] | 17.98 |
| SSIM | 0.68 [-0.09] | **0.68** [-0.09] | 0.47 |
| $E_h$ | 246.14 [20.11] | **245.35** [22.21] | 270.4 |
| $\Delta$ | 36.69 [-2.16] | 36.95 [-3.56] | **38.03** |
| **BSD100** | | | |
| PSNR | 24.06 [-0.78] | **24.06** [-0.78] | 20.56 |
| SSIM | 0.66 [-0.07] | **0.66** [-0.07] | 0.48 |
| $E_h$ | 142.04 [-0.03] | 142.03 [1.11] | **141.03** |
| $\Delta$ | 27.79 [-2.15] | 27.98 [-2.58] | **30.99** |
| **Set14** | | | |
| PSNR | **23.77** [-0.57] | 23.76 [-0.63] | 20.06 |
| SSIM | 0.68 [-0.08] | **0.68** [-0.09] | 0.5 |
| $E_h$ | **129.10** [-1.12] | 129.31 [0.7] | 130.06 |
| $\Delta$ | 30.28 [-1.88] | 30.34 [-2.55] | **31.24** |
| **Set5** | | | |
| PSNR | **26.49** [-1.03] | 26.45 [-1.14] | 21.48 |
| SSIM | 0.80 [-0.08] | **0.80** [-0.08] | 0.60 |
| $E_h$ | 66.88 [7.6] | **66.83** [8.71] | 91.95 |
| $\Delta$ | 27.68 [-1.58] | 28.00 [-2.18] | **28.13** |

version of EDSR that has only 4 residual blocks and 32 filters as opposed to 16 residual blocks and 64 filters in the original model. This reduced EDSR is trained on DIV2K dataset for the $\times 4$ super-resolution task. To compare with GAN, we additionally train the GAN counterpart of the reduced network with the same generator architecture and the discriminator of SRGAN. In Figure 7, one can visually compare some results of this experiment. As shown, using FAS loss to train the reduced network also improves the network's ability to recover fine details and textures. The objective quality metric values of different methods are reported in Table III, in which they are compared with the counterpart numbers before network simplification. One can see that network size reduction does cause performance numbers to drop, but it does not change the relative ranking of different methods. This agrees with the visual comparison in Figure 7. The FAS criterion still delivers higher contrast ($\Delta$) and lower high-frequency error ($E_h$) than MSE.

Note that GAN scores higher in sharpness metric $\Delta$ but has a significantly lower PSNR and SSIM. Although GAN generates high-frequency textures, it performs too poorly in objective quality metrics to keep image semantics intact.

*7) Results on Structured Datasets:* The gains made by the proposed FAS criterion over GAN discriminator in perceptual quality become more pronounced, if the images have some known priors. For example, when superresolving face images, the training process can make use of prior knowledge on the structure, shape and textures of the object in question. We use the Flickr Faces HQ dataset (FFHQ) consisting of 70,000 human face images to evaluate the performances of FAS and GAN. For the $\times$ 4 super-resolution task on this dataset, we have used 4 residual blocks for FANet and the generator network. The discriminator has the same architecture as SRGAN. The results are presented in Figure 8. As illustrated, the GAN-based results appear unnatural and suffer from severe distortions (note the reconstructed noses, mouths and teeth). On the other hand, using EDSR with the MSE loss function alone cannot reconstruct sharp details (see the areas around eyes and mouths). In contrast, FANet successfully recovers

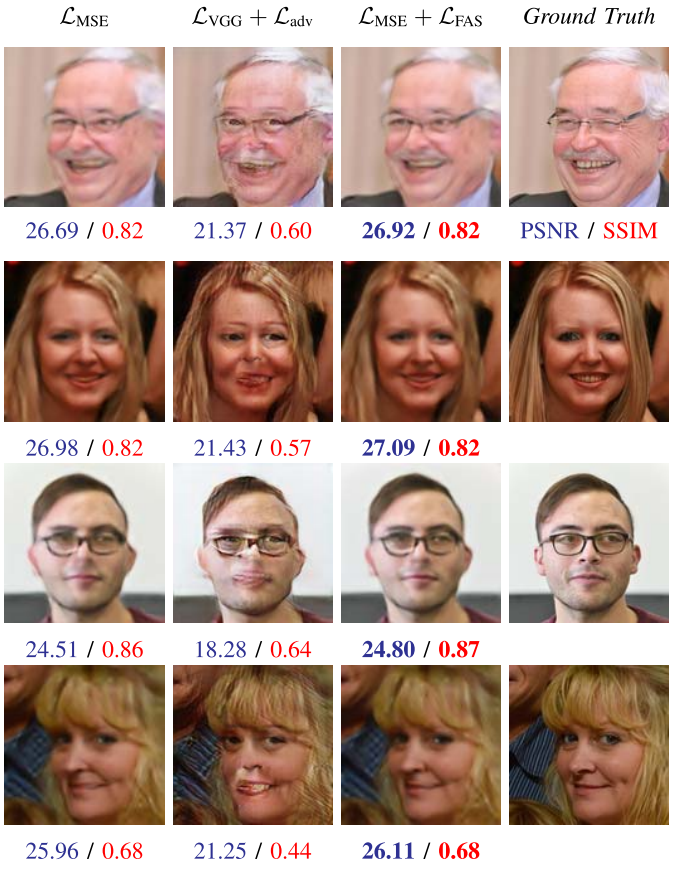



Fig. 8. Results of different methods on a structured dataset (FFHQ) for $\times 4$ super resolution.

these details with clarity and largely free of objectionable artifacts.

### B. Denoising

*1) Experiment Setting:* Another intensively researched image restoration task is denoising. For image denoising methods, including those of deep learning, blurring artifacts are inevitable; they or lack of them largely determine the quality of denoised images. We let our FANet method take up the challenge of preserving the sharpness and clarity of high-frequency details in the CNN denoising process. Specifically, to build the FANet for image denoising, we accentuate the EDSR model without upsampling layers (modified EDSR). The same hyperparameters in the super-resolution experiments are used to train the denoising FANet. The training data are generated by adding zero-mean Gaussian noise with variance ($\sigma^2$) 0.1 to the images. The trained denoising FANet is tested and evaluated below. These experimental results validate the effectiveness of the FAS accentuation method when being applied to CNNs for other image restoration tasks besides super-resolution.

*2) Quantitative Results:* To evaluate the efficacy and robustness of the denoising FANet, we add Gaussian noises of different variances to the validation images. This allows us to check how well the denoising FANet, which is designed for a fixed noise level, performs against different noise levels. The results are presented in Table IV. As shown in the table, when the CNN is integrated with FAS accentuation, all quality metrics improve for image denoising. This is consistent with our observations in the super-resolution case.

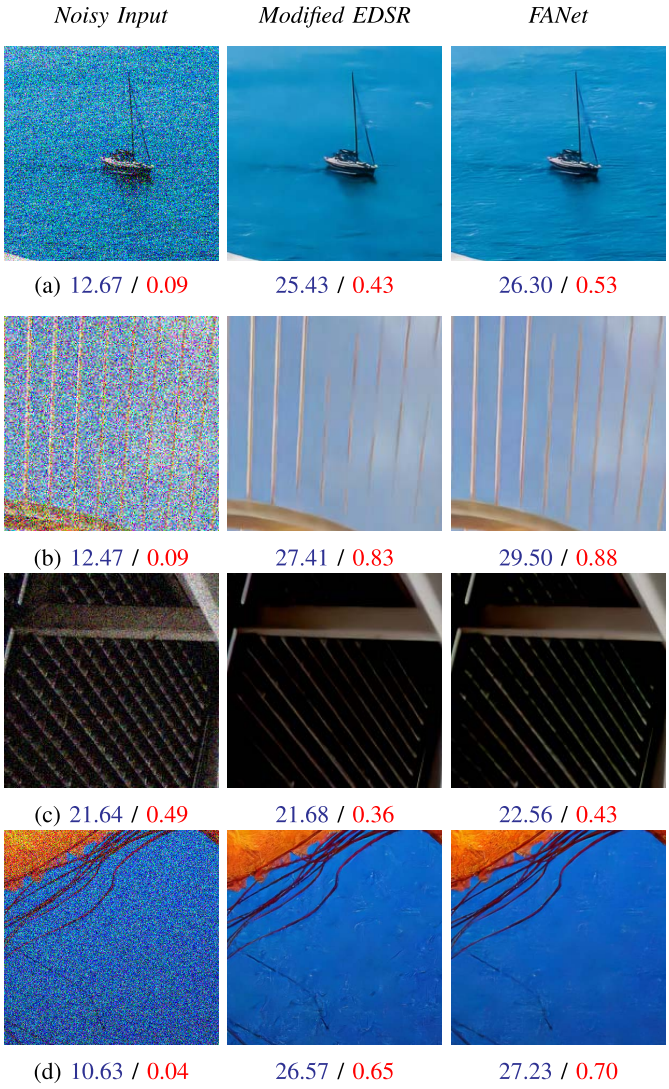



Fig. 9. Samples of denoising results.

TABLE IV
DENOISING PERFORMANCE RESULTS ON VARIOUS NOISE LEVELS FOR DIFFERENT ACCENTUATION LEVELS $\alpha$'S ($\alpha = 0$ CORRESPONDS TO THE MSE LOSS FUNCTION)

| Noise level ($\sigma^2$) | $\alpha$ | PSNR($db$) | SSIM | $E_h$ | $\Delta(db)$ |
|---|---|---|---|---|---|
| 0.01 | 0.0 | 24.27 | 0.66 | 24.40 | 38.40 |
| | 0.01 | 24.25 | **0.67** | 24.31 | **38.79** |
| | 0.07 | 24.03 | 0.66 | 24.30 | 38.75 |
| | 0.1 | **24.45** | 0.67 | **24.10** | 38.51 |
| 0.02 | 0.0 | 24.71 | 0.68 | 24.16 | 38.66 |
| | 0.01 | 24.64 | 0.68 | 24.08 | **38.90** |
| | 0.07 | 24.45 | 0.68 | 24.08 | 38.84 |
| | 0.1 | **24.84** | **0.69** | **23.85** | 38.68 |
| 0.05 | 0.0 | 26.38 | 0.73 | 23.79 | 38.66 |
| | 0.01 | 26.27 | **0.73** | 23.69 | **39.10** |
| | 0.07 | 26.08 | 0.72 | 23.74 | 38.95 |
| | 0.1 | **26.38** | 0.73 | **23.44** | 38.97 |
| 0.1 | 0.0 | 27.68 | 0.78 | 23.09 | 39.70 |
| | 0.01 | 27.66 | 0.78 | 23.01 | 39.85 |
| | 0.07 | 27.66 | 0.78 | 23.17 | 39.85 |
| | 0.1 | **27.68** | **0.78** | **22.85** | **39.89** |

*3) Qualitative Results:* We illustrate samples of the denoised images in Figure 9. As can be seen, the denoising FANet can effectively remove noises and at the same time

it also keeps edges and high-frequency textures sharp and clean. In perceptual quality FANet is clearly superior to the denoising CNN without FAS accentuation (the modified EDSR model of $\alpha = 0$). For example, in Figure 9 (a), the modified EDSR model fails to recover the sea wave texture and flattens the water surface, while FANet has much less over smoothing artifacts and recovers the wave structure approximately. In Figures 9 (b) and (c), the modified EDSR model is not able to recover the thin lines, which do not trouble FANet nearly as much. Similarly, in Figure 9 (d), FANet works equally well in restoring both low-frequency and high-frequency regions; the FANet recovered image is visually much more pleasant than that of the modified EDSR. Although the above comparison studies between with and without feature accentuation are carried out only on the EDSR architecture, the same conclusions should hold for other network architectures, simply because the FAS affects the optimization criterion that is independent of CNN architectures.

## VII. Conclusion

In this paper, we propose a novel design method for image restoration CNNs to achieve sharpness and clarity of high-frequency details. The key innovation is to construct a feature accentuation space that defines desired features and sensitizes reconstruction errors in these features. The FAS construction is done by efficient optimization techniques. As opposed to GANs, which is commonly used to generate high-frequency details in recovered images, the proposed FAS method has lower computational complexity, and more importantly it does not generate nonexistent features as GANs are prone to. Experiments show that our method can improve visual quality of restored images, especially on edges and high textures. The new method is general and it can be applied to many different restoration tasks, including super-resolution, denoising, deblurring, and etc.

## References


[1] R. A. Weale, "Vision. A computational investigation into the human representation and processing of visual information. David Marr," *Quart. Rev. Biol.*, vol. 58, no. 2, p. 299, Jun. 1983, doi: 10.1086/413352.

[2] M. R. Banham and A. K. Katsaggelos, "Spatially adaptive wavelet-based multiscale image restoration," *IEEE Trans. Image Process.*, vol. 5, no. 4, pp. 619–634, Apr. 1996, doi: 10.1109/83.491338.

[3] 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 *Proc. 8th IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 2, Jul. 2001, pp. 416–423.

[4] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, Mar. 2002, doi: 10.1109/97.995823.

[5] Z.-Q. Luo, W.-K. Ma, A. M.-C. So, Y. Ye, and S. Zhang, "Semidefinite relaxation of quadratic optimization problems," *IEEE Signal Process. Mag.*, vol. 27, no. 3, pp. 20–34, May 2010, doi: 10.1109/msp.2010.936019.

[6] M. Bevilacqua, A. Roumy, C. Guillemot, and M.-L.-A. Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," in *Proc. Brit. Mach. Vis. Conf.*, 2012, p. 135, doi: 10.5244/C.26.135.

[7] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *Curves and Surfaces*. Berlin, Germany: Springer, 2012, pp. 711–730, doi: 10.1007/978-3-642-27413-8_47.

[8] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013, doi: 10.1109/LSP.2012.2227726.

[9] S. Naik and N. Patel, "Single image super resolution in spatial and wavelet domain," *Int. J. Multimedia Appl.*, vol. 5, no. 4, pp. 23–32, Aug. 2013, doi: 10.5121/ijma.2013.5402.

[10] I. J. Goodfellow *et al.*, "Generative adversarial networks," 2014, *arXiv:1406.2661*. [Online]. Available: http://arxiv.org/abs/1406.2661

[11] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[12] H. Li, X. Wang, W. Liu, and Y. Wang, "Dictionary learning based image enhancement for rarity detection," in *Proc. 12th Int. Conf. Signal Process. (ICSP)*, Oct. 2014, pp. 891–894, doi: 10.1109/icosp.2014.7015132.

[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[14] L. Zhaoping, *Understanding Vision*. London, U.K.: Oxford Univ. Press, May 2014, doi: 10.1093/acprof:oso/9780199564668.001.0001.

[15] F. Chollet *et al.* (2015). *Keras*. [Online]. Available: https://github.com/fchollet/keras

[16] D. Masko and P. Hensman, "The impact of imbalanced training data for convolutional neural networks," Bachelor thesis, School Comput. Sci., Commun., KTH, 2015.

[17] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for neural networks for image processing," 2015, *arXiv:1511.08861*. [Online]. Available: http://arxiv.org/abs/1511.08861

[18] S. Diamond and S. Boyd, "CVXPY: A Python-embedded modeling language for convex optimization," *J. Mach. Learn. Res.*, vol. 17, no. 83, pp. 1–5, 2016.

[19] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Computer Vision—ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 694–711, doi: 10.1007/978-3-319-46475-6_43.

[20] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," 2016, *arXiv:1603.08155*. [Online]. Available: http://arxiv.org/abs/1603.08155

[21] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," 2016, *arXiv:1609.04802*. [Online]. Available: http://arxiv.org/abs/1609.04802

[22] X. Shen, S. Diamond, Y. Gu, and S. Boyd, "Disciplined convex-concave programming," 2016, *arXiv:1604.02639*. [Online]. Available: http://arxiv.org/abs/1604.02639

[23] E. Agustsson and R. Timofte, "NTIRE 2017 challenge on single image super-resolution: Dataset and study," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 126–135.

[24] N. Kodali, J. Abernethy, J. Hays, and Z. Kira, "On convergence and stability of GANs," 2017, *arXiv:1705.07215*. [Online]. Available: http://arxiv.org/abs/1705.07215

[25] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," 2017, *arXiv:1707.02921*. [Online]. Available: http://arxiv.org/abs/1707.02921

[26] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," 2017, *arXiv:1708.02002*. [Online]. Available: http://arxiv.org/abs/1708.02002

[27] S. Nah, T. H. Kim, and K. M. Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3883–3891.

[28] M. S. M. Sajjadi, B. Scholkopf, and M. Hirsch, "EnhanceNet: Single image super-resolution through automated texture synthesis," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4501–4510, doi: 10.1109/ICCV.2017.481.

[29] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Trans. Comput. Imag.*, vol. 3, no. 1, pp. 47–57, Mar. 2017, doi: 10.1109/TCI.2016.2644865.

[30] A. Agrawal, R. Verschueren, S. Diamond, and S. Boyd, "A rewriting system for convex optimization problems," *J. Control Decision*, vol. 5, no. 1, pp. 42–60, 2018.

[31] P. Liu, H. Zhang, K. Zhang, L. Lin, and W. Zuo, "Multi-level wavelet-CNN for image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 773–782, doi: 10.1109/CVPRW.2018.00121.

[32] R. K. Pandey, N. Saha, S. Karmakar, and A. G Ramakrishnan, "MSCE: An edge preserving robust loss function for improving super-resolution algorithms," 2018, *arXiv:1809.00961*. [Online]. Available: http://arxiv.org/abs/1809.00961

[33] G. Seif and D. Androutsos, "Edge-based loss function for single image super-resolution," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 1468–1472, doi: 10.1109/ICASSP.2018.8461664.

[34] J. Yu *et al.*, "Wide activation for efficient and accurate image super-resolution," 2018, *arXiv:1808.08718*. [Online]. Available: http://arxiv.org/abs/1808.08718
[35] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.
[36] C. Liu *et al.*, "Orthogonal decomposition network for pixel-wise binary classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6057–6066, doi: 10.1109/CVPR.2019.00622.
[37] K. Nazeri, H. Thasarathan, and M. Ebrahimi, "Edge-informed single image super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 3275–3284, doi: 10.1109/iccvw.2019.00409.
[38] C. Tian, L. Fei, W. Zheng, Y. Xu, W. Zuo, and C.-W. Lin, "Deep learning on image denoising: An overview," 2019, *arXiv:1912.13171*. [Online]. Available: http://arxiv.org/abs/1912.13171
[39] X. Zhang, S. Karaman, and S.-F. Chang, "Detecting and simulating artifacts in GAN fake images," 2019, *arXiv:1907.06515*. [Online]. Available: http://arxiv.org/abs/1907.06515
[40] C. Liu, Q. Gao, and X. Wu, "Exaggerated learning for clean-and-sharp image restoration," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, pp. 673–677, doi: 10.1109/ICIP40778.2020.9191132.
[41] Y. Liu, W. Yin, Y. Chen, H. Chen, and C. Shen, "Generic perceptual loss for modelling structured output dependencies," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5424–5432.
[42] T. Yang, P. Ren, X. Xie, and L. Zhang, "GAN prior embedded network for blind face restoration in the wild," 2021, *arXiv:2105.06070*. [Online]. Available: http://arxiv.org/abs/2105.06070
[43] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. 37th Asilomar Conf. Signals, Syst. Comput.*, Nov. 2003, pp. 1398–1402, doi: 10.1109/acssc.2003.1292216.

**Seyed Mehdi Ayyoubzadeh** received the B.Sc. degree in electrical engineering from the Sharif University of Technology, Tehran, Iran, in 2018. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, McMaster University, Hamilton, ON, Canada. His main research interests include deep learning, computer vision, and optimization.

**Xiaolin Wu** (Fellow, IEEE) received the B.Sc. degree in computer science from Wuhan University, China, in 1982, and the Ph.D. degree in computer science from the University of Calgary, Calgary, AB, Canada, in 1988. In 1988, he started his academic career. Since 1988, he has been a Faculty Member with the University of Western Ontario, New York Polytechnic University (NYU-Poly), and McMaster University. He is currently a Professor with the Department of Electrical and Computer Engineering, McMaster University. His research interests include image processing, data compression, digital multimedia, low-level vision, and network-aware visual communication. He has published over 350 research papers and holds four patents in these fields. He holds the NSERC Senior Industrial Research Chair. He serves as an Associated Editor for IEEE TRANSACTIONS ON IMAGE PROCESSING.