Robust deep learning-based multi-image super-resolution using inpainting

Henry Yau* **and Xian Du**
University of Massachusetts, Institute for Applied Life Sciences, Department of Mechanical and Industrial Engineering, Amherst, Massachusetts, United States

Abstract. Traditional super-resolution techniques are generally presented as optimization problems with variations in the choice of optimization methods and cost functions. Even for the overdetermined cases, the problem is ill-conditioned. The situation is worsened when considering underdetermined cases with unknown regions due to occlusions or lack of data. Deep learning-based methods have shown promise in solving a similar problem. One recent advancement has come in the form of partial convolutions, which were developed to perform infilling of holes in images. When used in an appropriate deep neural network, this particular variant of the convolutional filter has shown great promise in approximating missing spatial information. The method described is formulated as a two-stage process. Lower resolution images are first registered and placed on a high-resolution grid. The problem is then treated as an in-painting task where the missing regions are reconstructed using a deep neural network with partial convolutional filters. We compare our method against deep learning-based single image super-resolution methods and classical multi-image super-resolution techniques using two similarity metrics and show that our method is more robust to occlusions and errors in registration while also producing higher quality outputs. © *2021 SPIE and IS&T* [DOI: 10.1117/1.JEI.30.1.013005]

Keywords: super-resolution; deep learning; inpainting; remote sensing.

Paper 200276 received Apr. 16, 2020; accepted for publication Dec. 9, 2020; published online Feb. 1, 2021.

1 Introduction

For any imaging system, there exist constraints that limit the spatial resolution of the device. For example, restrictions may be placed on the telecommunications bandwidth or the physical dimensions of components of the imaging device, such as lens size and device mass. A more obvious constraint may be cost, limiting the size, and/or resolution of the sensor. Even without constraints on the imaging device itself, the process of capturing a scene with an imaging device is typically idealized with the assumption that targets can be resolved to the maximum resolution of the imaging sensor. However, due to the distortions and noise from atmospheric turbulence,[1] motion blur, camera blur, and other causes, the theoretical maximum resolving power of an imaging device is never reached using a single image.

It is of course desirable to have higher spatial resolution images to perform analysis with. By combining multiple low-resolution (LR) noisy images, the high-resolution (HR) image can be approximated. This process, called super-resolution (SR) restoration or reconstruction, was first described in the seminal work by Ref. 2 in the frequency domain. Multi-image SR (MISR) has been a well-researched topic since that time and has great use in fields where obtaining multiple LR images may be easier than a single HR image, such as remote sensing, medical imaging, microscopy, and from video sources such as in computer vision. Reference 2 describes the aliasing relationship that exists between the discrete Fourier transform (DFT) of the LR images and the continuous Fourier transform of the HR image. References 3 and 4 consider when the LR images are blurred and use a weighted least squares method and Tikhonov regularization.

More recently, the spatial-domain variants, such as Refs. 5 and 6, have been preferred almost exclusively as they are applicable to a wider range of observational models. For a thorough

*Address all correspondence to Henry Yau, hy2215@columbia.edu

1017-9909/2021/$28.00 © 2021 SPIE and IS&T

Downloaded From: https://www.spiedigitallibrary.org/journals/Journal-of-Electronic-Imaging on 02 Jul 2021
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

review of the problem and proposed methods, see Refs. 7 and 8. In general, the spatial SR techniques can be broadly categorized into two types. The first are methods which form a sparse linear system relating the LR images to an HR representation. The system represents an observational model that can be solved using methods such as constrained least squares, maximum *a posteriori*,[9,10] maximum likelihood, projection onto convex sets,[11] and iterative backprojection.[12] Regularizers such as total variance (TV)[6] are commonly implemented as the problem is generally ill-posed. We will refer to this method class as classical super-resolution (CSR) for the remainder of the text.

The second type is example-based methods, which use machine learning. The rapid growth in the availability of high-performance GPUs led to an exponential increase in research into deep learning. Deep neural networks (DNN) such as convolutional neural networks (CNN)[13] have been used extensively in various image-based tasks, such as classification, object detection, and segmentation. For the SR problem, some researchers attempted to learn the end-to-end mapping from the LR image space to the HR image space, such as SRCNN.[14] State-of-the-art results for single images have been achieved with generative adversarial networks (GAN) such as with SRGAN[15] and later ESRGAN.[16] Reference 15 also presents SRResNet, which is a 16-block ResNet.[17] These techniques can be viewed as the image upscalers as they take a single LR image input to generate an HR output. For multi-image approaches, there have been several differing approaches. Reference 18 has a two-stage process: a traditional SR reconstruction method and a CNN to perform noise removal. EvoNet[19] also uses a two-step process. The input images are first upscaled using a ResNet network, then the upscaled images are registered and combined into an intermediate HR image. A separate CNN is then applied to denoise the final image. The MAGiGAN[20] system is a very involved algorithm that uses several specialized processes that achieve state-of-the-art performance in remote sensing images. The process can be coarsely summarized as preprocessing, upscaling to an HR grid, refining the HR image with an image degradation model, and finally refine the HR image with a GAN network.

The work here differs from other approaches as it is designed to handle underdetermined problems with multiple LR images. For example, in Earth observation (EO) images, imaging satellites make multiple passes over the same patch. However, due to dust or clouds and the shadows they cast, parts of the patch are occluded. This makes the already ill-posed problem even more so. Unlike the single image-based methods, the multiple passes provide more information to reconstruct the SR representation. Using a data-based approach to infilling, more plausible image data can be generated than the interpolation-based CSR methods. The procedure can be divided into two stages. First, the LR images are registered to a HR grid to form a partial SR reconstruction. This intermediate SR image can be at a resolution higher than the intended output and is likely very sparse. Then, a deep learning model with partial convolutional layers (PConv) developed in Ref. 21 is used to solve for the regions not present in the LR images and remove noise. Reference 21 proposes an upscaling method where the HR images are generated using a single LR image whose pixels are placed on a subpixel corner of an HR grid. This work extends that idea using multiple LR images.

2 Proposed Method

2.1 *Image Observation Model*

Although an image observation model is not explicitly defined in our method, a brief overview is useful to understand what the DNN is attempting to replicate. An observation model describes the relationship between a continuous natural scene that is captured as bandlimited signals, represented as LR raster images. It is not possible to recover the continuous signal, the aim is to recover a high-resolution representation at a sampling rate higher than the Nyquist rate of the imaging device. We are primarily interested in the relationship between the set of LR images and the ideal HR image. A more detailed model may include several properties that negatively affect the quality of the sampled image, such as noise and blur. Typical features of the standard observation model will be detailed here, outlining the path from a continuous image to the LR representations. To relate all the LR images to the ideal HR image, the LR images must

Downloaded From: https://www.spiedigitallibrary.org/journals/Journal-of-Electronic-Imaging on 02 Jul 2021
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

be transformed to the same space because the camera or target is in different positions for each LR frame. This is communicated via a warping matrix F, which we assume to contain only translations but in general can represent rigid body motions and skew.

Different types of blur can be accounted for with a blur matrix H. The imaging sensor (e.g., CMOS or CCD) discretizes the scene by capturing the integral of the scene at each sensing element. From the viewpoint of transforming the idealized HR image to a lower resolution, this can be envisioned with a discretization matrix D, which samples a subset of pixels in the HR image to a single LR pixel. Finally, an additive noise term V_k is included to represent the image sensor noise.

The notation used here follows Refs. 5 and 6, where the images are represented as the column vectors written in a lexicographical order. The k'th observed LR images are given as $\bar{Y}_k$ for $1 \leq k \leq N$, where N is the total number of LR observations and $\bar{Y}_k$ is represented as a column vector. The desired HR image $\bar{X}$ is also written as a column vector. The observation model to solve for can then be written as follows:

$$\bar{Y}_k = D_k H_k F_k \bar{X} + \bar{V}_k \quad \text{for } 1 \leq k \leq N, \tag{1}$$

where $\bar{V}_k$ is the nonhomogeneous additive Gaussian noise, F_k is the warping matrix with dimensions, H_k is the blur matrix, and D_k is the decimation or discretization matrix. In most circumstances, Eq. (1) is an ill-posed problem.

In the set of LR images, there may be pixels that are not usable and should not contribute to the SR reconstruction. This can exist in the form of shadows, occluders, such as dust and clouds, or corrupted pixels. To account for this, a binary mask M_k of the appropriate size is included in the observation model for each LR image. For each pixel in the LR image, the corresponding value in M_k is either one or zero depending on whether the pixel should be used or ignored, respectively:

$$\bar{Y}_k^{\text{obs}} = M_k \circ (D_k H_k F_k \bar{X} + \bar{V}_k) \quad \text{for } 1 \leq k \leq N, \tag{2}$$

where "$\circ$" is the Hadamard product operator given as

$$(A \circ B)_{ij} = A_{ij} B_{ij}. \tag{3}$$

Rather than solving for the HR image through the aforementioned optimization techniques, we reformulate the problem so the HR image can be approximated using a DNN. This is done by explicitly computing the inputs for the DNN using the observation model in to produce an intermediate HR image X_{int} and using the set of LR occlusion masks M_k to produce the intermediate occlusion map M_{hr}:

$$X_{\text{int}} = \underset{1 \leq k \leq N}{\text{Median}}[D_k F_k (M_k \circ Y_k)], \tag{4}$$

$$M_{\text{hr}} = \begin{cases} 1 & \text{if } \sum_{1 \leq k \leq N} D_k F_k M_k > 0 \\ 0 & \end{cases}, \tag{5}$$

where F_k and D_k can be computed using an image registration and the desired upscaling factor. By training a deep learning model to relate X_{int} and M_{hr} to the HR representation of $\bar{X}$, the model learns not only the noise and blur but also to approximate the absent information. Creating X_{int} and M_{hr} and the DNN is discussed in the following section.

2.2 *Proposed Super-Resolution Method*

Our SR method can be divided into two main procedures: image registration and reconstruction. In image registration, the subpixel shifts between LR images are computed so they can be represented in the same spatial domain. As all possible shift values are possible, the registered HR image will not necessarily align with a regular HR grid. Therefore, a nonuniform grid is typically used and the shifted LR images are interpolated onto this grid. However, the input to the

Downloaded From: https://www.spiedigitallibrary.org/journals/Journal-of-Electronic-Imaging on 02 Jul 2021
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

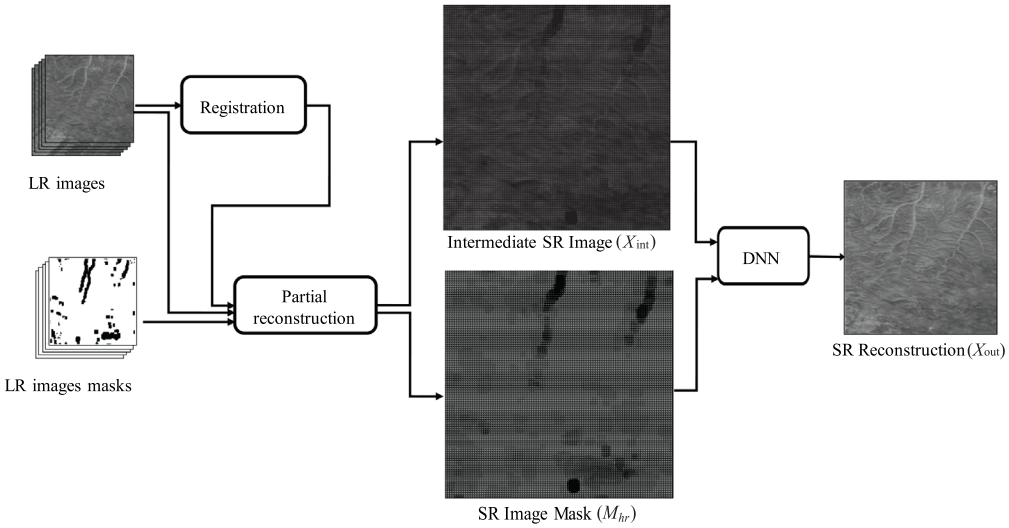

Fig. 1 Workflow of our procedure. A set of low-resolution images and their quality masks are registered and are used to form an intermediate image X_{int} and its mask M, which are used as inputs to the DNN to produce the output X_{out}.

reconstruction stage requires information to exist in discrete pixel locations, so we generate an intermediate HR image on a grid at a higher resolution than the desired output. For the reconstruction step, we propose using a deep learning network with PConv layers, which performs infilling and noise reduction on the intermediate HR image. The entire procedure is shown in Fig. 1 with the following sections detailing the steps in the proposed SR method.

2.2.1 *Image registration*

In the observational model described in Eq. (2), it is necessary to relate the LR images in the same domain by computing the warping matrix F_k. This process called image registration, transforms an image to the same spatial domain of a target image. Researchers developed several dozens methods for image registration such as pyramid iterative back-projection (PIBP),[12] SIFT,[22] and cross-correlation methods. A more general mapping that considers all affine transformations may be valuable for certain tasks, such as refining a target moving through the field of view in a video, e.g., reconstructing an SR image of moving product in a roll-to-roll process.[23]

In this work, we limit our focus to LR image transforms with only in-plane translations. With this restriction, we adopt a phase correlation method, which is robust to noise. The phase shift between two signals can be computed by finding the maximum cross-correlation between the reference signal and the signal of interest. From the convolution theorem, the Fourier transform of the cross-correlation is equal to the product of the Fourier transform of the reference signal with the complex conjugate of the Fourier transform of the target signal. The cross-correlation between images $f(x, y)$ and $g(x, y)$ is described as

$$[H]r(f, g) = \mathcal{F}^{-1}\left\{\frac{\mathcal{F}\{f\} \circ \mathcal{F}\{g\}^*}{|\mathcal{F}\{f\} \circ \mathcal{F}\{g\}^*|}\right\}, \tag{6}$$

where $\mathcal{F}$ is the Fourier transform and A^* is the complex conjugate of the matrix A. The peak of $r(f, g)$ then gives the location of the phase difference between the two images:

$$(\Delta x, \Delta y) = \underset{(x,y)}{\arg\max}\{r(f, g)\}. \tag{7}$$

To compute the Fourier transform, one could use the fast Fourier transform (FFT), however, to obtain subpixel registration, one would need to upscale the result of the product $\mathcal{F}\{f\} \circ \mathcal{F}\{g\}^*$ using zero padding. This can become computationally wasteful. For example,

Downloaded From: https://www.spiedigitallibrary.org/journals/Journal-of-Electronic-Imaging on 02 Jul 2021
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

to achieve subpixel accuracy within 1/10th of a pixel, an FFT representation 10 times larger than the original is required. Therefore, a matrix product must be computed with matrices 10 times larger than the initial image. The 10 times increase in image size leads to a 1000 times increase in computation time and a 200 times increase in memory use.

Other more efficient techniques have already been developed in Ref. 24 and others. In Ref. 24, rather than upscaling the representation, three methods are developed that exploit a feature of the frequency domain allowing for the same accuracy at a fraction of the computational and memory cost. In these methods, a DFT matrix of an equivalent upscaling factor is constructed to locally shift about an initial estimate to determine the peak cross-correlation.

An initial estimate of the phase difference is first computed using the standard FFT methods with a two-times upscaling factor, which is performed using zero padding. The true peak is therefore within a 1.5×1.5 pixel region. Using two-dimensional DFT matrices, the Fourier transform within that neighborhood can be computed without the zero padding upscaling. For two images whose dimensions are (N, N), and with an integer κ upsampling factor, the cross-correlation within the 1.5×1.5 pixel neighborhood can be computed through matrix multiplication of three matrices. The product, $D_r(\mathcal{F}\{f\} \circ \mathcal{F}\{g\}^*)D_c$, where D_r and D_c of DFT matrices of dimensions $(1.5\kappa, N)$ and $(N, 1.5\kappa)$, respectively. The computed subpixel translations can then be used to generate the warping matrix F_k in Eq. (2).

Instead of applying a transformation that interpolates LR data, a larger HR grid is created and filled with the median value of the LR data for each subpixel. The median is chosen here as it rejects extrema while directly using actual observed data. This reasoning will also be used in choosing the norm for the loss function. Using a larger grid size, potentially more LR samples are used as inputs to the deep learning network. The examples in the results section use an LR image size of (128,128) and a partial reconstruction is formed by placing the LR data into a uniform HR grid six times larger than the LR data (768,768) though the final output image is a 3× reconstructions of dimension (384,384).

In Fig. 2, the left image shows a 3×3 pixel path from an LR image whose data is inserted into a uniform HR grid six times the density (18×18) of the LR grid on the right. Each numbered square on the left represents a pixel with an area of 300 m × 300 m. Each pixel on the left is represented by 6×6 subpixels on the right with each subpixel representing an area of 50 m × 50 m. If the LR pixel has a registration shift of between 0 and 1/6 pixel in both directions, its value would be copied to the bottom left subpixel of the corresponding numbered square in the HR grid. The range for each subpixel is needed as we wish to avoid interpolation. In Fig. 2, the pixels of the LR image that contain data (1,4,7) are shifted according to the registration and placed in the upper left corner of the corresponding HR grid. The white squares represent regions where no data are available. These portions of the image will be reconstructed with the deep learning network. The right figure illustrates at least seven LR images used to construct the intermediate HR image.

2.2.2 *Reconstruction model*

If feed forward mapping from HR to LR of the observational model is interpreted as a low pass filter, solving the inverse problem can be interpreted as a high pass filter with the result that noise

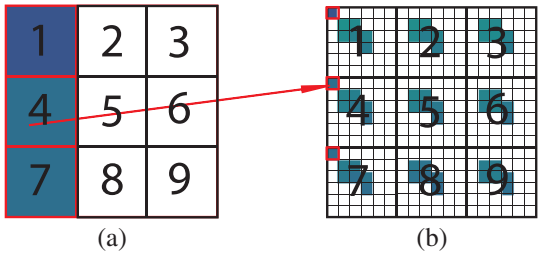

Fig. 2 (a) Low-resolution 3×3 pixel grid. (b) Intermediate 6× high-resolution image after subpixel registration. Pixels of the left image are shifted according to registration and placed in the corresponding subpixels, indicated with red/bold outline.

Downloaded From: https://www.spiedigitallibrary.org/journals/Journal-of-Electronic-Imaging on 02 Jul 2021
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

becomes amplified. This is typically done through optimization and then the application of a noise filter but is not suitable for the underdetermined case. A DNN solution provides a better alternative by incorporating information from a number of training examples. The work presented here utilizes the PConv layer developed in Ref. 21. The PConv layer can be seen as an extension of the convolutional layer that uses a mask to limit regions of the output. Like a traditional convolutional filter, the PConv is a set of weights $\mathbf{W}$ and biases $\mathbf{b}$ applied to a sliding window. For the current patch of pixels X, there is also a corresponding binary mask M that indicates the empty regions. A zero output is produced when the mask contains only zero values, otherwise a weighted convolution is performed. The output of a PConv kernel is

$$[H]x = \begin{cases} \left\langle \mathbf{W}, (X \circ M)\frac{\text{sum}(U)}{\text{sum}(M)} + \mathbf{b} \right\rangle_F & \text{if } \text{sum}(M) > 0 \\ 0 \end{cases}, \tag{8}$$

where U is a matrix of all ones with the dimensions of M and $\langle , \rangle_F$ is the Frobenius inner product. The mask is then updated such that if the mask contains at only zero values (in the sliding window), the corresponding mask output is zero, otherwise it is set to one. This simple update to the convolutional layer creates a powerful tool to synthesize data. Each subsequent application of PConv layers reduces the masked area until the entire voided area is filled.

As described in a section of Ref. 21, we use the U-Net architecture as the general framework for the DNN model. The U-Net structure was introduced by Ref. 25 for the end-to-end biomedical image segmentation. The network consists of a contraction path, which encodes features in feature maps, and an expansive path, which reconstructs an output image. The contraction path is essentially a traditional CNN with convolutional layers with rectilinear linear unit (ReLU) activation functions and max pooling layers. Rather than traditional convolutional layers, the PConv layer is used. Moving along the contraction path, the pooling layers halve the spatial dimensions of the outputs while the number of the PConv layers increases. The expansive path consists of the so-called "up-convolutions", which upsample the input and apply convolutional filters. The innovation of U-Net is the skip connections that carry information from the contraction path to the expansive path.

This can be implemented by concatenating the outputs at each depth of the contraction path with the corresponding upconvolution outputs of the expansive path. A PConv layer with a leaky ReLU activation function is applied to generate the output for each layer of the expansive path. The DNN model follows the one described in Ref. 21 with slight modifications. First, the output dimensions are smaller than the input dimensions. To address this, one could decide not to use upscaling on the final expansion layer and concatenate a downsampled first contraction layer output. One may also keep the original structure and add an additional convolutional layer, which acts as a downsampler. From our experience, this produced noticably noisier images with longer training times. We therefore chose to simply reduce the number of expansion layers. An additional change was to use a $1 \times 1 \times 1$ convolutional layer with a sigmoid activation function appended at the output to produce a single channel grayscale image. The kernel dimension and activation functions are shown in Table 1. An input image is fed into the network starting at PConv1, and the SR image is produced as the output of the final convolutional layer labeled Conv2D.

Other network structures can be used by replacing any convolutional filter with the Pconv filter, including the specialized SR networks described in Sec. 1. However, swapping the filters will double the memory requirements as each filter also includes an equivalent sized mask.

2.2.3 *Cost functions and regularization*

Given a set of LR input images, after registration the intermediate SR input X_{int} and its associate mask M_{hr} are fed into the DNN to produce the output image X_{out}. An appropriate cost function must be prescribed to recover the HR image. In the classical SR literature, the most common loss function is one based on minimizing the L_2 norm between the set of low-resolution images Y_k and $M_k \circ (D_k H_k F_k \bar{X} + \bar{V}_k)$ from Eq. (2).[26]

Downloaded From: https://www.spiedigitallibrary.org/journals/Journal-of-Electronic-Imaging on 02 Jul 2021
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

Table 1 PConv Unet model. PConv1 indicates the first PConv filter followed by a max pooling layer. PConv1c is the corresponding module consisting of upsampling, concatenation with the output of PConv1 followed by a PConv filter.

	Layer type	Kernel dimension	Activation
Expansive path	PConv1	$7 \times 7 \times 64$	ReLU
	PConv2	$5 \times 5 \times 128$	ReLU
	PConv3	$3 \times 3 \times 256$	ReLU
	PConv4	$3 \times 3 \times 512$	ReLU
	PConv5	$3 \times 3 \times 512$	ReLU
	PConv6	$3 \times 3 \times 512$	ReLU
	PConv7	$3 \times 3 \times 512$	ReLU
	PConv8	$3 \times 3 \times 512$	ReLU
Contraction path	PConv7c	$3 \times 3 \times 512$	LeakyReLU
	PConv6c	$3 \times 3 \times 512$	LeakyReLU
	PConv5c	$3 \times 3 \times 512$	LeakyReLU
	PConv4c	$3 \times 3 \times 512$	LeakyReLU
	PConv3c	$3 \times 3 \times 256$	LeakyReLU
	PConv2c	$3 \times 3 \times 128$	LeakyReLU
	PConv1c	$3 \times 3 \times 64$	LeakyReLU
	Conv2D	$1 \times 1 \times 1$	Sigmoid

The authors in Ref. 6 show that an L_2 loss exacerbates the effects of outliers, which is not desirable and suggest instead using the L_1 norm for the loss. The authors also show that the L_2 loss performs better than the L_1 loss only in the uncommon situation when an image registration is very accurate and the noise is only pure additive Gaussian. Using an L_1 norm will yield the median solution whereas the L_2 norm would yield the mean solution. This reasoning was applied to the image registration step where the median pixel value is used as opposed to the mean pixel value. The L_1 loss is also used in this work. Two loss functions related to the available pixel information and the holes are described by Eqs. (9) and (10), respectively,

$$L_{\text{present}} = \frac{1}{N_{\text{GT}}} \|M \circ (X_{\text{out}} - X_{\text{GT}})\|_1, \tag{9}$$

$$L_{\text{hole}} = \frac{1}{N_{\text{GT}}} \|(1 - M) \circ (X_{\text{out}} - X_{\text{GT}})\|_1, \tag{10}$$

where X_{out} is the output of the DNN, X_{GT} is the ground truth image, and N_{GT} is the number of elements in X_{GT}. These two loss functions are intended to maximize the peak signal-to-noise ratio (PSNR). To improve the perceptual quality of another metric such as structural similarity index measure (SSIM), we may also include a perceptual term.

The SR problem is ill-posed even for the square and over-determined cases. If the problem is underdetermined, when there is not sufficient data in the LR images, then there exist an infinite number of solutions. In the square and overdetermined cases, a small noise in the input LR images will lead to large changes in the output HR image.[6] To address this, traditionally a regularization term is added to the cost function. Tikhonov regularization, also known as a ridge regression, is commonly used. A criticism of this type of regularization is that sharp edges can become smoothed thus removing detail.

Downloaded From: https://www.spiedigitallibrary.org/journals/Journal-of-Electronic-Imaging on 02 Jul 2021
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

An alternative that addresses this problem is called TV regularization, which was proposed in Ref. 27, applied to SR in Refs. 6 and 26, and infilling in Ref. 21. TV can be defined as a criterion that penalizes the total energy change in the image, which prevents oversmoothing and preserves edges. This can be implemented as an L_1 norm on magnitude of the gradient. A discretized TV regularizer can be approximated using a 1-pixel neighborhood with

$$L_{\mathrm{TV}} = \sum_i \sum_j (\|X_{\mathrm{out}}[i+1,j] - X_{\mathrm{out}}[i,j]\|_1 + \|X_{\mathrm{out}}[i,j+1] - X_{\mathrm{out}}[i,j]\|_1). \tag{11}$$

The original work on PConv for infilling in Ref. 21 uses a perceptual loss function described in Ref. 28 (called content loss). The perceptual loss function uses the feature maps of the images when projected into the feature spaces of a VGG-16[29] pretrained on ImageNet. Due to being trained for classification, the higher layers in the network have feature spaces that represent the high level content of an image while the lower layers relate more closely to per pixel representation. Our perceptual loss can then be written as a weighted sum of the L_1 norms between the feature maps of the output and ground truth image as

$$L_{\mathrm{perceptual}} = \sum_{0<p<P-1} \omega_p \|\Psi_p^{X_{\mathrm{out}}} - \Psi_p^{X_{\mathrm{GT}}}\|_1, \tag{12}$$

where P is the number of layers in the classification CNN, $\Psi_p^{X_{\mathrm{out}}}$ is the feature map of layer p of the CNN for the output image X_{out}, and w_p is the weight of the layer. The complete loss function is then a weighted sum of the individual components as

$$L_{\mathrm{tot}} = \alpha_0 L_{\mathrm{TV}} + \alpha_1 L_{\mathrm{perceptual}} + \alpha_2 L_{\mathrm{present}} + \alpha_3 L_{\mathrm{hole}}. \tag{13}$$

2.3 *Data*

The results presented in the next section use the PROBA-V image dataset from Ref. 30. The dataset includes RED and NIR spectral bands at 100 and 300 m resolutions. The LR 300-m resolution images are 128×128, and the HR 100-m resolution images are 384×384. Both sets of images are provided at 14-bits with each image accompanied by a 1-bit quality mask, which marks occlusions such as clouds and dust. These quality masks are also used to construct the cleared mask used in computing the clear PSNR (cPSNR) score. The data set consists of 1160 samples with multiple LR images for each HR image. The number of LR images varies for each HR example from 14 to 30. This dataset is split 80/10/10 for training, validation, and testing, respectively. The training set is augmented using the flips and orthogonal rotations. The validation images are used to evaluate the performance of the proposed method during the training. Our PConv U-Net model is implemented with the Pytorch[31] framework and is trained until the total loss on the validation set stops decreasing. The remaining testing images are used to evaluate the performance of all methods, which will be discussed in the following section.

2.4 *Performance Metrics*

PSNR is a metric based on the mean squared error (MSE) of a subject and target given in dB. The PSNR is important when the actual pixel value is of scientific importance such as in the test EO data. A higher score indicates better performance. The high-resolution ground truth images used may also exhibit occlusions. These are marked with a binary quality mask discussed in the previous section. The mask allows us to compute the PSNR for only clear pixels or a cPSNR. The SSIM is a metric that measures perceptual similarity. As with PSNR, a higher score indicates better performance with a maximum score of 1 indicating identical data with perfect structural similarity. The images are brightness equalized to the reference and cropped.

Downloaded From: https://www.spiedigitallibrary.org/journals/Journal-of-Electronic-Imaging on 02 Jul 2021
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

3 Results

We compare our method against other methods including a number of DNN-based techniques. Baseline performance is provided using bicubic interpolation (BC) of a single image constructed with the median clear pixels for every point on the LR image space. This is also the input used on all tested single image SR methods. Reference 15 proposed the SRResNet network, a ResNet-based method that minimizes the MSE. Those authors also introduced SRGAN, a GAN-based single image SR method. ESRGAN[16] is considered to be the current state-of-the-art single image SR method. We also compare our method to the classical SR approach motivated by the works of Refs. 5 and 6. We will refer this generically as the CSR method in the remainder of the paper.

Input for the CSR method is the same as that for our method, an image twice as large as the ground truth with partial data. This is then treated as an inpainting problem, solving the holes using least squares with a TV regularizer. Only partial results for CSR are provided as the mean cPSNR score is well below even the BC method, likely due to poor registration. We attempted using a more accurate affine registration method with the build-in MATLAB®[32] function "imregtform"; however, for this intensity-based method, the majority of test cases the registration was poor and a valid least squares solution could not be found. In the few cases where the LR images were able to be registered, the results were on par or better than our method as seen in Fig. 4.

The average cPSNR and SSIM scores for the test set are presented in Table 2. We see that our proposed method has the highest cPSNR and SSIM scores by a fair margin. Scatter plots are shown in Fig. 3 comparing our method with a other methods with a single cross representing the scores of one test image and the red line indicating equal scores. We see there are a few outliers with much higher scores using our method, but for the majority our method is only slightly better. 109 out of 117 examples (93.16%) have a better cPSNR score using our method versus the BC. 105 out of 117 examples (89.74%) have a better SSIM score than the BC method. A slightly lower rate is found when comparing our method to ESRGAN with our method

Table 2 Average performance scores of our method compared with other methods. Best scores are highlighted in bold.

	Bicubic[a]	ESRGAN[b]	SRGAN[b]	SRResNet[b]	CSR[c]	Our method
cPSNR	32.9980	33.9187	33.2359	33.9217	30.6262	**35.4724**
SSIM	0.8384	0.8572	0.8280	0.8568	0.7881	**0.8703**

[a]As implemented in MATLAB® 2018.
[b]As implemented by Github repository: https://github.com/open-mmlab/mmediting.
[c]As implemented by Ref. 33.

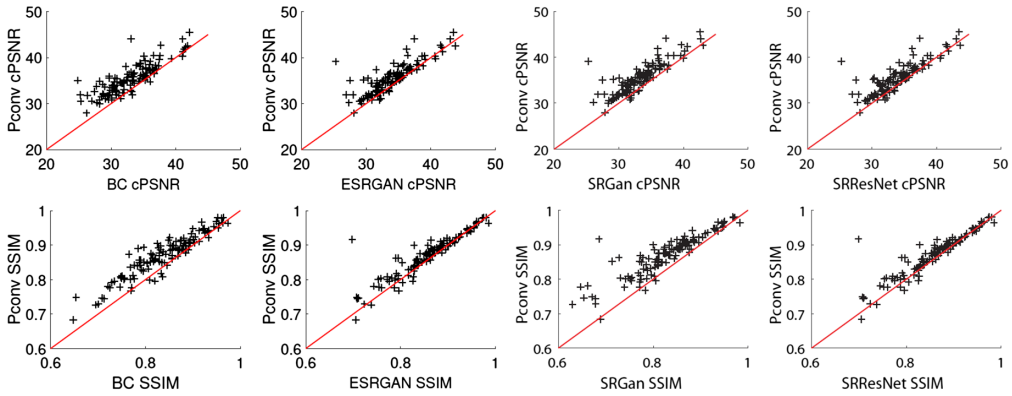

Fig. 3 Scatter plots of the performance metrics of test examples for our method (vertical axes) compared with others methods (horizontal axes). Diagonal line shows the same scores. Our method outperforms (above line) other methods in the majority of test cases, with a few significant outliers.

Downloaded From: https://www.spiedigitallibrary.org/journals/Journal-of-Electronic-Imaging on 02 Jul 2021
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

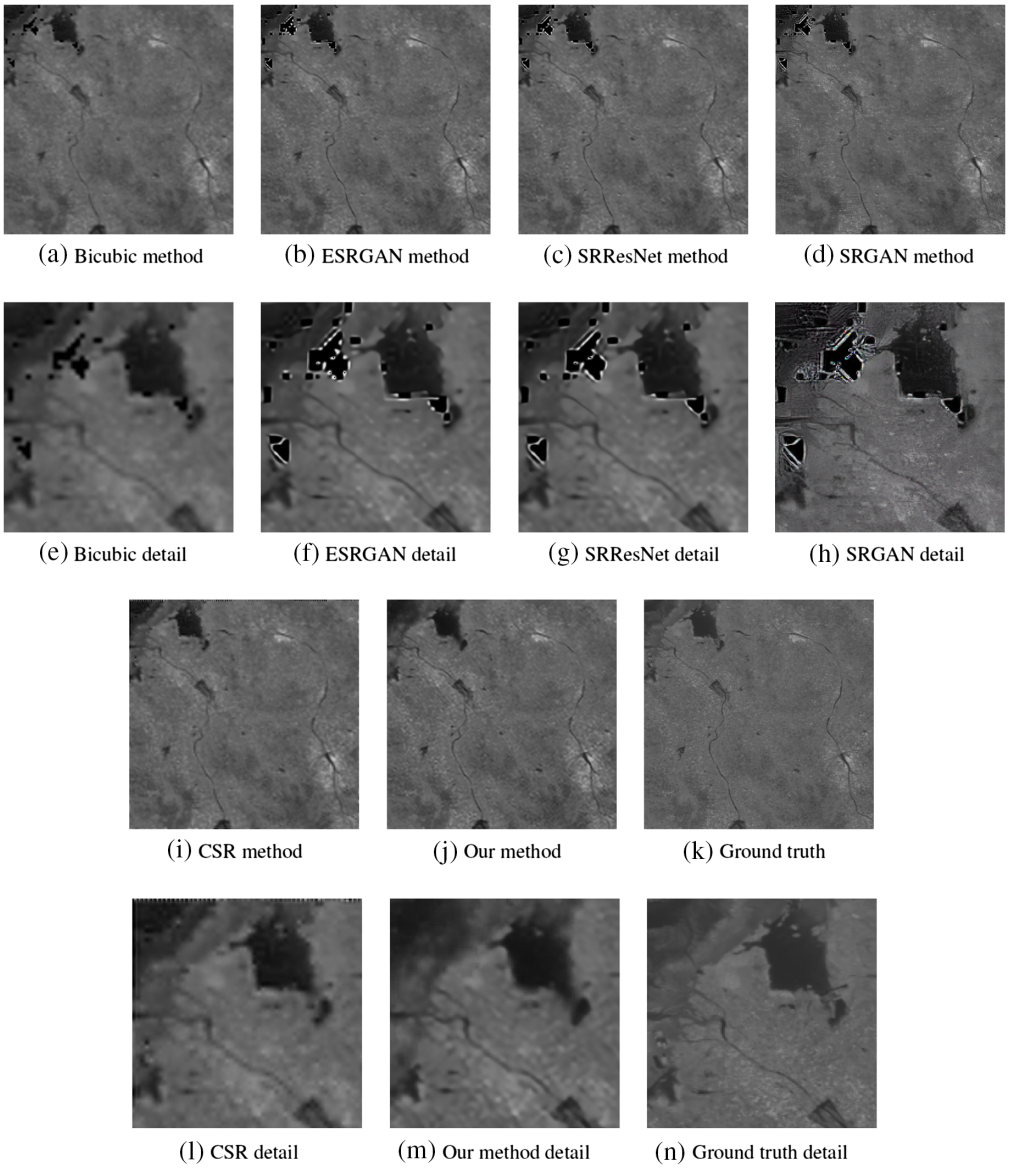

Fig. 4 Example outputs of multiple methods. Top row shows full images, bottom row shows details. With holes in the input data, other deep learning methods may extend the holes beyond the quality mask. (a) Bicubic interpolation of a single image, (b) ESRGAN result of a single image, (c) SRResNet result of a single image, (d) SRGAN result of a single image, (e) Zoom-in highlighted holes of image (a), (f) Zoom-in highlighted holes of image (b), (g) Zoom-in highlighted holes of image (c), (h) Zoom-in highlighted holes of image (d), (i) CSR result of multiple images, (j) Our method result of multiple images, (k) Ground truth of the high resolution image, (l) Zoom-in highlighted holes of image (i), (m) Zoom-in highlighted holes of image (j), and (n) Zoom-in highlighted holes of image (k). Note (l) shows that with manually selected LR images registered with affine transformations CSR perform well.

achieving better performance for cPSNR in 100 out of 117 samples (85.47%) and for SSIM in 90 out of 117 samples (76.92%). It is somewhat surprising the cPSNR scores for the single image DNN methods are much better than BC as they use the same input with no additional data given.

Figure 4 shows how holes data are treated by DNN without using the partial convolutional filters. The holes expand into areas defined by the cleared mask leading to poorer cPSNR scores. As stated earlier, the CSR method uses a more accurate registration method and the end result is better than our method.

The SRGAN method generates a more noisy image than ESRGAN or SRResNet. Figure 5 shows block artifacts when using classical SR with our partially constructed input.

Downloaded From: https://www.spiedigitallibrary.org/journals/Journal-of-Electronic-Imaging on 02 Jul 2021
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

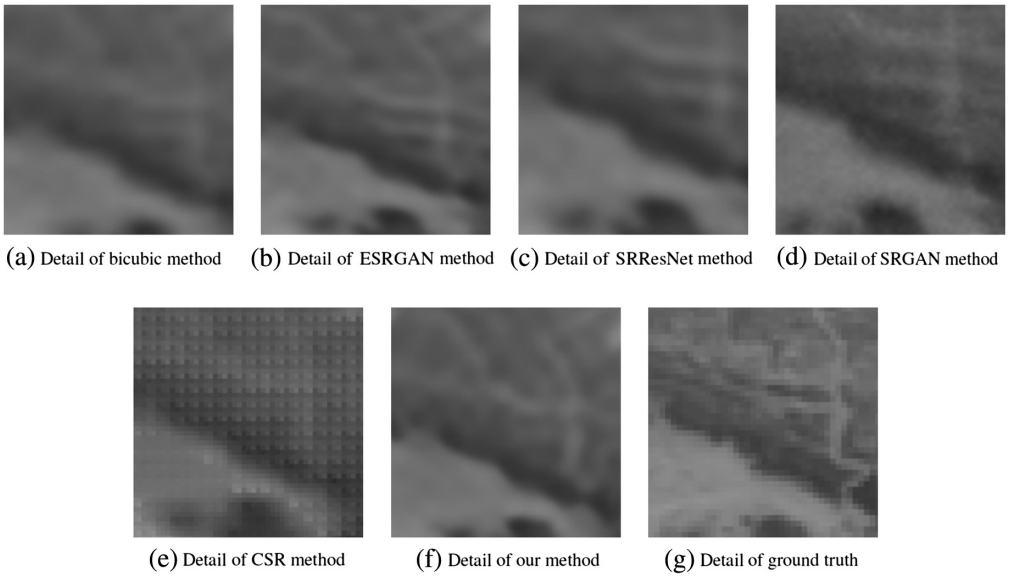

Fig. 5 Details of various methods and ground truth. (a) Bicubic interpolation of a partially constructed image, (b) ESRGAN result of the partially constructed image, (c) SRResNet result of the partially constructed image, (d) SRGAN result of the partially constructed image, (e) CSR result of the partially constructed input, (f) our method result of the partially constructed input, and (g) ground truth of the constructed input. Note the blocking pattern on (d) and noise and (c).

4 Discussion

The referenceless metrics such as PIQUE and BRISQUE scores indicate that ESRGAN produces the most natural looking images; however, though they may not be very accurate. An illustration of this is shown in Fig. 6. The figure shows a field with center pivot irrigation. This is clearly seen in the HR image; however, in the ESRGAN result, the field is transformed into a series of vertical

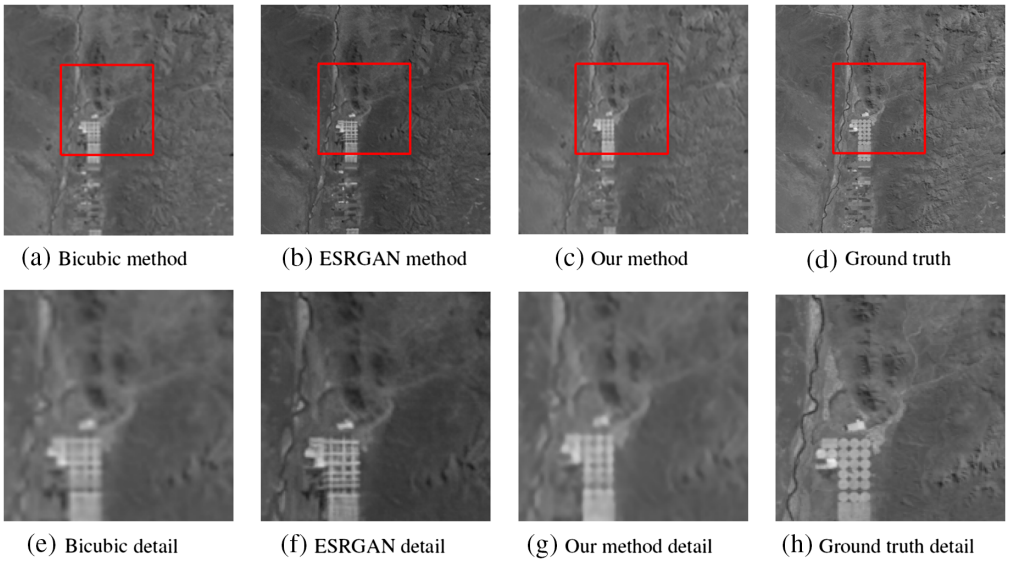

Fig. 6 SR results of a field with center pivot irrigation. Top row shows full images, bottom row shows details. (a) Bicubic interpolation of the field, (b) ESRGAN result of the field, (c) our method result of the field, (d) ground truth of the field, (e) zoom-in highlighted center pivot irrigation of image (a), (f) zoom-in highlighted center pivot irrigation of image (b), (g) Zoom-in highlighted center pivot irrigation of image (c), and (h) zoom-in highlighted center pivot irrigation of image (d). Note that GAN methods such as ESRGAN may add extraneous false detail as shown in (f).

Downloaded From: https://www.spiedigitallibrary.org/journals/Journal-of-Electronic-Imaging on 02 Jul 2021
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

and horizontal lines. Our result, while not as clear as the ground truth image, at least indicates the circular structures. This also indicates that methods that use DNN-based upscaling on LR images prior to reconstruction may in fact be introducing fallacious data into the image.

Although the same image database was used, we did not use the same images for validation as Ref. 30 so a direct comparison between cPSNR scores cannot be made. However, we may compare the improvement of the methods for the average cPSNR score over the BC method. Reference 30 has a 1.7% increase in mean cPSNR score from BC to their multi-image deep learning method. Our method has a roughly 4.9% increase in mean cPSNR score from BC.

The seemingly poor performance of the CSR methods for the test cases is due to poor registration. With LR images containing occlusions or with entire regions missing, the MATLAB® function imregtform fails for many of the test cases. Using the constructed images as the inputs for the SISR methods and performing the inpainting optimization leads to results better than the BC but below the DNN methods. Results were not shown as they did not reflect multi-image input intended for the method. By manually selecting the LR input images, results on par with our method can be obtained. However, manually determining which LR input images to use is laborious and was not performed for the entire test set.

When using a DNN structure that maintains the same input dimension to output dimension then applies a downsampling layer, we obtain outputs similar to the CSR outputs shown in Fig. 6. This is likely due to the final concatenation of the contraction path, maintaining more information from the initial image. This final concatenation was skipped for our proposed network and produced a much lower loss. This can be interpreted as the network learning the essential representation of the image and ignoring high spatial dimensional information from the input. A consequence is a much softer image, however, a more accurate image as the registration errors from the input are ignored.

In this work, we have proposed using a U-Net DNN with PConv layers to perform SR reconstruction on an intermediate image constructed using multiple LR image with quality masks. We demonstrate the method outperforms the state-of-the-art single image SR algorithms in terms of cPSNR and SSIM scores for satellite imagery and also outperforms an MISR method that uses a traditional CNN. Using the intermediate image as an input to the CSR inpainting problem produces poor results. This indicates our DNN is able to compensate for poorly registered LR images. This also suggests that an improved registration algorithm for our method would also yield better results. Affine transformations are more realistic than assuming the registration to be strictly translational. The greatest potential gain in performance will likely come an improved registration method. The method described herein may be very robust; however, using a human-curated set of LR images and classical methods is likely better. Therefore, creating a method for choosing whether or not an LR image is viable to be registered or better yet creating a robust and accurate affine or potentially homographic registration method would likely yield impressive results.

The image quality of our method appears slightly soft when compared with the ground truth and ESRGAN. This is reflected in the poorer PIQUE and BRISQUE scores. Incorporating a style loss likely produce more natural looking images; however, this would come at the cost of PSNR score. Future work in this area includes examining performance on other domains such as machine vision and medical images. Tailoring a DNN specifically for the SR problem would likely be beneficial as the missing information is much less than the traditional in-painting task, so fewer PConv layers are needed and more convolutional layers can be used for a noise reduction. The field of a deep learning-based image registration is still in its infancy and much more work can be done. It is also worth investigating other network architectures in addition to U-Net. A single DNN end-to-end solution is possible if image registration can be done using part of the DNN. This would address the major flaw in the current system of having poor registration.

Acknowledgments

This research was funded by the National Science Foundation (NSF) GOALI Grant No. 1916866. The authors have no conflicts of interest to declare.

Downloaded From: https://www.spiedigitallibrary.org/journals/Journal-of-Electronic-Imaging on 02 Jul 2021
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

References

1. A. H. Al-Hamadani et al., "Effects of atmospheric turbulence on the imaging performance of optical system," *AIP Conf. Proc.* **1968**(1), 030071 (2018).
2. T. S. Huang, Ed., "Superresolution images reconstructed from aliased images," Chapter 7 in *Adv. Comput. Vision and Image Process.* pp. 317–339, JAI Press (1984).
3. S. P. Kim, N. K. Bose, and H. M. Valenzuela, "Recursive reconstruction of high resolution image from noisy undersampled multiframes," *IEEE Trans. Acoust. Speech Signal Process.* **38**, 1013–1027 (1990).
4. S. P. Kim and W. Y. Su, "Recursive high-resolution reconstruction of blurred multiframe images," *IEEE Trans. Image Process.* **2**, 534–539 (1993).
5. M. Elad and A. Feuer, "Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images," *IEEE Trans. Image Process.* **6**, 1646–1658 (1997).
6. S. Farsiu et al., "Fast and robust multiframe super resolution," *IEEE Trans. Image Process.* **13**(10), 1327–1344 (2004).
7. S. C. Park, M. K. Park, and M. G. Kang, "Super-resolution image reconstruction: a technical overview," *IEEE Signal Process Mag.* **20**, 21–36 (2003).
8. J. Yang and T. Huang, *Image Super-Resolution: Historical Overview and Future Challenges*, pp. 1–33, CRC Press (2017).
9. R. R. Schultz and R. L. Stevenson, "Extraction of high-resolution frames from video sequences," *IEEE Trans. Image Process.* **5**, 996–1011 (1996).
10. R. C. Hardie, K. J. Barnard, and E. E. Armstrong, "Joint map registration and high-resolution image estimation using a sequence of undersampled images," *IEEE Trans. Image Process.* **6**, 1621–1633 (1997).
11. H. Stark and P. Oskoui, "High-resolution image recovery from image-plane arrays, using convex projections," *J. Opt. Soc. Am. A* **6**, 1715–1726 (1989).
12. M. Irani and S. Peleg, "Improving resolution by image registration," *CVGIP Graphical Models Image Process.* **53**, 231–239 (1991).
13. Y. LeCun and Y. Bengio, "Convolutional networks for images, speech, and time series," in *The Handbook of Brain Theory and Neural Networks*, M. A. Arbib, Ed., pp. 255–258, MIT Press, Cambridge, Massachusetts (1998).
14. C. Dong et al., "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.* **38**(2), 295–307 (2016).
15. C. Ledig et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *IEEE Conf. Comput. Vision and Pattern Recognit.* (2017).
16. X. Wang et al., "ESRGAN: enhanced super-resolution generative adversarial networks," *Lect. Notes Comput. Sci.* **11133**, 63–79 (2019).
17. K. He et al., "Deep residual learning for image recognition," in *IEEE Conf. Comput. Vision and Pattern Recognit.* (2016).
18. J. Wu et al., "Multiple-image super resolution using both reconstruction optimization and deep neural network," in *IEEE Global Conf. Signal and Inf. Process.*, pp. 1175–1179 (2017).
19. M. Kawulok et al., "Deep learning for multiple-image super-resolution," *IEEE Geosci. Remote Sens. Lett.* **17**(6), 1062–1066 (2020).
20. Y. Tao and J.-P. Muller, "Super-resolution restoration of MISR images using THE UCL MAGIGAN system," *Remote Sens.* **11**, 52 (2019).
21. G. Liu et al., "Image inpainting for irregular holes using partial convolutions," *Lect. Notes Comput. Sci.* **11215**, 89–105 (2018).
22. Mahesh and M. V. Subramanyam, "Automatic feature based image registration using sift algorithm," in *Third Int. Conf. Comput., Commun. and Networking Technol.*, pp. 1–5 (2012).
23. X. Du, H. David, and A. Brian, "Real time imaging of invisible micron-scale monolayer patterns on a moving web using condensation figures," *IEEE Trans. Ind. Electron.* **67**(5), 4077–4087 (2020).
24. M. Guizar-Sicairos, S. T. Thurman, and J. R. Fienup, "Efficient subpixel image registration algorithms," *Opt. Lett.* **33**, 156–158 (2008).

Downloaded From: https://www.spiedigitallibrary.org/journals/Journal-of-Electronic-Imaging on 02 Jul 2021
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

25. O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," *Lect. Notes Comput. Sci.* **9351**, 234–241 (2015).
26. M. K. Ng et al., "A total variation regularization based super-resolution reconstruction algorithm for digital video," *EURASIP J. Adv. Signal Process.* **2007**, 074585 (2007).
27. L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D* **60**(1), 259–268 (1992).
28. L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," arXiv:1508.06576 (2015).
29. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv:1409.1556 (2014).
30. M. Märtens et al., "Super-resolution of Proba-V images using convolutional neural networks," *Astrodyn* **3**, 387–402 (2019).
31. A. Paszke et al., "Automatic differentiation in pytorch," in *Proc. 31st Conf. Neural Inf. Process. Syst. (NIPS)*, *Long Beach, California* (2017).
32. MATLAB, version 9.4.0 (R2018a), The MathWorks Inc., Natick, Massachusetts (2018).
33. O. Hanson, "Robust and fast super resolution," https://faculty.idc.ac.il/toky/old_courses/videoProc-07/projects/SuperRes/srproject.html (2007).

Henry Yau was a post-doctoral researcher in the Mechanical and Industrial Engineering Department at the University of Massachusetts Amherst. He received his BS degree in applied mathematics from the University of Colorado Boulder, his MS degree in mechanical engineering from the University of Colorado Denver, and his doctorate in mechanical engineering from Columbia University in the City of New York. His research interests include intelligent control systems, reinforcement learning, GNC, and applications of AI deep learning to these problems.

Xian Du is an assistant professor in the Mechanical and Industrial Engineering Department at the University of Massachusetts Amherst. He received his PhD in the program of Innovation of Manufacturing Systems and Technology from the Singapore-MIT Alliance. He is a recipient of the 2020 NSF Career award. His current research interests include high-resolution, large-area and fast-speed sensing, machine vision and pattern recognition technologies for roll-to-roll flexible electronics printing process and personalized health monitoring devices.

Downloaded From: https://www.spiedigitallibrary.org/journals/Journal-of-Electronic-Imaging on 02 Jul 2021
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use