Analyzing Perception-Distortion Tradeoff in Remote Sensing Image Super Resolution

Abstract

High spatial and temporal resolution earth observation images are desirable for many remote sensing applications. Despite many advances have been made in launching medium to high spatial resolution satellites, the majority of public available remote sensing images are still within a spatial resolution of less than 10 m, such as Sentinel-2 (10-60 m), Landsat (30-m), and MODIS (250-1000m), which limits our ability to quantify spatiotemporal changes with high details. Finding out solutions to enhance the spatial resolution of these available remote sensing archives through a cost-effective approach will be a great contribution to earth observation and remote sensing communities. Although deep learning convolutional neural networks have been shown to outperform on non-remote sensing benchmark databases, only a few research has been carried out for remote sensing images. Generative Adversary networks (GANs) recently show superior performance when restoring the texture details. Work focused on achieving better perceptual performance is still missing on remote sensing images. In this paper, we investigate the perception-distortion tradeoff by conducting experiments on two remote sensing products: Landsat and NAIP. We further discuss whether perceptual-oriented methods such as (GANs) should be paid attention to on remote sensing image SR. Our results show that perceptual-oriented methods have their superiority especially when the resolution of images is lower than 100m (e.g. the x4 image of Landsat-30m has resolution 120m). However, perceptual-oriented methods are very sensitive to luminance changes such as shadow, a common feature in relatively high-resolution remote sensing images, and leads to unpleasant artifacts. Also, the use of perception loss sometimes eliminates small items (e.g., cars) that have similar color with their background.

1. Related Work

1.1. Remote Sensing Image Super-Resolution

In the field of remote sensing image analysis, the conflict between the limited access to a satisfying open-source dataset and the requirement for higher resolution earth observation (EO) images arises the research about remote sensing image super-resolution (SR).

The most commonly-used optical dataset in the area is Landsat [25], which is the fundamental data source for understanding historical change and its effect on the environmental process. Although the dataset provides the longest record of moderate spatial resolution (30m) data of the earth from 1984 to present [26], it is far from enough to fulfill all kinds of requirement of high-resolution observations.

The most widely accepted and developed remote sensing image super-resolution technology is data fusion. This method depends on the redundant information among multiple images and combines data from different remote sensing devices like optical remote sensing data and radar remote sensing data [13]. The first discussion about combining satellite imagery with different spatial resolutions started from 1991 [3]. After that, more papers bring about the variations of this method [29, 7, 19, 22], and it proves to improve the quantitative remote sensing applications such as land cover [29, 6, 13]. Recently, with the development of deep learning and neural networks, convolution neural networks (CNNs) have been shown to outperform these. Although CNNs have achieved attractive results on non-remote sensing benchmark databases, much less research has been carried out for remote sensing images[26].

1.2. Neural Network Solution

As Dong et al. [5] proposed SRCNN, a simple three layer CNN, to learning the mapping from Low Resolution (LR) to High Resolution (HR) image end-to-end, achieving superior performance against previous work, a variety of network architectures have been introduced into SR problem [14, 17, 32].

Deep networks SR can be roughly divided into two classes based on their main purpose. One class aims at optimizing distortion measures such as Peak Signal-to-Noise Ratio (PSNR), Mean-Square Error (MSE), etc. to restore more accurate images. Networks based on pixel-wise loss functions are within this category. Another group of studies attempt to produce more photo-realistic images that are superior in terms of perceptual quality. Usually, an extra

perceptual loss is considered.

Most existing CNN based SR models are designed to achieve better PSNR performance. The trend of development is increasing the complexity of the network. In [18], the authors proposed SRResNet with 16 residual units to optimize MSE. Kim et al. [15] proposed a deeply-recursive convolutional network (DRCN). Further, densely connected network [11] combined with residual blocks employed in SR [32]. EDSR [20] proposed by Lee et al. improves the performance significantly by removing the batch normalization layers in SRResNet and increasing the number of residual blocks to 23. Recently, DBPN [8] using deep architectures to simulate iterative back-projection and further improves performance with dense connections is shown wonderful performance in 8 scale SR.

Perceptual-oriented approaches proposed in SR are relatively less than PSNR-oriented approaches. To optimize perceptual similarity, Johnson et al. [12] proposed perception loss to enhance the perceptual quality by minimizing the error in feature space instead of pixel space. SRGAN [18] is later proposed that combines perceptual loss and adversarial loss to generate realistic textures during SR. [28] is similar to SRGAN but using a local texture matching loss. [30] further improved the performance of SRGAN by removing the BN layers and expanding the network size and also employing a more effective relativistic average GAN. Generative adversarial networks (GANs) have shown the greatest advantages in photo-realistic SR image generation. However, studies for restoring high-frequency textures and producing high perceptual quality SR images with neural networks are still missing.

1.3. Perception-Distortion Tradeoff

There is no existing algorithm that can produce SR images with high accuracy and high perceptual quality simultaneously. PSNR-oriented methods optimizing the pixel-wise reconstruction measure tend to produce over-smoothed images that lack high-frequency textures, which have been shown to correlate poorly with the human perception [18, 28]. In contrast, perceptual-oriented methods are often inferior in terms of distortion measure e.g. PSNR [18, 12, 28, 4]. This leads to two different types of quality measures: full-reference measures and non-reference measures. SR algorithms are typically evaluated by full-reference measures, i.e. distortion measures such as PSNR, MSE, the Structural Similarity Index (SSIM), etc. These measures are considered not in line with human opinion score. Besides human evaluation (e.g. user study), non-reference measures can be used for perceptual quality evaluation, such as Ma's score [21], NIQE [24] and BRISQUE [23]. A common interpretation of the perception-distortion tradeoff phenomenon is a shortcoming of the existing distortion measures [31]. Recently, Blau et al. [2] mathematically proved that distortion and perceptual quality are at odds with each other and has nothing to do with the measure criteria. In this work, we analyze the tradeoff between perception-distortion in remote sensing images and discuss whether perceptual-oriented methods should be employed in remote sensing image SR problem.

2. Approach

The goal of our proposed model is to produce a high-resolution, super-resolved remote sensing image by using a deep neural network model. The model uses only low-resolution sensor image as input. This can be viewed as single-image super-resolution (SISR) problem where the input low-resolution image is the only correspondent image to the target high-resolution image; when performing the actual super-resolution task no high-resolution images are provided. In the training process, pairs of low-resolution images and high-resolution images are given. Low-resolution images can be generated by applying image blurring filters to existing high-resolution images and then performing a downsampling.

The large number of parameters in a deep neural network allows the network to characterize complex problems. Both adding network layers (producing deeper network) and adding filter numbers (producing thicker layer) can increase the network complexity. In our study we also want our neural network to have enough complexity to capture the mapping from LR to SR and adding more layers is a straightforward solution. However more parameters will increase the difficulty in training the model [10], and backpropagation can be found to fall short in training very deep networks, causing deeper networks having errors. The ResNet proposed by He et al [10] used shortcut connection. After non-linear operations, the weights can be directly added to another deeper layer. This shortcut copy the weights from the shallower layer and directly add up to the deeper weight, After the addition, the network performs a non-linear operation on the weighted sum.

2.1. Network Architecture

Using ResNet for solving SISR tasks is natural, but we want to study how ResNet should be applied in SISR models and what modifications are required. Ledig [18] et als SRResNet structure employed skip connection idea from the ResNet. Compared to the original ResNet, they removed the ReLU module after the addition operation (the layer connected with and the source by the shortcut) and they preserved the batch normalization layers.

Lim et al [20] also used the ResNet-like skip connection structure in their SISR model EDSR. EDSR further removed the batch normalization layers in the skip connections and the saved memory allows it to have more total layers than the SRResNet. Consider the fact that ResNet was

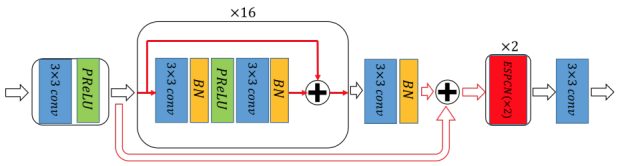


Figure 1: Architecture of the SRResNet-resembling network

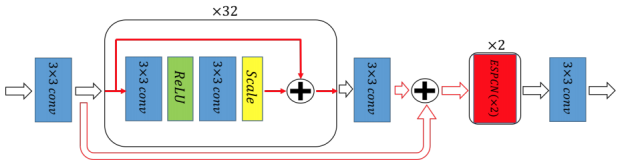


Figure 2: Architecture of the EDSR-rsembling network

designed for tasks like object detection and image classification, where inner representations are highly abstract and these representations can be insensitive to the shift brought by batch normalization. As for image-to-image tasks like SISR, since the input and output are strongly related, if the convergence of the network is of no problem, then such a shift may harm the final performance. Simplifying the ResNet by removing batch normalization can be reasonable. However, it is still not sure how different modifications of backbone ResNet will perform in a setup with certain loss functions and in the context of remote sensing images. Except for regular depth increasing, EDSR also increases the number of output features of each layer on a large scale. To release the difficulties of training such wide ResNet, the residual scaling trick is employed.

In our study, we construct different modified ResNet as training backbone for generating high-resolution remote sensing images. Figure 1 shows the detailed structure of our first network to evaluate. It resembles the SRResNet structure. This architecture follows the guideline proposed by Radford et al [27]. There are in total 16 residual blocks in the network. In each residual block, we use 3x3 convolutional kernels with 64 filter channels. A batch normalization is added after the convolution. The feature maps are activated then using ParametricReLU [9] as the activation function. The activation is followed by another convolutional layer and a batch normalization layer. This skip connection links featuremaps of the start of a residual block directly to the end, using the elementwise sum operation. Outside the residual blocks we also have a ParametricReLU activation layer. As is pointed by Radford et al., the network does not contain fully connected hidden layers.

Figure 2 shows the EDSR-resembling network we construct. In its residual block we remove the batch normalization layer. Besides we follow the choice of Lim et al to use ReLU as the activation function and remove the activation outside the residual blocks. There are in total 32 such residual blocks in the architecture.

2.2. Loss Functions and Adversarial Learning

Another important factor in training a good SISR model is the choice of the loss function. Both SRResNet and EDSR use pixel-wise distance as the loss function for model optimization. The difference is that SRResNet uses mean squared error (MSE) which in other words is L2 loss in image pixel space while EDSR uses the L1 loss in pixel space. In this study, we want to use Generative Adversarial Networks (GANs) so that we can optimize the model based on the adversarial loss given by the discriminator. By using the GAN architecture, the image produced by the model will have a more natural look regarding the existing images. The GAN architecture includes a generative model and a discriminative model. Both of these models will be based on deep convolutional neural networks. The generative model can be originated from existing SISR model. The discriminator network will be trained to solve the maximization problem in the equation below:

$$\min_G \max_D E_{x\sim P_{data}} logD(x) + E_{x\sim P_{model}} log(1 - D(x))$$

Ledig et al proposed SRGAN based [18] by using the SRResNet as the generative network. The incentive and intuition behind the SRGAN are from optimizing the forward Kullback-Leibler divergence (KLD). This average effect of the forward KLD is well-known as the regression-to-the-mean problem, which is very common when the solution is inadequate in practice. The Kullback-Leibler divergence (KLD) is defined between the conditional empirical distribution P_{data} and the conditional model distribution P_{model}. Although minimizing the backward KLD will lead to better visual results with respect to non-reference perceptual quality assessment, in most low-level computer vision tasks, P_{data} is an empirical distribution and P_{model} is an intractable distribution. For this reason, the backward KLD is unpractical for optimizing deep architectures. Generative adversarial nets (GAN) proposed by Goodfellow et al. use the objective function below to implicitly handle this problem in a game theory scenario.

However, using SRResNet as the generative network in GAN can be suboptimal in SISR task because it has little change over the initial ResNet. So we consider design the generative model with a modified ResNet, similar to EDSR, and build a GAN-like model to handle the remote sensing image super-resolution task. As is in SRGAN, the total loss l^{SR} (termed as perceptual loss in [18]) used by the generative model is the weighted summation of the content loss and the adversarial loss:

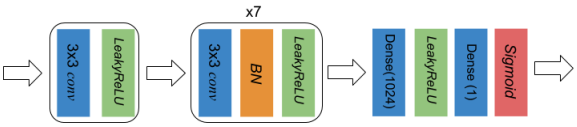


Figure 3: Architecture of the Discriminative network

$$l^{SR} = \underbrace{l_X^{SR}}_{\text{content loss}} + \underbrace{10^{-3} l_{Gen}^{SR}}_{\text{adversarial loss}}$$

The content loss for SRGAN is defined as the element-wise L2 loss in VGG feature space. The features were extracted from pretrained VGG models instead of raw pixels so that the high-level content is preserved. And the adversarial loss ensures the generated images look real.

We plan to have a performance comparison on the experiment results using GAN with both backbone structures we discussed as the generative network. We call the network with EDSR-resembling generator EDSR-GAN to be compared with SRGAN. The structure of the discriminative network in our study is shown in Figure 3. We use the same discriminator architecture for both of the generators. We follow Ledig et al's SRGAN design. The network uses LeakyReLU activation in the discriminator for all layers and contains batch normalization layers.

We empirically find that the unbalanced training data leads to bad performance when images are within city region. Since city samples are much less than other area. So we balance our loss with this equation:

$$\theta_G = \underset{\theta_G}{\operatorname{argmin}}(w_1 \frac{1}{N_1} \sum_{N_1} l^{SR} + w_2 \frac{1}{N_2} \sum_{N_2} l^{SR})$$

where N_1 is the total number of images over city area, N_2 is the number of images of other area. $w_1 = 0.8$ and $w_2 = 0.2$

Results have been slightly improved.

3. Experiments

We evaluate the performance of PSNR-oriented methods: SRResNet and EDSR, perceptual-oriented methods: SRGAN and our proposed EDSR-GAN on two different resolution (0.6m and 30m) remote sensing products, then discuss the perception-distortion tradeoff in remote sensing images SR.

3.1. Data

The United States Department of Agriculture (USDA) National Agricultural Imagery Program (NAIP) aerial imagery with 0.6 m spatial resolution (NAIP-0.6m), and Landsat-8 OLI level-2 surface reflectance product with 30 m spatial resolution (Landsat-30m) archived in Google Earth Engine[1] were directly used for multi-scale experimental tests in this study. Samples from two datasets are shown in Fig. 4

(a) Landsat-30m (b) NAIP-0.6m

Figure 4: HR Samples from Landsat-30m and NAIP-0.6m datasets

NAIP-0.6m dataset and Landsat-30m dataset are trained and tested separately. Landsat-30m in general has more high-frequency texture information over the city area compared to NAIP-0.6m due to its coarser resolution. This is an interesting feature when we investigate the perceptual and distortion tradeoff.

Downloaded Landsat-30m (NAIP-0.6m) images are cropped into 480x480 patches. 500 HR patches from Landsat-30m (NAIP-0.6m) are split into a training set (450), a validation set (30) and testing set (15). We obtain LR images by down-sampling HR images using bicubic interpolation. All experiments are performed with a scale factor of 4x between low- and high-resolution images. The size of HR image I^{HR} is $480 \times 480 \times C$ where $C = 3$ is color channels (RGB), then the size of corresponding LR image I^{LR} is $120 \times 120 \times C$. Images are all in tiff format. Landsat-30m is level-2 surface reflectance product. Its image type is float64 ranging from 0 to 10000. NAIP-0.6m is commercial product ranging from 0 to 255. Before feeding into the network, linear interpolation is applied to convert all images into float32 ([0, 1]). Also, as is mentioned in section 2.2, in our dataset, there are much more forests and farmland than cities. We notice that such unbalanced dataset has an influence on our results. Thus we add more weights to the loss of the city samples during the training.

3.2. Training Details

We fine-tune SRResNet, SRGAN and EDSR on Landsat-30m (NAIP-0.6m) training dataset. Although remote sensing images seem to be different from non-remote sensing dataset (e.g. ImageNet [27], DIV2K [1]), our experiments show that it's beneficial to initialize with weights from a pretrained model due to the limited training data and training time. The pretrained models are obtained from online

[1]https://earthengine.google.com/

materials: SRResNet and SRGAN[2], EDSR[3]. Then the fine-tuned EDSR model is employed as the initialization for the generator of EDSR-GAN.

For training, The mini-batch size is set to 16. The training images are again cropped into 128x128 HR patches as input images. We fine-tuning SRResNet, SRGAN and EDSR with a learning rate of 10^{-5} which is smaller than the original 10^{-4} [18, 20], since we're not training from scratch. When training EDSR-GAN, we initialize the learning rate as 10^{-4} and decayed by a factor of 2 every 10^5 mini-batch update. For optimization, we use Adam [16] with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$. We implemented these networks with the PyTorch framework and train them using NVIDIA Tesla K40c GPU.

3.3. Distortion and Perception Measures Evaluation

Full-reference measures we used in this work is PSNR and SSIM. For non-reference measures we use BRISQUE, Ma's score and NIQE. Ma's score is a recent metric specifically designed for SR quality assessment. All metrics are evaluated on the luminance channel in YCbCr color space. The results are shown in table 1

We notice that perceptual-oriented methods achieve high scores using non-reference metrics while their PSNR and SSIM score are not as good as PSNR-oriented methods. However, as we will discuss in qualitative results, the over-sharpening effect caused by perceptual-oriented method actually will influence its perceptual quality. while these metrics didn't reflect such an effect. Different from [2], we think a more "perceptual" distortion metric is still needed due to the nature of the SR. SR is different from fake image generation. We do care if it is in line with LR images and if we can get more useful information from SR. Accuracy should be considered. Especially in remote sensing image SR.

3.4. Qualitative Results

We qualitatively compared the result of two PSNR-oriented and two perceptual-oriented models on two different remote sensing products: Landsat-30m and NAIP-0.6m. When comparing the results of NAIP-0.6m, we use enlarged patches. While for the Landsat-30m SR image, we evaluate it in a whole. This is because the coarse resolution makes even the HR image look "noise", due to the complexity of the surface features. Especially over the city area, we are not able to see different buildings but very little tiny points or squares. Such natural "noise" makes the detailed comparison less obvious. Also, due to the size of the image we used here, Landsat can be used directly for land use detection, etc. While NAIP with the same size is too small for such task but it's high resolution make it possible for object detection, etc. Different usage decides that our emphasis are different.

Since regular image reading software is not able to show Landsat images details clearly due to its low contrast (Landsat is level-2 surface reflectance product). We use QGIS[4], an open source geographic information system to do the analysis.

3.4.1 Landsat-30m

Qualitative results of Landsat-30m over two different areas: "smooth" area such as farmland and "noise" area such as city are shown in Fig. 5 top and bottom respectively. PSNR and Ma's score are also provided for reference. We see that no matter what the surface type is, perceptual-oriented methods look much better than PSNR-oriented methods. While the performance within the same class (e.g., SRGAN vs. EDSR-GAN, and SRResNet vs. EDSR) is similar. We notice that remote sensing images differ from natural images because of the complexity of the surface features. Buildings may have high reflection depending on their material while the ground is not a good reflector but absorber. The low contrast and various surface reflection make it's hard to restore the textures precisely. Perceptual-oriented methods generate high-frequency texture by enhancing the "semantic edges" but will introduce more noise. However, compared to the "noise" nature of the HR images, SRGAN and EDSR-GAN outperform SResNet and EDSR. Although the restored textures are not precise for both of SRGAN and EDSR-GAN so that it's hard to compare which one is better, it's good enough to distinguish the grassland from cities compared with SRResNet and EDSR.

3.4.2 NAIP-0.6m

Some representative qualitative results for NAIP-0.6m are shown in Fig. 6. Different from the results of Landsat-30m, EDSR and EDSR-GAN show superior performance in both accuracy and perceptual quality compared to SRResNet and SRGAN. However, perceptual-oriented methods doesn't obviously outperform PSNR-oriented methods. Although perceptual-oriented methods still generate visually better images than PSNR-oriented method, they are more sensitive to the changes of the luminance. Sometimes, slightly changes (e.g., shadow) will be detected as texture and then be enhanced. Such over-sharpening leads to unpleasant artifacts as is shown in Fig. 6 middle and bottom images. However, the perception score

In sum, when remote sensing images has low resolution, producing high frequency texture information is more important. The noise and imprecise introduced by the method itself do not have significant impact on the usage of the data.

[2]https://github.com/mseitzer/srgan/tree/master/resources/pretrained

[3]https://github.com/thstkdgus35/EDSR-PyTorch

[4]https://www.qgis.org/en/site/

	NAIP				landsat			
method	SRResNet	EDSR	SRGAN	EDSR-GAN	SRResNet	EDSR	SRGAN	EDSR-GAN
PSNR	32.2545	**33.3648**	29.1281	28.7281	**41.9061**	41.8931	38.9993	39.2799
SSIM	0.66893	**0.89518**	0.82545	0.7884	**0.94836**	0.82378	0.90917	0.91172
Ma's	4.4087	4.5259	6.4298	**6.4523**	2.8144	2.7111	4.3224	**4.7406**
NIQE	8.8035	8.5127	**5.1982**	5.4045	5.6204	5.7767	3.6606	**3.6072**
BRISQUE	64.2224	58.8081	**29.1929**	29.3357	63.8464	68.3178	**26.8060**	28.7601

Table 1: We evaluate the four methods mentioned in this paper using five types of measures on two datasets: NAIP(the left block) and Landsat(the right block). The number in bold is the best result for each measure. For PSNR, SSIM and Ma's measurement, higher score means better quality. While for NIQE and BRISQUE, lower score reflects better perceptual quality.

When the image resolution is high, shadow becomes crucial. Perceptual-oriented methods are not robust to slight luminance change which will lead to obvious artifacts such as house cannot be detected. Experiments also reveals that training based on perceptual loss will sometimes eliminate small items which have similar color with its background. But overall, perceptual-oriented methods should be considered into remote sensing image SR.

4. Conclusion

In this work, we investigated the tradeoff between PSNR-oriented methods and perceptual-oriented methods on remote sensing images. Our experiments were conducted on two types of remote sensing products: Landsat-8 OLI level-2 surface reflectance product with 30m spatial resolution and NAIP aerial imagery with 0.6m spatial resolution.

We fine tuned SRResNet, EDSR, SRGAN and further, extended EDSR with SRGAN's adversarial loss as generator network, named EDSR-GAN. To address the unbalanced training data problem, we applied loss balance, i.e. adding more weights on the loss of city samples. Our results proof that Perceptual-oriented methods are less accuracy compared to PSNR-oriented methods but has higher perception quality. When input LR image has resolution lower than 100m (e.g., Landsat-30m), dataset perception-oriented methods outperform PSNR-oriented methods. While the performance SRGAN vs. EDSR-GAN and EDSR vs. SRResNet are similar. When input LR image resolution is relatively high (e.g. less than 5m, NAIP-0.6m dataset), EDSR-GAN and EDSR performs better than SRGAN and SRResnNet, respectively. However perceptual-oriented methods don't show superior performance over PSNR-oriented methods. This is because perceptual-oriented methods are sensitive to luminance changes, shadow which is a common feature in remote sensing images will lead to oversharpening. Besides, training with perceptual loss will sometimes cause small items fused into it's background.

We think a more "perceptual" metric is still needed and can be found. A user study can be employed to help define such metric. For example, design 100 images with different types of distortion. Results would show us that noise resides at which part of the image and what type of distortion may influence human perception to what degree. Moreover, a better perceptual loss function that can take the changes of the luminance (e.g. shadow) into consideration is needed.

References

[1] E. Agustsson and R. Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1122–1131, July 2017.

[2] Y. Blau and T. Michaeli. The perception-distortion tradeoff, 2017.

[3] P. S. Chavez, S. C. Sides, and J. A. Anderson. Comparison of three different methods to merge multiresolution and multispectral data: Tm spot pan. *Photogrammetric Engineering Remote Sensing*, 57(3):265–303, 1991.

[4] R. Dahl, M. Norouzi, and J. Shlens. Pixel recursive super resolution. *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

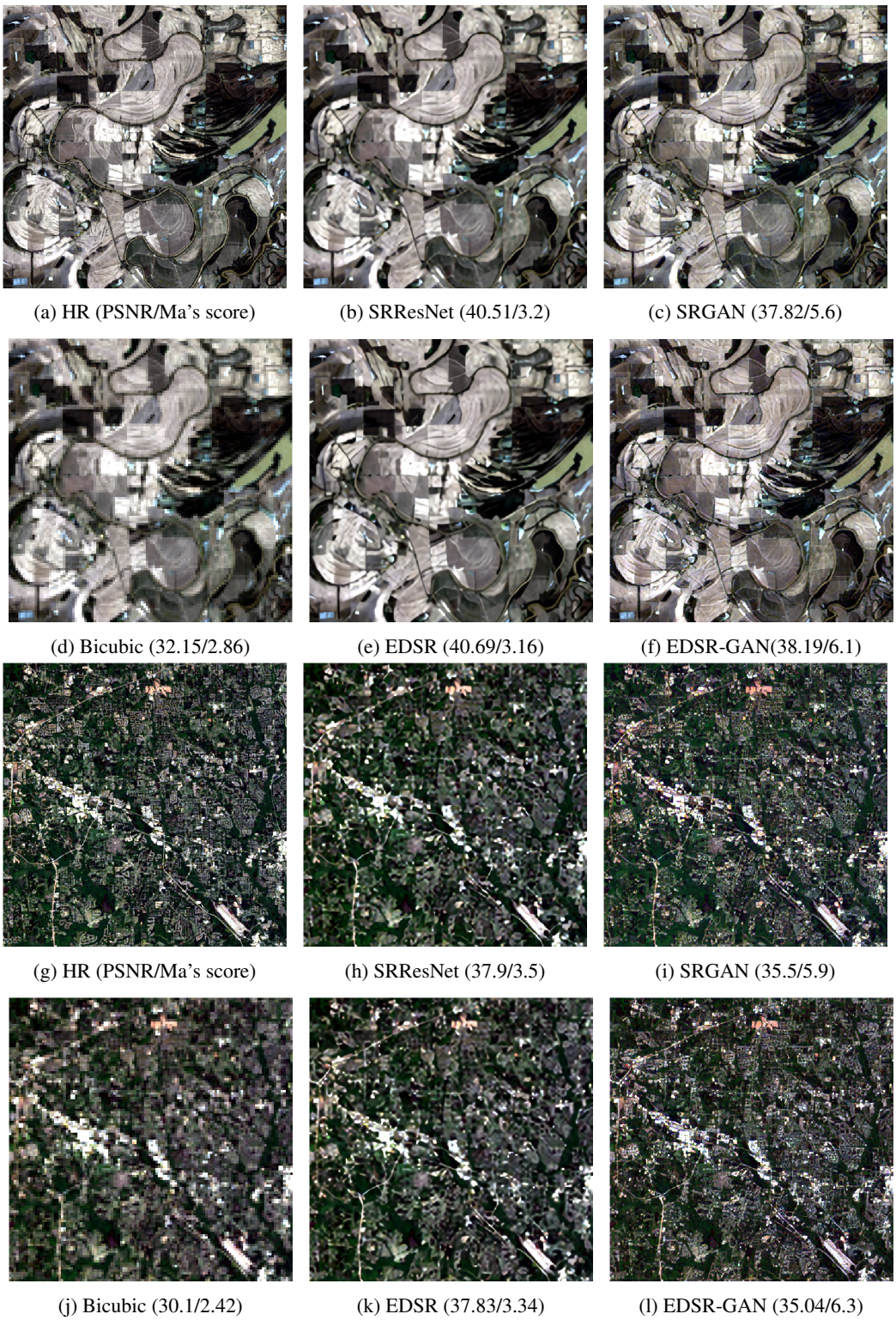
[5] C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295307, Feb 2016.

[6] J. K. Gilbertson, J. Kemp, and A. V. Niekerk. Effect of pan-sharpening multi-temporal landsat 8 imagery for crop type differentiation using different classification techniques. *Computers Electronics in Agriculture*, 134:151–159, 2017.

[7] A. Grochala and M. Kedzierski. A method of panchromatic image modification for satellite imagery data fusion. *Remote Sensing*, 9(6):639, 2017.

[8] M. Haris, G. Shakhnarovich, and N. Ukita. Deep back projection networks for super-resolution. In *Conference on Computer Vision and Pattern Recognition*, 2018.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet



(a) HR (PSNR/Ma's score) (b) SRResNet (40.51/3.2) (c) SRGAN (37.82/5.6)

(d) Bicubic (32.15/2.86) (e) EDSR (40.69/3.16) (f) EDSR-GAN(38.19/6.1)

(g) HR (PSNR/Ma's score) (h) SRResNet (37.9/3.5) (i) SRGAN (35.5/5.9)

(j) Bicubic (30.1/2.42) (k) EDSR (37.83/3.34) (l) EDSR-GAN (35.04/6.3)

Figure 5: (a) and (g) are reference HR image from Landsat-30m, (b)-(f) represents different SR model, same for (h)-(l). Quantitative measures are shown in parenthesis as (PSNR/Ma's score)

Figure 6: Qualitative results of NAIP-0.6m. The reference HR images from NAIP-0.6m are shown on the left. On the right are 3 times enlarge patches of the yellow box. Quantitative measures are shown in the parenthesis as (PSNR/Ma's score)

classification. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pages 1026–1034, Washington, DC, USA, 2015. IEEE Computer Society.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[11] G. Huang, Z. Liu, L. v. d. Maaten, and K. Q. Weinberger. Densely connected convolutional networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017.

[12] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. *Lecture Notes in Computer Science*, page 694711, 2016.

[13] N. Joshi, M. Baumann, A. Ehammer, R. Fensholt, K. Grogan, P. Hostert, M. R. Jepsen, T. Kuemmerle, P. Meyfroidt, and E. T. A. Mitchard. A review of the application of optical and radar remote sensing data fusion to land use mapping and monitoring. *Remote Sensing*, 8(1):70, 2016.

[14] J. Kim, J. K. Lee, and K. M. Lee. Accurate image super-resolution using very deep convolutional networks. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2016.

[15] J. Kim, J. K. Lee, and K. M. Lee. Deeply-recursive convolutional network for image super-resolution. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2016.

[16] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2014.

[17] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017.

[18] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, volume 2, page 4, 2017.

[19] Z. Li, H. K. Zhang, D. P. Roy, L. Yan, H. Huang, and J. Li. Landsat 15-m panchromatic-assisted downscaling (lpad) of the 30-m reflective wavelength bands to sentinel-2 20-m resolution. *Remote Sensing*, 9(7):755, 2017.

[20] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee. Enhanced deep residual networks for single image super-resolution. In *The IEEE conference on computer vision and pattern recognition (CVPR) workshops*, volume 1, page 4, 2017.

[21] C. Ma, C.-Y. Yang, X. Yang, and M.-H. Yang. Learning a no-reference quality metric for single-image super-resolution. *Computer Vision and Image Understanding*, 158:116, May 2017.

[22] G. Masi, D. Cozzolino, L. Verdoliva, and G. Scarpa. Pan-sharpening by convolutional neural networks. *Remote Sensing*, 8(7):594, 2016.

[23] A. Mittal, A. K. Moorthy, and A. C. Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, Dec 2012.

[24] A. Mittal, R. Soundararajan, and A. C. Bovik. Making a completely blind image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, March 2013.

[25] NASA. Landsat—nasa. `https://www.nasa.gov/mission_pages/landsat/main/index.html`. Accessed December 4, 2018.

[26] D. Pouliot, R. Latifovic, J. Pasher, and J. Duffe. Landsat super-resolution enhancement using convolution neural networks and sentinel-2 for training. *Remote Sensing*, 10(3):394, 2018.

[27] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks, 2015.

[28] M. S. M. Sajjadi, B. Scholkopf, and M. Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[29] H. Song, B. Huang, Q. Liu, and K. Zhang. Improving the spatial resolution of landsat tm/etm+ through fusion with spot5 images via learning-based super-resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 53(3):1195–1204, 2015.

[30] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, C. C. Loy, Y. Qiao, and X. Tang. Esrgan: Enhanced super-resolution generative adversarial networks, 2018.

[31] Z. Wang and A. C. Bovik. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE Signal Processing Magazine*, 26(1):98–117, Jan 2009.

[32] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu. Residual dense network for image super-resolution, 2018.