Perceptual improvements for Super-Resolution of Satellite Imagery

Daniel Bull
Dept of Computer Science
University of Waikato
Hamilton, New Zealand
danbull1@yahoo.co.nz

Nick Lim
Dept of Computer Science
University of Waikato
Hamilton, New Zealand
nlim@waikato.ac.nz

Eibe Frank
Dept of Computer Science
University of Waikato
Hamilton, New Zealand
eibe@waikato.ac.nz

***Abstract*—Super-resolution of satellite imagery poses unique challenges. We propose a hybrid method comprising two existing deep network super-resolution approaches, namely a feed-forward network called DeepSUM, and ESRGAN, a GAN-based approach, to super-resolve multiple low-resolution images by a factor of four to obtain a single high-resolution image. We also introduce a novel loss function, called variation loss, to better define edges and textures to create a sharper, perceptually better output. Using our hybrid, we inherit some of the advantages of both deep learning approaches, resulting in super-resolved images that better show boundaries, textures, and details.**

***Index Terms*—Super-Resolution, Computer Vision, Remote Sensing, Deep Neural Network**

Super-resolution (SR) is the process of obtaining a higher-resolution image from a single or multiple lower-resolution images. It has a wide variety of applications including medical imaging, security imaging, and satellite remote sensing [Yang et al., 2019]. However, super-resolution is an inherently ill-posed problem as a multiplicity of solutions exist for the low-resolution image [Dong et al., 2015]. To overcome this issue, sophisticated super-resolution methodologies attempt to exploit contextual information to infer the missing high-resolution components. While most super-resolution studies focus on generic photo imagery, there is a substantial body of work focusing on super-resolving satellite imagery, the topic of this paper, which presents its own unique challenges.

Many studies address super-resolution as part of the wider context of image processing using techniques such as de-noising, compressing and imprinting imagery. Before the advent of modern machine learning, interpolation-based methods such as bicubic interpolation and Lanczos resampling were used for super-resolution [Yang et al., 2019], where the value of each pixel is estimated based on the surrounding pixels. Although these methods are fast and do not require other data to work from, they suffer from low accuracy and blur high-resolution details due to a lack of information about how to resolve the pixel values.

With the advent of deep convolutional neural networks (CNN) came the ability to learn pixel values from very high-level feature maps. These super-resolution methods can be divided into single-image SR (SISR) and multi-image SR (MISR). As the names suggest, SISR infers higher resolution data from a single image, whereas MISR takes advantage of the information gain presented by multiple complementary images of the same scene to better infer pixel values [Molini et al., 2019]. One of the first SISR methods using deep CNNs [Dong et al., 2015] learnt features via hidden layers, rather than explicitly learning a dictionary of image features. In the model, called Super-Resolution Convolutional Neural Network (SRCNN), the low-resolution image is first upscaled to the desired size using bicubic interpolation. The image is then flattened and passed through the CNN to generate an output. This method was found to be better than any of the previous SR methods available at the time.

Several recent studies have looked at using Generative Adversarial Networks (GANs) to enhance the product of super-resolved data. These include [Hoque et al., 2019], [Jiang et al., 2019], [Ledig et al., 2017], and [Wang et al., 2018]. GANs utilise adversarial networks consisting of a generator network and a discriminator network. The generator network produces a high-resolution image similar to the original high-resolution image, whereas the discriminator acts as a judge and determines whether the output image is plausibly fake. In this way, the discriminator guides the generator to produce steadily more realistic images, until the discriminator is unable to distinguish the output of the generator and the ground truth image. In the context of super-resolution, this will likely produce an image that is not the same as the ground truth but has perceptual qualities that appear to be identical to the human visual system. GANs offer the possibility of photo-realistic images at large up-scaling factors [Ledig et al., 2017]. However, while GANs can successfully generate photo-realistic images, they have a tendency to "hallucinate" details.

I. Remote Sensing and Satellite Imagery

Remote sensing is the use of electromagnetic energy to measure the physical properties of distant objects. The history of remote sensing can be traced back to World War I and World War II when millions of aerial photographs had to be manually analysed for military purposes [Moore, 1979]. The development of remote sensing platforms progressed rapidly through the twentieth century. A key moment was the launch of the first Landsat satellites in 1972, which were the first dedicated earth landcover imaging satellites. For the first time,

978-1-6654-0645-1/21/$31.00 ©2021 IEEE

repetitive images of the earth were easily available for analysis. The first Landsat satellites (1 and 2) carried a green and red sensor and two NIR sensors. More recent Landsat satellites (Landsat 8 and 9) can acquire data from 11 spectral bands at between 15 and 30m resolution, vastly increasing the amount of information that can be acquired [Wulder et al., 2019]. A parallel effort by the European Space Agency (ESA) created the SPOT satellites with a relatively high resolution of 2.5 RGB bands, and more recently the Sentinel series of satellites, whose imagery we use in this study. In the last decade, cube satellites, such as those launched by *Planet* provide a higher return rate with images available up to two times daily [Planet Labs Inc, 2021].

Satellite data is comparatively cheap compared to aerial photography, so it is used for tasks that require large area coverage or regular coverage over time. However, it does suffer from the issue that commercial satellite images generally have a lower resolution than aerial or drone imagery, so it cannot always resolve features to a sufficiently high level. In 2020, the highest-resolution commercially available satellite data was from satellite WorldView-3 with a 30cm ground sample distance (GSD). Other satellites had submeter imagery products of GSD 50cm including WorldView-2, GeoEye-1, Pleiades. This resolution is still not sufficient for tasks such as traffic monitoring and animal tracking [Zhu et al., 2020].

One unique issue with satellite imagery is cloud cover. In 12 years of observations by the Moderate Resolution Imaging Spectroradiometer (MODIS), it was found that 67% of the Earth's surface is covered by clouds on average. (Note that this is lower overland with only 55% of the area covered in cloud, and cloud cover is much lower on average during late summer and early autumn [Meraner et al., 2020].) Additionally, satellite data also suffer from noise, caused partially by inaccuracies in the point spread function (PSF) of the imaging system and motion blur caused by satellite movement sensor scanning. Moreover, the imagery is likely to be processed and resampled, which causes further blur [Zhu et al., 2020] especially when compensating for atmospheric distortion.

These issues are typically not taken into consideration in general super-resolution methods, which often use down-sampled higher-resolution data to generate low-resolution training data [Hoque et al., 2019], [Johnson et al., 2016], [Wang et al., 2018]. This model assumes that either the degradation model can be characterised in some way or has a limitation in that the degradation model might not match reality [Molini et al., 2019]. In this paper, we propose a fusion method combining both a convolutional neural network method (namely DeepSUM) and a GAN approach (namely ESRGAN), with the goal to perform perceptually consistent and accurate super-resolution of satellite imagery. [1]

[1]For more detail on the work presented in this paper, see the first author's MSc thesis, which is available at the University of Waikato research repository [Bull, 2021].

II. Methodology

DeepSUM [Molini et al., 2019] is a CNN developed by a team at *Politecnico di Torino* for super-resolving multiple unregistered temporal images to a single higher-resolution image by a factor of three. In the original algorithm, the low-resolution images with 128×128-pixels were up-sampled to 384×384-pixels. DeepSUM employs a supervised deep learning approach, where the CNN learns the residual between a bicubic interpolation and the ground truth. Using multiple images, DeepSUM aims to explore the extra information provided by the temporal depth.

The method was the winner of the PROBA-V super-resolution challenge issued by the European Space Agency (ESA) [Märtens et al., 2019]. In the PROBA-V challenge, the teams were given multiple images from each of 78 Earth locations that needed to be super-resolved and checked against a high-resolution image taken from the same satellite. The satellite data used by PROBA-V consisted of top-of-atmosphere reflectances for the red and NIR spectral bands at 300m (LR) and 100m (HR) resolution. Each image came with a quality map indicating pixels affected by cloud, shadow, ice, water etc. Each data point contained one HR image and several LR A recorded within 30 days of each other. This set of images was referred to as an *image set*. At each location, there were up to 19 different LR images and at least 9 HR images of the area. A unique feature of this data was that the HR and LR images were separately acquired by the same satellite, as opposed to using artificial data, i.e., data that had been previously downsampled from an HR image [Molini et al., 2019]. The competition expected LR images to be super-resolved from 128×128-pixels to a 384×384-pixel image.

Unlike DeepSUM, the ESRGAN is a GAN-based super-resolution algorithm that is designed to generate perceptually consistent super-resolution images [Wang et al., 2018]. As we will observe in the next section, the output of DeepSUM lacks the pixel-by-pixel variation of the ground truth images, and some of the sharpness and detail is not recreated by the network, particularly when a mean-squared error (MSE) loss is used. Hence, we use ESRGAN to infer additional details in the super-resolution image. We should note here that ESRGAN is not designed for imprinting clouds, and as such, would not be suitable as a drop-in replacement for DeepSUM.

A. DeepSUM Algorithm Changes

We adapted the DeepSUM algorithm to super-resolve Sentinel-2 data by a factor of 4 (rather than 3 as in the original algorithm), using aerial imagery as ground truth. This change was made to better match the data, as 10m pixels from Sentinel-2 resolve well to 4×2.5m pixels with no rounding required. This change also means the up-sampling factor is the same as used in ESRGAN, which makes the methods consistent when used jointly. Several changes were made to the algorithm to facilitate using these datasets with a different up-scaling factor.

Tweaks were also required for the DeepSUM algorithm to work with the different number of images in each image set in

our problem (8 vs. 9). In the *Fusion Net* subnetwork, feature maps from each of the individual images are combined to create a single image. The original algorithm used the best 9 images in a set, and these were reduced down to a single image using four $3\times3\times3$ 3D convolutional layers. A reduced image set size necessitated a minor architectural change. In place of the four $3 \times 3 \times 3$ 3D convolutions, three $2 \times 3 \times 3$ 3D convolutions were used, followed by a single $3\times3\times3$ 3D convolution. This is illustrated in Figure 1.

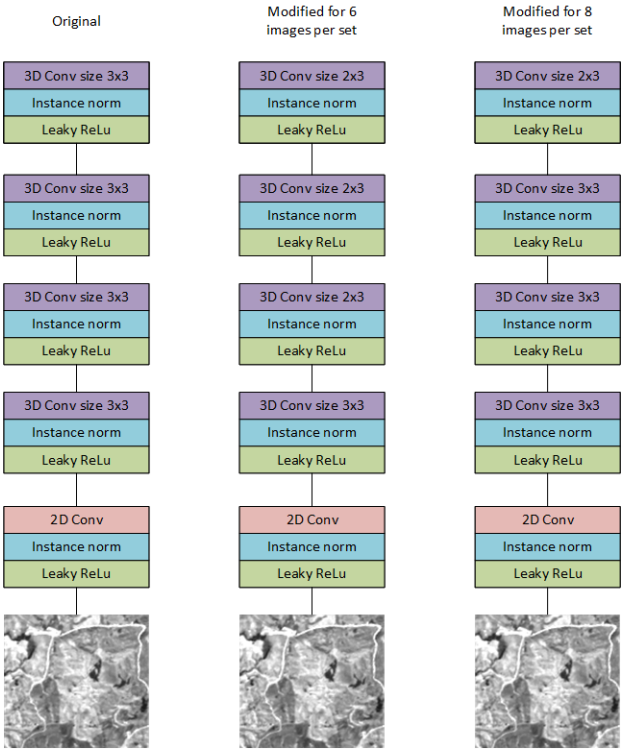

Fig. 1: Fusion block modifications for different image set sizes.

Another notable change in our implementation is the creation of different loss functions, and the use of additional metrics, namely SSIM, perceptual loss, and variation loss.

B. Loss Functions

Mean squared error (MSE or ℓ_2) and mean absolute error (MAE or ℓ_1) are dominant loss functions used across machine learning that are applied in a wide range of applications including super-resolution. The reason these losses are so popular is that they are convex and often available pre-packaged in software libraries that make them easy to use [Zhao et al., 2016]. In addition, MSE is simple and easy to understand, and relatively computationally cheap [Wang and Bovik, 2009]. In the context of image processing, MSE and MAE are the error signal between the original image and the distorted image. In super-resolution, this is the error between the HR and upsampled LR images.

Unfortunately, pixel-wise loss functions such as MSE struggle to handle the uncertainty inherent in recovering lost high-frequency details. As discussed in detail by [Ledig et al., 2017], pixel loss functions can appear overly smooth due to the pixel-wise average of possible solutions in the pixel space. By favouring an average over the plausible HR solutions, a significant reduction of high-frequency details occurs [Lugmayr et al., 2020]. This leads to the issue that although using an MSE loss may have a high peak signal-to-noise ratio (PSNR), it correlates poorly with image quality as perceived by a human observer. To address this shortcoming, the structural similarity index measure (SSIM) and the multi-scale structural similarity index measure (MS-SSIM) are often proposed as an alternative to MSE for image processing tasks. SSIM attempts to mimic aspects of the human visual system in focusing on the structural similarity between the images rather than the luminance and the contrast [Wang et al., 2004]. Intriguingly, although SSIM and MS-SSIM are commonly-cited functions to measure image distortion, they are not commonly used as loss functions when training super-resolution models. This is even though differentiable versions exist and exhibit obvious advantages over MSE in the context of image quality [Zhao et al., 2016].

However, SSIM is also pixel based and does not capture stylistic differences between the output and ground-truth image. Ideally, in super-resolution, fine details are inferred from visually ambiguous low-resolution imagery [Johnson et al., 2016]. Both MSE and SSIM have been found to correlate poorly with human assessment of visual quality, as both capture low-level differences between pixels.

Our preliminary studies showed that the super-resolved images obtained from DeepSUM often produced a smooth output as pixel values trended towards the mean. Highly textured objects appeared largely monochrome and lacked the texture of a true image. To rectify this, and to better replicate the variation and the perceptual qualities of the ground truth, we introduce a new loss function to encourage pixel-by-pixel variance, which we dub "variation loss". We calculate the variation loss using the following steps. For each mini-batch of output imagery and ground truth encountered during training of the network, we produce 9 copies of the minibatch, and each of those 9 copies is obtained by moving the original image by each of $-1, 0, 1$ in the X and Y dimension respectively. Following this, the off-centre copies of the original mini-batch (and the images which were not moved i.e., moved by (0,0)) are stacked together and variance matrices are created using only the stack dimension as illustrated in Figure 2. This effectively creates a mini-batch of image variances. The output imagery and the prediction imagery are compared, and the difference is calculated using mean squared error (or alternatively, absolute error). In our experiments, the variation loss is combined with other loss functions using a *variance factor* hyper-parameter to increase or decrease its effect on the overall loss. Mathematically, assume N is the number of samples, $x = (x_i|i = 1, 2, \cdots, N)$ is the predicted output, and $y = (y_i|i = 1, 2, \cdots, N)$ is the ground truth. Then, V_x is the variance of the output and V_y is the variance of the ground-truth HR image. Variation loss can be defined by the

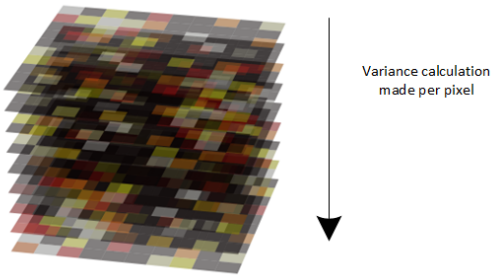

Fig. 2: Variation is captured by stacking copies of an image on top of each other, where copies are obtained by moving each image by 0 or 1 pixels in an X and Y direction, and calculating the variation through the z axis.

equation:

$$L_{var}(x, y) = \frac{1}{N} \sum_{i=1}^{N} (V_{xi} - V_{yi})^2 \quad (1)$$

In our experiments, results from using a variation loss component in the loss function showed an improvement in output quality in two ways. Land cover texture appeared more realistic and similar to the ground-truth HR image. A second, more surprising effect, was that the features were crisper and less undefined. This effect can be explained by the variation loss working to preserve high variance in boundary areas by forcing the high pixel intensities higher and the low pixels intensities lower, and so enhancing the edge effect, which is blurred by a pixel-based loss.

We also consider the *perceptual loss function* measured by the learned perceptual image patch similarity metric (LPIPS), which aims to capture differences based on high-level feature representations rather than pixel-based differences. It does so by employing style transfer. In this paradigm, content is defined as the larger spatial structure in the image, whereas style refers to the colours and local structures of the image. The insight that allows this transfer is that higher layers (or the layers closer to the output) in a deep neural network capture the high-level content in terms of objects and their arrangement in the input image, but do not contain information about detailed pixel values. The key finding here is that style and content representations are separable. When CNNs are trained for object recognition, they develop a representation of an image that is increasingly explicit, i.e., further along the network, feature maps are increasingly about content rather than style [Gatys et al., 2015].

To apply style from one image to the content of another image, loss functions must be devised that allow this transfer. The method used by [Johnson et al., 2016] to aid super-resolution is to create a network with two components: an *image transfer network* and a *loss network*. As existing image classification networks that are publicly available have already learnt to encode perceptual and semantic information, such a pre-trained network is used as a fixed *loss network*. The networks most used are VGG networks pre-trained on ImageNet or the MS-COCO dataset [Johnson et al., 2016].

III. Results and Discussion

In our experiments, low-resolution satellite data of the North Island of New Zealand and corresponding high-resolution aerial photography is used to train DeepSum [Molini et al., 2019] as well as ESRGAN and our hybrid. When interpreting the results we present, note that large PSNR and SSIM values, and a small LPIPS value, are usually preferable.

Figure 3 illustrates some results obtained in our experiments. Row 1 shows up-sampled satellite images obtained using basic bicubic upsampling. Row 2 is the output of super-resolution using ESRGAN. Observe that there are significant artefacts in the regions with clouds. Row 3 is the output of super-resolution using DeepSum. Row 4 is generated using our fusion method, combining the output of DeepSum and ESRGAN using variation loss as a component in the loss function. Row 5 is the high-resolution ground truth from the aerial photography.

Table I shows a comparison of the peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), and learned perceptual image patch similarity (LPIPS) of the four approaches to super-resolution, namely bicubic upsampling, DeepSum, ESRGAN, and the hybrid of DeepSum and ESRGAN. Note here that we are reporting the values for the network trained using MSE loss. Also note here that the ESRGAN-only approach is not suited for satellite image super-resolution and is provided only for reference.

Table II compares the different metrics when trained using different loss functions. Here, we see that for the DeepSUM-only network, the variation loss (combined with MSE using $0.4\times$MSE$+0.6\times$Variation loss) gives us the best LPIPS score. Interestingly, the DeepSUM network trained using perceptual loss results in a poorer LPIPS score compared to the network trained using variation loss. We would like to note also that as one can see in Figure 4, the output with variation loss tends to create output with stronger edges and boundaries, and arguably better textures, which unfortunately is not captured by the three metrics used.

IV. Conclusion and Future Work

Our results show the trade-off between optimising a result against pixel-based metrics such as SSIM and PSNR and optimising for more perceptually-based metrics. Different loss functions run in our experiments show that the original CNN algorithm DeepSUM can be improved. The addition of ESRGAN to the process shows how the data can be made to appear more photo realistic. In particular, the GAN creates more realistic textures and fine detail; however, this comes at some uncertainty as to the veracity of minor detail. Using the novel variation loss function introduced in this

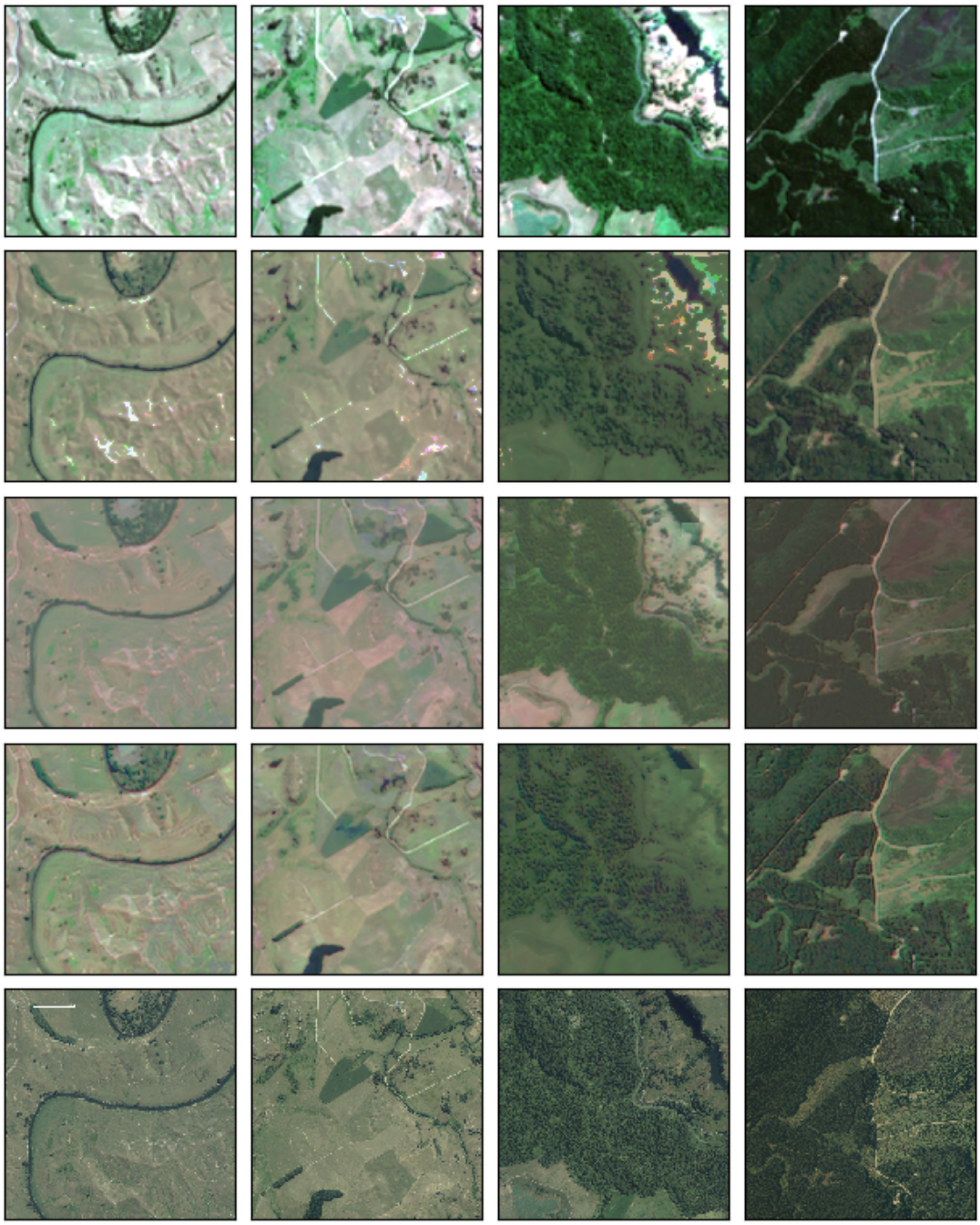

Fig. 3: Result of the super-resolution of the satellite images. Row 1 shows up-sampled satellite images obtained using bicubic upsampling (baseline). Row 2 is the output of super-resolution using ESRGAN. Row 3 is the output of super-resolution using DeepSum. Row 4 is generated using our fusion method, combining the output of DeepSum and ESRGAN using variational loss as the loss function. Row 5 is the high-resolution ground truth from the aerial photography.

	bicubic upsample			DeepSUM only			ESRGAN only			DeepSUM then ESRGAN		
land use class	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
overall result	19.1	0.30	0.34	**20.1**	**0.31**	0.35	19.5	0.22	0.21	19.7	0.25	**0.20**
farmland	18.1	0.44	0.25	**20.1**	**0.46**	0.25	19.3	0.30	0.16	19.7	0.40	**0.15**
bush	19.8	0.23	0.38	**20.1**	**0.24**	0.40	19.6	0.18	0.24	19.6	0.19	**0.22**
mixed	18.1	0.38	0.30	**20.2**	**0.38**	0.30	19.8	0.27	0.17	19.8	0.30	**0.16**

TABLE I: The performance of the different approaches on the accuracy and perceptual metrics.

	bicubic upsample			DeepSUM only			DeepSUM then ESRGAN		
DeepSUM loss func	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
MSE loss	19.1	0.30	0.34	20.1	0.31	0.35	19.7	0.25	0.20
MSE + Variation loss				19.0	0.23	0.23	19.0	0.21	0.21
SSIM loss				**20.3**	**0.33**	0.33	18.1	0.19	0.26
Perceptual loss				17.6	0.32	0.33	19.4	0.21	**0.19**

TABLE II: The effect of loss function on DeepSUM output and ESRGAN output.

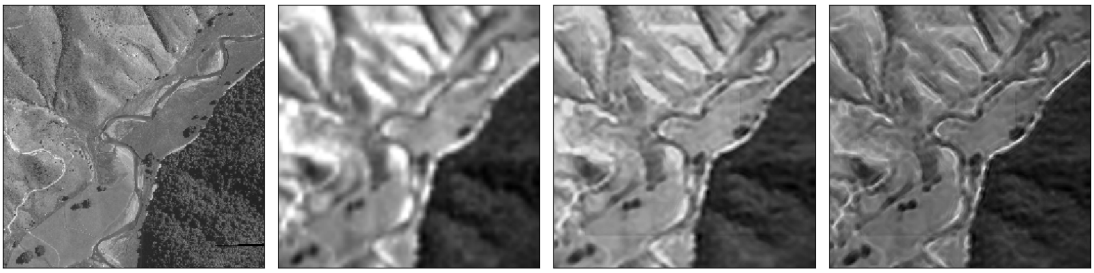

Fig. 4: Images showing the effect of the variation loss on image quality. Zoomed-in samples taken from the top right of our test image-52. Top-left to bottom-right: HR/ground-truth image, LR image bicubically upsampled, output using MSE loss only without variation loss, and output using MSE loss including variation loss.

paper with DeepSUM produces a crisper final output with stronger edges and boundaries, and arguably better textures. However, this result is not reflected in measurement metrics. Note that, as each output image is an amalgamation of several temporarily different inputs, the result cannot possibly be a true representation of any real image. In this sense, using a GAN to make the DeepSUM output appear more realistic-looking to a human is appropriate, as the output will never correspond exactly to reality anyway.

REFERENCES

[Bull, 2021] Bull, D. (2021). Super-resolution of satellite imagery. Master's thesis, University of Waikato.

[Dong et al., 2015] Dong, C., Loy, C. C., He, K., and Tang, X. (2015). Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307.

[Gatys et al., 2015] Gatys, L. A., Ecker, A. S., and Bethge, M. (2015). A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, pages 1646–1654.

[Hoque et al., 2019] Hoque, M. R. U., Burks, R., Kwan, C., and Li, J. (2019). Deep learning for remote sensing image super-resolution. In *2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, pages 0286–0292. IEEE.

[Jiang et al., 2019] Jiang, K., Wang, Z., Yi, P., Wang, G., Lu, T., and Jiang, J. (2019). Edge-enhanced gan for remote sensing image superresolution. *IEEE Transactions on Geoscience and Remote Sensing*, 57(8):5799–5812.

[Johnson et al., 2016] Johnson, J., Alahi, A., and Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer.

[Ledig et al., 2017] Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al. (2017). Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690.

[Lugmayr et al., 2020] Lugmayr, A., Danelljan, M., Van Gool, L., and Timofte, R. (2020). Srflow: Learning the super-resolution space with normalizing flow. In *European Conference on Computer Vision*, pages 715–732. Springer.

[Meraner et al., 2020] Meraner, A., Ebel, P., Zhu, X. X., and Schmitt, M. (2020). Cloud removal in sentinel-2 imagery using a deep residual neural network and sar-optical data fusion. *ISPRS Journal of Photogrammetry and Remote Sensing*, 166:333–346.

[Molini et al., 2019] Molini, A. B., Valsesia, D., Fracastoro, G., and Magli, E. (2019). DeepSUM: Deep neural network for super-resolution of unregistered multitemporal images. *IEEE Transactions on Geoscience and Remote Sensing*, 58(5):3644–3656.

[Moore, 1979] Moore, G. K. (1979). What is a picture worth? a history of remote sensing/quelle est la valeur d'une image? un tour d'horizon de télédétection. *Hydrological Sciences Bulletin*, 24(4):477–485.

[Märtens et al., 2019] Märtens, M., Izzo, D., Krzic, A., and Cox, D. (2019). Super-resolution of proba-v images using convolutional neural networks. *Astrodynamics*.

[Planet Labs Inc, 2021] Planet Labs Inc (2021). Planet website. https://www.planet.com. Accessed: 2021-03-03.

[Wang et al., 2018] Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., and Change Loy, C. (2018). Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0.

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

[Wang et al., 2004] Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612.

[Wulder et al., 2019] Wulder, M. A., Loveland, T. R., Roy, D. P., Crawford, C. J., Masek, J. G., Woodcock, C. E., Allen, R. G., Anderson, M. C., Belward, A. S., Cohen, W. B., Dwyer, J., Erb, A., Gao, F., Griffiths, P., Helder, D., Hermosilla, T., Hipple, J. D., Hostert, P., Hughes, M. J., Huntington, J., Johnson, D. M., Kennedy, R., Kilic, A., Li, Z., Lymburner, L., McCorkel, J., Pahlevan, N., Scambos, T. A., Schaaf, C., Schott, J. R., Sheng, Y., Storey, J., Vermote, E., Vogelmann, J., White, J. C., Wynne, R. H., and Zhu, Z. (2019). Current status of landsat program, science, and applications. *Remote Sensing of Environment*, 225:127–147.

[Yang et al., 2019] Yang, W., Zhang, X., Tian, Y., Wang, W., Xue, J.-H., and Liao, Q. (2019). Deep learning for single image super-resolution: A brief review. *IEEE Transactions on Multimedia*, 21(12):3106–3121.

[Zhao et al., 2016] Zhao, H., Gallo, O., Frosio, I., and Kautz, J. (2016). Loss functions for image restoration with neural networks. *IEEE Transactions on computational imaging*, 3(1):47–57.

[Zhu et al., 2020] Zhu, X., Talebi, H., Shi, X., Yang, F., and Milanfar, P. (2020). Super-resolving commercial satellite imagery using realistic training data. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 498–502. IEEE.