Super-Resolution Algorithm Applied in the Zoning of Aerial Images

J. A. Baldion(✉), E. Cascavita, and C. H. Rodriguez-Garavito

Automation Engineering Program, La Salle University, Bogotá, Colombia
{jbaldion06,ecascavita17,cerodriguez}@unisalle.edu.co

Abstract. Nowadays, multiple applications based on images and unmanned aerial vehicles (UAVs), such as autonomous flying, precision agriculture, and zoning for territorial planning, are possible thanks to the growing development of machine learning and the evolution of convolutional and adversarial networks. Nevertheless, this type of application implies a significant challenge because even though the images taken by a high-end drone are very accurate, it is not enough since the level of detail required for most precision agriculture and zoning applications is very high. So, it is necessary to further improve the images by implementing different techniques to recognize small details. Hence, an alternative to follow is the super-resolution method, which allows constructing an image with the information from multiple images. An efficient tool can be obtained by combining drones' advantages with different image processing techniques. This article proposes a method to improve the quality of images taken on board in a drone by increasing information obtained from multiple images that present noise, vibration-induced displacements, and illumination changes. These higher resolution images, called super-resolution images, allow supervised training processes to perform different zoning methods better. In this study, GAN-type networks show the best results to recognize visually differentiated ones on an aerial image automatically. The quality measure of the super-resolution image obtained by different methods was defined using sharpness and entropy metrics, and a semantic confusion matrix measures the accuracy of the following semantic segmentation network. Finally, the results show that the super-resolution algorithm's implementation and the automatic segmentation provide an acceptable accuracy according to the defined metrics.

Keywords: Super-resolution · Zoning · Modified linear interpolation · Semantic segmentation

1 Introduction

Technological development has allowed the advance in Unmanned Aerial Vehicles (UAV) and their use in a wide range of applications [1], such as precision agriculture, search and rescue, remote sensing, and infrastructure inspections. Likewise,

Supported by Universidad de La Salle Bogotá-Colombia.

H. Florez and M. F. Pollo-Cattaneo (Eds.): ICAI 2021, CCIS 1455, pp. 346–359, 2021.
https://doi.org/10.1007/978-3-030-89654-6_25

UAVs have advantages when capturing images because they can cover extensive zones and take bursts of overlapped images quickly. Nevertheless, these features are not enough for most practical applications related to precision agriculture, for instance, crop health analysis base on leaf plant symptoms. One approach for improving details in regions of interest on UAV images consists of creating from multiple low-resolution overlapped images a suitable Super-resolution image using SISR (Single Image Super-Resolution) [16,18] or MIRS (Multiple Images Super Resolution) [2,14,15,17] algorithms.

Different approaches based on interpolation [19], reconstruction [20], and learning [21,22] tackle the lack of details observed in images taken from UAVs. Several works focus on using Convolutional Neural Networks to decompose, hierarchically, intrinsic and context features from a low-resolution image to generate an image with more information expressed by its level of details.

Improving the quality and resolution of the images captured from drones on flying opens an enormous range of applications like the diagnosis of crops [23], control production in agriculture [24], fault detection [25,26], analysis of human settlements [27], among others.

A super-resolution image in the context of zoning of a geographical area can provide more visual information than a single UAV image, let it recognize water bodies, rural extension, and delimitation of crop areas. At this point, the authors propose a variant of the traditional linear interpolation method for making a super-resolution image, comparing this with super-resolution images generated by other methods base on neural networks. Finally, the best super-resolution method is determined, and automatic zoning from the PIX2PIX [3] network performs a semantic segmentation with the super-resolution image. The improved zoning is measured to check if the precision grows up using the methodology proposed.

2 Generation of Super-Resolution Images

Some applications like zoning for development territorial planning use very high-resolution images. Since UAV's camera captures its field of view from a high height, image resolution is not enough for trim details, so super-resolution methods offer a way to increase visual information. This section describes different techniques of super-resolution, such as a modified Linear interpolation method proposed by the authors, and other techniques found in the literature, such as ESRGAN [4] and PIX2PIX [3].

2.1 Modified Linear Interpolation

Modified linear interpolation is dependent on linear interpolation, but with some modifications made by the authors. In the first instance, Linear interpolation is a MISR method; therefore, it needs several images from the same geographic area to generate a super-resolution image. One hundred ninety-six captures of the same scene form an image bank (Available in: https://github.com/juanbaldion/

SuperResolutionImages.git), where the first capture is the reference image, and each one of the others is selected as a test image to implement this method. The image shown in Fig. 1 has a size of 3648×5472 pixels, but since the interpolation of this scene has a high computational cost, only a section of this image will be treated with this technique; this zone of interest will have a size of 912×1513 pixels. Afterward, each image will be loaded by trimming the section around a specific coordinate point, but due to changes in camera pose caused by the drone overflight, it is necessary to transform every capture to a standard top perspective. Therefore, relevant points in the images, reference and test, are detected using SIFT features [5] to extract local descriptors and then return matching feature points. Then, to calculate the change of perspective between each pair of images, I_{ref} and I_{test_i}, a parametric estimator RANSAC (Random Sample Consensus) [6] is implemented to obtain the Homography matrix ${}^{i}H_0$. The reference image must be enlarged in this method, considering that a higher amplitude factor will cause a higher computational cost to continue with the super-resolution process, the expansion of the base image will only be three times its original size, introducing intermediate black pixels within the pixels of the base image. The filling process is done by taking one by one of the black pixels of the enlarged image I_{ref} located in coordinates $[u, v]$. Then, multiplying by the homography matrix ${}^{0}H_i$ and the scale factor $\mathcal{K}_s$, it is possible to know the coordinates in which each pixel $[u, v]$ is projected on I_{test_i}, $[x, y]$. Later, the linear interpolation is made on the corresponding pixel $[x, y]$, so that it is possible to obtain the intensity $I_{test_i}(x, y)$ according to the current pixel and its j neighbours (x_j, y_j). Hence, already having the intensity $I_{test_i}(x, y)$, it will be replaced in the black pixel in question ${}^{i}I_{ref}(u, v)$. This process is performed for each of the pixels that are in black Fig. 2. Finally, in order to improve accuracy in interpolated information added to super-resolution image I_{ref}, the process is repeated with all burst of images taken around the zone of interest, and the $I_{ref}(u, v)$ are averaged across interpolated intensities extracted from all test images i^{th} ${}^{i}I_{ref}(u, v)$.

$$[x, y] =^{i} H_0 \mathcal{K}_s \begin{bmatrix} u \\ v \end{bmatrix} ;^{i} I_{test}(x, y) = \sum_{n=1}^{j} I_{test}(x_j, y_j) \frac{|(x, y) - (x_j, y_j)|}{\sum_{n=1}^{j} |(x, y) - (x_j, y_j)|} \quad (1)$$

(a) Random image. (b) Zone of Interest.

Fig. 1. (a) Random image of the 196 captures, (b) Interest zone magnified three times compared to the base image.

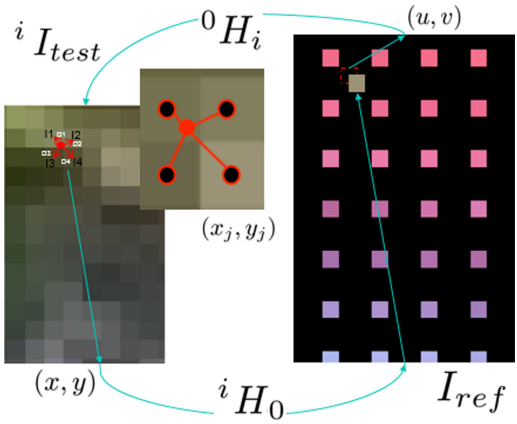


Fig. 2. Interpolation based on homography matrix.

The process develops each pixel's average in all the channels, resulting in a super-resolution image containing combined information of all the processed images. At this point, the result is a super-resolution image, but some sections of the image encounter strange intensity patterns. The method called blind deconvolution is used to improve regions that contain this distortion [7].

2.2 ESRGAN

It is a GAN (Generative Adversarial Network) type architecture. This type of network is composed of 2 convolutional neural networks that solve two different tasks. A U-Net type generator network which create the desired content. Additionally, there is the discriminating network. This network is used to know whether the output image is real or false, i.e., whether an image was taken from the set of images or is generated from the generator network. These networks are in opposition. One always seeks to deceive the other and in this way both networks improve over time.

2.3 PIX2PIX

As in ESRGAN, this is also a GAN-based architecture, the difference is that it is conditional, which is known as cGAN.

This architecture stands for "picture to picture", which means that this network is going to be in charge of creating images, not modifying the ones that already exist. The architecture is in charge of translating images from one domain to another domain.

3 Automatic Image Zoning

Firstly, an image storage bank with 250 samples from free databases was collected constrained to exhibit the following geographical areas: Rural zone, Green zone,

Roads, and Water bodies, taking into account the dimensions of each image for standardization. However, because deep learning required an extensive data set, data augmentation was implemented, and then automatic zoning was done based on semantic segmentation using the PIX2PIX network mentioned above. To do this, we must have output images to which the network must be adjusted. The tool used for labelling was Inkscape which helped to paint the areas of interest mentioned above on the input images, as shown in Fig. 3. The automatic image zoning continues modifying the PIX2PIX network; a convolutional layer, a normalization layer, and an activation layer are added to use over the segmented images. Since the training process is a time-consuming task, it halts after 267 epochs, because the error remains constant at 0.17%. The time needed for this was 35 h, 6 min, and 18 s. The result of this process is shown in Fig. 4.

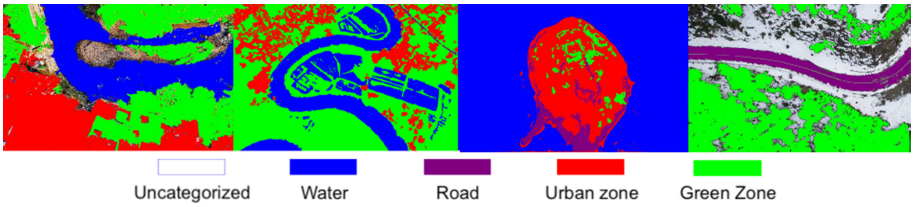


Fig. 3. Manual segmentation of the input images.

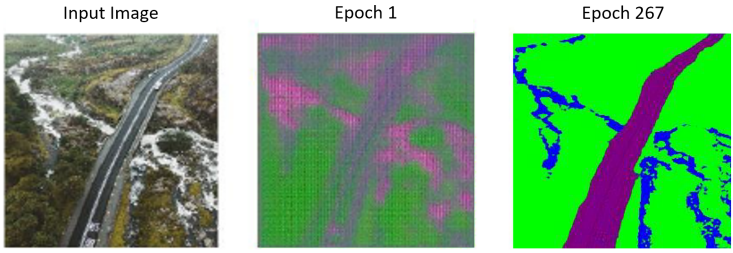
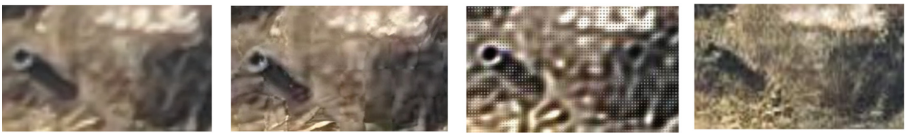


Fig. 4. Segmentation training.

4 Summary Results

4.1 Super-Resolution

By implementing ESRGAN, PIX2PIX, and the Modified Linear Interpolation methods over Fig. 1 and set a ROI, the results are shown in Fig. 5.



(a) Original (b) ESRGAN (c) Mod. Lin. Interp. (d) PIXTOPIX

Fig. 5. Prediction of the implemented techniques, (a) Original Image, (b) ESRGAN Network, (c) Technique, (d) PIX2PIX Network

Metric in Super-Resolution Algorithms. Two metrics are used to measure each of the resulting images in the super-resolution process, allowing them to determine with greater precision which yields the most accurate super-resolution image and is closest to reality.

Entropy Metric. Entropy is a statistical measure of randomness that can be used to characterize the texture of an image. It is defined as the corresponding states of the level of intensity at which individual pixels can adapt. It is used in the details of the super-resolution image under test, as it provides a better comparison of image details. A higher entropy value means more precise information, i.e., an image that offers more detail [8]. Equation 2 defines the entropy used in this study. Where: $H(S_m)$ is a numerical value representing the entropy of the image in the gray scale, and$P_n(S_m)$ is the probability density, which is calculated from the image's histogram. The quantitative results for this metric are shown in Table 1. The image obtained by modified linear interpolation offers more details than the original image; for this reason, it is determined that this method successfully generates an image of super-resolution.

$$H(S_m) = -\sum_{n=1}^{256} P_n(S_m) * log_2(P_n(S_m) \tag{2}$$

Sharpness Metric. Sharpness is defined as the clarity of image detail [9]. There are different techniques for calculating this metric, but the magnitude of the average gradient is used in this case. This metric allows the calculation rate of intensity change at pixel level [10]. In this study, sharpness is mathematically defined by Eq. 3. Where: SP means the sharpness level of the image, a value close to 0 means that the image is completely blurred, and a value above 20 indicates that the image has good sharpness [11]. ΔFx and ΔFy are the gradients in the X and Y directions. This metric calculates the overall gradient for the whole test image.

$$SP = \overline{\sqrt{\Delta Fx^2 + \Delta Fy^2}} \tag{3}$$

The result of applying all metrics in a test image is shown in Table 1. As can be seen, interpolation shows a higher rate of intensity change at the pixel level, which allows us to determine that this method has greater clarity in its details than other methods.

Table 1. Entropy of the original image and the image resulting from each of the methods.

Super-resolution method	Entropy metric	Sharpness metric
Original	7.4256	13.58014
Interpolation	7.7345	28.117
ESRGAN	7.4188	14.5721
PIX2PIX	7.4201	13.2368

4.2 Zoning

The PIX2PIX network is tested by entering images of the training set observing the behaviour of the network as it is shown in Fig. 6.

Metrics for Zoned Images. Different metrics are used, such as pixel by pixel accuracy and (accuracy, sensitivity, and precision of the image) from the confusion matrix to know the margin of error for the implemented method. That will be done with three images from the test set, see Fig. 7.

Pixel by Pixel Metric. This metric consists of comparing the pixels of two images in the value of their intensity. The input image must be painted manually since the output image is already segmented. Then, the size of the images is adjusted to 1500×1500 pixels. Both images are subtracted pixel by pixel, taking into account the pixel position to be measured, where the result will be a 3-colour image; white and gray are pixels that had differences, and black are areas where

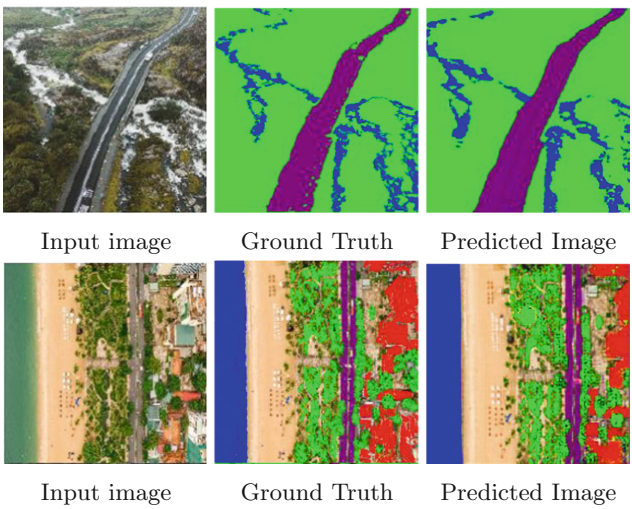


Fig. 6. Result of the network trained after 267 epochs.

the images were painted the same way. This process is repeated for each of the three images with which the network is tested. When there are pixels of all three colours, it must be calculated how many pixels did not match each of the methods' input image and output image. That is done using Eq. 4.

$$PA_c = 100\% - \frac{K_c * 100}{s} \tag{4}$$

where: PA_c corresponds to the success percentage per channel, K_c is the number of pixels where there were differences between the two images, and s is the number of pixels of the corresponding colour in the image. One example of applying this metric can be seen in Fig. 8. Where averaging each of the percentages, and the result is 79.765% accuracy per pixel. Figure 9 shows the results of applying pixel to pixel metric over the three test images.

Confusion Matrix. It shows the segmentation's performance, describing how many categories of the network were correctly predicted, taking into account the natural values of the manually painted image and the one that was predicted by the network [12]. It was necessary to count in each image how many pixels were correctly classified and how many were not to create this matrix. Finally, the metrics of accuracy, precision, and sensitivity are implemented [13]. These are measured for all the zoned images and are given by Eqs. 5–7. Where: Ex is accuracy, Pr is precision, and Se is sensitivity. M represents the number of classes. $Total$ represents the number of samples submitted for classification. TP represents the values marked adequately as positive. FP is the values wrongly marked as positive, and FN the values erroneously Super-resolution algorithm applied in the zoning of aerial images marked as negative. Tables 2, 3, 4 and 5 summarize the results applying segmentation metrics based on Confusion Matrix.

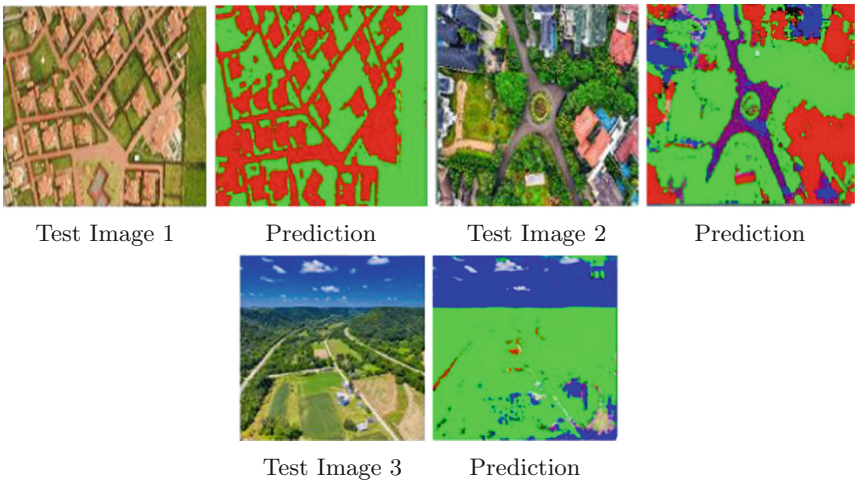


Fig. 7. Zoning Results.

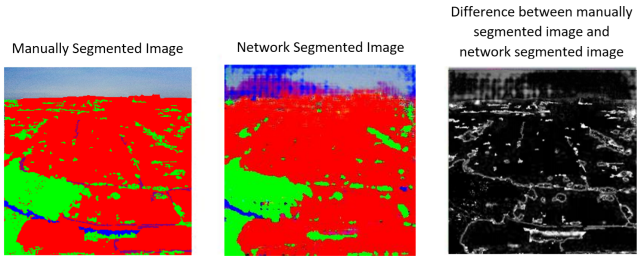


Fig. 8. Result between manually segmented image and network segmented image.

$$Ex = \frac{\sum_{i=1}^{M} TPi}{Total} \tag{5}$$

$$Pr = \frac{\sum_{i=1}^{M} \frac{TPi}{TPi+FPi}}{M} \tag{6}$$

$$Se = \frac{\sum_{i=1}^{M} \frac{TPi}{TPi+FNi}}{M} \tag{7}$$

5 Analysis of Results

After evaluating all super-resolution methods throughout the metrics described above, it can be observed that the best result is the modified linear interpolation because it shows more detailed information within the zone of analysis. Furthermore, this technique resembles a closer reconstruction of reality than the other methods that create or estimate information from a base image. In this way,

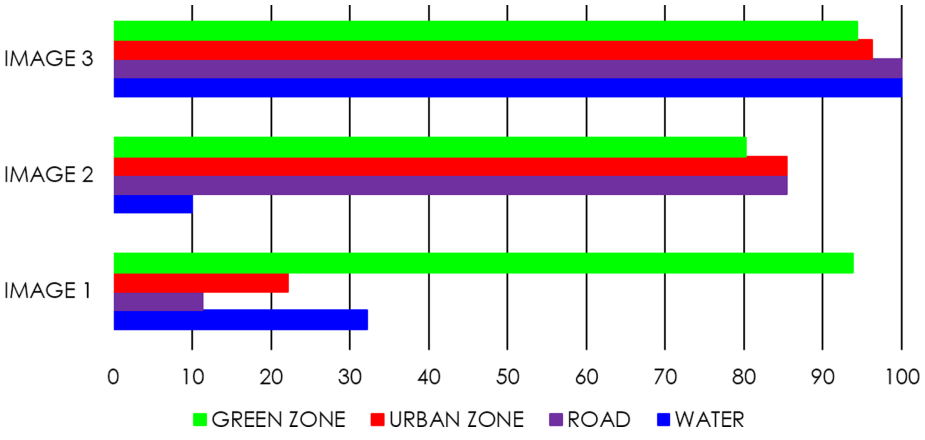


Fig. 9. Zoning Results.

Table 2. Confusion matrix for images 1.

Class	Water	Road	Urban zone	Green zone	Total
Water	51364	236	189	2105	53894
Road	429	539	198	246	1412
Urban zone	47219	3289	1169821	216945	1537274
Green zone	1513	1647	128784	408674	440618
Total	100525	5711	1398992	527970	2033198

modified linear interpolation shows a higher entropy value than the PIX2PIX and ESRGAN images, corresponding to greater detail at the pixel level. Moreover, it can be analysed that the best method, according to the values of the metric sharpness, is the modified linear interpolation too, where high values indicate the clarity of details.

6 Implementation

It consists of taking the image in Fig. 1 as a reference and considering the results obtained by the different metrics, applying the modified linear interpolation output to the PIX2PIX network for zoning by automatic segmentation. The entropy metric measures the level of noise of each image, having that in the original image is 0.70388, while the image in super-resolution is 0.70722. Although the difference is minimal, at the pixel level, it is equivalent to a higher resolution. The accuracy of zoned super-resolution is verified pixel by pixel, and in this way, the percentage of the successful zones is quantified. The visual result of this process can be seen in Fig. 10, and the performance of the super-resolution zoned image is shown in Fig. 11.

Table 3. Confusion matrix for images 2.

Class	Water	Road	Urban zone	Green zone	Total
Water	31577	9354	28662	5334	74927
Road	293	395478	12347	2183	410301
Urban zone	27465	4184	518859	10635	561143
Green zone	1783	2149	19572	1115371	1138875
Total	61118	411165	579440	1133523	2185246

Table 4. Confusion matrix for images 3.

Class	Water	Road	Urban zone	Green zone	Total
Water	0	0	0	0	0
Road	293	395478	12347	2183	410301
Urban zone	0	0	1329416	47894	1377310
Green zone	0	0	89175	706136	795311
Total	0	0	1418591	754030	2172621

Table 5. General metrics data for zoning

Metric	Image 1	Image 2	Image 3
Accuracy	0.7411	0.7824	0.9252
Precision	0.5391	0.7391	0.9684
Sensitivity	0.7261	0.7223	0.9632

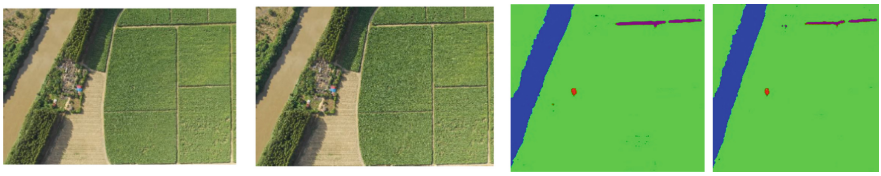


Image of sugar cane field) | Interpolation of sugar cane field | Original segmentation | Super-res. segmentation

Fig. 10. Original and interpolated image; Original and super-resolution zoned image

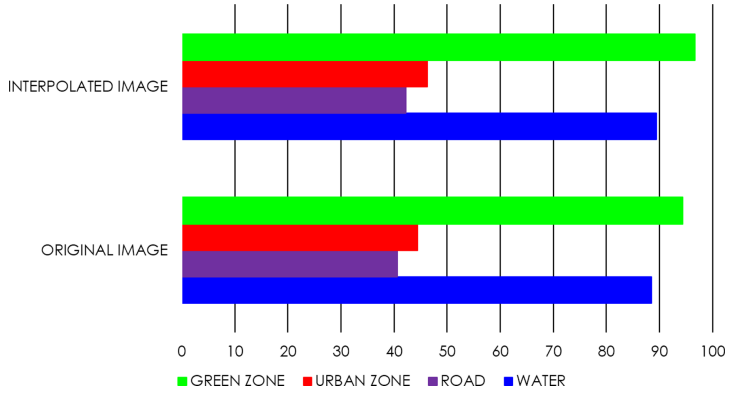


Fig. 11. Pixel-to-pixel comparison between the zoned original image and the super-resolution zoned image.

7 Conclusions

In this work, several ways of addressing two problems related to images captured from drones have been presented, super-resolution and zoning. As for the super-resolution, a considerable visual improvement was obtained by using the modified linear interpolation method, but this method is the one that takes more time for its execution, considering the available comprehensive set of images of the same scene with slight pose and noise variations among the images. The output image from the modified linear interpolation method gives the best results since the final image was obtained from multiple images that are slightly rotated or displaced, which allows the scene to be explored in greater detail. Simultaneously, using convolutional neural network methods, information is estimated from a base image, the intensity of the pixels in each channel. However, it is not real. It was not obtained from many samples but only from one. That is why the modified linear interpolation method proposed by the authors gives the best results because it resembles a closer reconstruction of reality. However, when ESRGAN was used, an improvement could be seen in the objects' textures in the images. As future work, a new methodology could involve the three methods used in this study. We recommend improving the resolution of an input image by first using ESRGAN to obtain better textures. Then implement the modified linear interpolation to provide much more detail to the image, and finally, as this image will have a very high quality, use that image as the target image with which to train the PIX2PIX network so that the network will learn the patterns of an image with much better information. Although the PIX2PIX architecture gave good results in zoning, it was not so in super-resolution. Showing that, even though it is an excellent and versatile model, it does not necessarily give good results for all cases. That is why there are so many types of neural network architectures since each one specializes in a different task and using it in a task for which it was not adequately designed can lead to inefficient performance. Based on the results obtained, it can be determined that the category with the best segmentation was the green area, then the urban area, followed by water, and finally roads and highways. That happens because the green zone tones are very characteristic; it is difficult to find shades of green that are not green zone. Reddish and white tones characterize the urban zone. Water takes mainly blue tones, but water can be white if the sun shines on it or yellowish if the water is dirty. Finally, the roads and highways take only gray and white tones on different scales, confused with water when there are shadows or urban areas when it has much light. One of the advantages of using convolutional neural networks is that they learn in a global rather than local context. Meaning that, in this specific case of zoning, the network does not learn to associate an intensity of one pixel with another pixel's intensity, but rather the network analyses a pixel and its neighbours and, based on that analysis, it makes the prediction. That is an enormous advantage, since, if it were not so, the predicted image would be a mixture of many colours everywhere, since there are many tonalities of the pixel that can correspond to different categories, as it happens with the different tonalities of the water, of the trees or the buildings.

References

1. Shakhatreh, H., et al.: Unmanned aerial vehicles (UAVs): a survey on civil applications and key research challenges. IEEE Access **7**, 48572–48634 (2019)
2. Del Gallego, N.P., Ilao, J.: Multiple-image super-resolution on mobile devices: an image warping approach. EURASIP J. Image Video Process. **2017**(1), 1–15 (2017)
3. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125–1134 (2017)
4. Wang, X., et al.: Esrgan: enhanced super-resolution generative adversarial networks. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops, pp. 63–79 (2018)
5. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision **60**(2), 91–110 (2004)
6. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM **24**(6), 381–395 (1981)
7. Shaw, P., Rawlins, D.: The point-spread function of a confocal microscope: its measurement and use in deconvolution of 3-D data. J. Microscopy **163**, 151–165 (1991)
8. Kopriva, I., Ju, W., Zhang, B., Xiang, D.: Single-channel sparse nonnegative blind source separation method for automatic 3d delineation of lung tumor in PET images. IEEE J. Biomed. Health Inf **21**, November 2017
9. Panetta, K., Gao, C., Agaian, S.: No reference color image contrast and quality measures. IEEE Trans. Consum. Electron. **59**(3), 643–651 (2013)
10. Mlnsa, P., Rodríguez, J.: The Essential Guide to Image Processing. Academic Press, Estados Unidos (2009)
11. Atkins, B.: Sharpness: What is it and How it is Measured (2006). www.imatest.com/docs/sharpness/#:~:text=Image%20sharpness%20can%20be%20measured,distance%3B%20see%20Figure%203
12. Jalife, A., Calderón, G., Fierro, A., Nakano, M.: Clasificación de Imágenes Urbanas Aéreas: Comparación entre Descriptores de Bajo Nivel y Aprendizaje Profundo. Información tecnológica **28**(3), 209–224 (2017)
13. Tharwat, A.: Classification assessment methods. Applied Computing and Informatics (2020)
14. Ur, H., Gross, D.: Improved resolution from subpixel shifted pictures. CVGIP: Graph. Models Image Process. **54**(2), 181–186 (1992)
15. Farsiu, S., Robinson, M.D., Elad, M., Milanfar, P.: Fast and robust multiframe super resolution. IEEE Trans. Image Process. **13**(10), 1327–1344 (2004)
16. Li, K., Yang, S., Dong, R., Wang, X., Huang, J.: Survey of single image super-resolution reconstruction. IET Image Proc. **14**(11), 2273–2290 (2020)
17. Protter, M., Elad, M., Takeda, H., Milanfar, P.: Generalizing the nonlocal-means to super-resolution reconstruction. IEEE Trans. Image Process. **18**(1), 36–51 (2008)
18. Yang, J., Wright, J., Huang, T.S., Ma, Y.: Image super-resolution via sparse representation. IEEE Trans. Image Process. **19**(11), 2861–2873 (2010)
19. Zhang, L., Wu, X.: An edge-guided image interpolation algorithm via directional filtering and data fusion. IEEE Trans. Image Process. **15**(8), 2226–2238 (2006)
20. Zhang, K., Gao, X., Tao, D., Li, X.: Single image super-resolution with non-local means and steering kernel regression. IEEE Trans. Image Process. **21**(11), 4544–4556 (2012)

21. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 136–144 (2017)
22. Ledig, C., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4681–4690 (2017)
23. Garcia-Ruiz, F., Sankaran, S., Maja, J.M., Lee, W.S., Rasmussen, J., Ehsani, R.: Comparison of two aerial imaging platforms for identification of Huanglongbing-infected citrus trees. Comput. Electron. Agric. **91**, 106–115 (2013)
24. Huang, Y., Thomson, S.J., Hoffmann, W.C., Lan, Y., Fritz, B.K.: Development and prospect of unmanned aerial vehicle technologies for agricultural production management. Int. J. Agric. Biol. Eng. **6**(3), 1–10 (2013)
25. Mohamadi, F.: U.S. Patent No. 8,880,241. Washington, DC: U.S. Patent and Trademark Office (2014)
26. Sampedro, C., Martinez, C., Chauhan, A., Campoy, P.: A supervised approach to electric tower detection and classification for power line inspection. In 2014 International Joint Conference on Neural Networks (IJCNN), pp. 1970–1977. IEEE, July 2014
27. Santos, M.J., Disney, M., Chave, J.: Detecting human presence and influence on neotropical forests with remote sensing. Remote Sens. **10**(10), 1593 (2018)