SEMANTIC SEGMENTATION OF REMOTE SENSING IMAGERY USING AN ENHANCED ENCODER-DECODER ARCHITECTURE

N. Aburaed [a,b] , M. Al-Saad [a] , M. Q. Alkhatib [a], M. S. Zitouni[a], S. Almansoori [c], and H. Al-Ahmad [a]

[a] College of Engineering and IT, University of Dubai, Dubai, UAE - (nour.aburaed, minaalsaad, mqalkhatib)@ieee.org (mzitouni, halahmad)@ud.ac.ae

[b] Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow, UK

[c] Mohammed Bin Rashid Space Centre, Dubai , UAE - saeed.almansoori@mbrsc.ae

KEY WORDS: Deep Learning, Semantic Segmentation, RUNET, UNET, Remote Sensing, Squeeze and Excitation

ABSTRACT:

Semantic segmentation is one of most the important computer vision tasks for the analysis of aerial imagery in many remote sensing applications, such as resource surveys, disaster detection, and urban planning. This area of research still faces unsolved challenges, especially in cluttered environments and complex sceneries. This study presents a repurposed Robust UNet (RUNet) architecture for semantic segmentation, and embeds the architecture with attention mechanism in order to enhance feature extraction and construction of segmentation maps. The attention mechanism is achieved using Squeeze-and-Excitation (SE) block. The resulting network is referred to as SE-RUNet. SE is also tested with the classical UNet, termed SE-UNet, to verify the efficiency of introducing SE. The proposed approach is trained and tested using "Semantic Segmentation of Aerial Imagery" dataset. The results are evaluated using Accuracy, Precision, Recall, F-score and mean Intersection over Union (mIoU) metrics. Comparative evaluation and experimental results show that using SE to embed attention mechanism into UNet and RUNet significantly improves the overall performance.

1. INTRODUCTION

The field of remote sensing has witnessed a notable surge in significance across diverse applications, including but not limited to environmental monitoring, land use and land cover classification, urban planning, and disaster management. With each passing day, the role of remote sensing becomes increasingly vital in addressing various real-world challenges and fostering sustainable development. One of the crucial aspects that underpins the successful extraction of valuable information from remote sensing data lies in the accurate and efficient segmentation of these complex images. Semantic segmentation, in particular, has emerged as a cornerstone in this endeavor, offering a sophisticated technique that enables the precise identification and classification of distinct objects and features within an image.

In the realm of remote sensing, the task of semantic segmentation can be considered as a classification problem at pixel-level (Panda and Rosenfeld, 1978). The ability to delineate and label individual pixels based on their semantic meaning unlocks a wealth of spatial information, allowing researchers and practitioners to gain unprecedented insights into the distribution, composition, and characteristics of the features present in the scenes captured by remote sensing instruments (Panda and Rosenfeld, 1978).

Traditional semantic segmentation methods, such as threshold-based segmentation (Ma et al., 2013, Aburaed et al., 2018), edge-based segmentation (Kurbatova and Laylina, 2019), and region-based segmentation (Cavallaro et al., 2016) often rely on manual feature engineering and are computationally intensive. Recently, Deep Learning (DL) techniques, such as Convolutional Neural Networks (CNNs), have shown great potential for improving the efficiency and accuracy of semantic segmentation in remote sensing (Xianyang et al., 2019, Talal et al., 2018, Al Saad et al., 2020a). Despite the significant progress made in DL-based semantic segmentation methods for remote sensing imagery, several challenges remain. One of the main challenges is the availability of labeled training data, which is essential for training DL models. Remote sensing datasets are often large and require significant resources for labeling. Additionally, the variability of remote sensing data, including differences in illumination, seasonality, and sensor characteristics, can affect the performance of semantic segmentation methods.

In this context, this study presents a new technique for semantic segmentation of remote sensing images that leverages DL techniques. Particularly, the use of attention mechanism in semantic segmentation models is explored by using Squeeze-and-Excitation (SE), which is a technique that has not been studied in the context of semantic segmentation thus far. SE has the potential to boost the performance of DL models that are popularly used for semantic segmentation, such as UNet. Additionally, the advanced version of UNet, called Robust UNet (RUNet), which was devised for Single Image Super Resolution (SISR) (Hu et al., 2019, Aburaed et al., 2022), is repurposed for semantic segmentation. Thus, SE is embedded into UNet, called SE-UNet, and embedded into RUNet, called SE-RUNet, to test its efficiency. The quantitative evaluation is performed using Accuracy, Precision, Recall, F-score, and mean Intersection over Union (mIoU).

The remainder of the paper is organized as follows: Section 2 presents the literature review of semantic segmentation based on DL approaches in the field of remote sensing, Section 3 explains the methodology and the dataset used in this research work, Section 4 illustrates and discusses the results, and finally, Section 5 summarizes and concludes the paper.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-1-W1-2023-1015-2023 | © Author(s) 2023. CC BY 4.0 License.

2. LITERATURE REVIEW

In the recent years, significant progress has been made in developing DL-based semantic segmentation methods for remote sensing imagery. DL methods have been shown to outperform traditional machine learning and statistical approaches for semantic segmentation due to their ability to learn complex and hierarchical representations from large datasets (Matrone et al., 2020). One popular DL approach for semantic segmentation is the Fully Convolutional Network (FCN) (Shelhamer et al., 2017). FCN is a deep neural network architecture that uses convolutional layers to learn feature representations and upsampling layers to generate dense pixel-wise predictions. FCN has been successfully applied to remote sensing imagery for land use and land cover classification (Mboga et al., 2020), urban object detection (Zhang and Chi, 2020), and change detection (Zhang et al., 2021).

Another DL approach for semantic segmentation is the UNet architecture. UNet is a CNN that uses an encoder-decoder structure with skip connections to preserve spatial information and reduce the loss of spatial resolution during feature extraction. UNet has been used for various remote sensing applications, including crop monitoring (Fan et al., 2022), forest mapping (Guo et al., 2021), building detection (Liu et al., 2020), road segmentation (Al Saad et al., 2020b), and oil spill detection (El Rai et al., 2020). An advanced version of UNet called RUNet was devised by (Hu et al., 2019). However, it was built for the purpose of enhancing images using Single Image Super Resolution rather than segmentation (Aburaed et al., 2022).

In addition to FCN and UNet, there are other DL-based semantic segmentation methods that have been applied to remote sensing imagery, including DeepLab (Wang et al., 2022), SegNet (Chen and Lu, 2019), and Mask R-CNN (Wu et al., 2021). DeepLab (Chen et al., 2018) uses dilated convolutions to increase the receptive field of convolutional layers and capture multiscale context information. SegNet (Badrinarayanan et al., 2017) is a variant of FCN that uses an encoder-decoder structure with max-pooling and up-sampling layers. Mask R-CNN (He et al., 2017) is a region-based DL approach that can detect and segment objects in remote sensing imagery.

Despite the significant progress made in this field of research, several challenges remain. One of the main challenges is the availability of labeled training data, which is essential for training DL models. Remote sensing datasets are often large and require significant resources for labeling. Additionally, the variability of remote sensing data, including differences in illumination, seasonality, and sensor characteristics, can affect the performance of semantic segmentation methods.

In this work, RUNet is repurposed to achieve semantic segmentation. Attention mechanism is introduced to the architecture by SE to enhance feature extraction and construction of segmentation maps. Additionally, SE is also used with the classic UNet architecture to verify its effectiveness. These methods are tested using "Semantic Segmentation of Aerial Imagery" dataset, which captures complex landscapes of Dubai city, and it is therefore an interesting case study to test the networks' robustness in a cluttered environment.

3. PROPOSED METHODOLOGY

3.1 Network Architecture

This section explains the intricate details of UNet and RUNet architectures, along with how SE is injected into their layers.

3.1.1 Encoder: Let X be the input image of size $m \times n \times c$, where m, n, and c are the height, width, and channels, respectively. Also, let Y be the ground truth segmentation map, and $\hat{Y}$ be the predicted segmentation map. Y and $\hat{Y}$ have the same height and width as X, but $c = 3$ for X, while for Y and $\hat{Y}$ the number of channels is equal to the number of classes. The network takes X as an input and passes it through an encoder network, which downsamples the feature maps and extracts hierarchical representations of the input. the features are downsampled using convolution operation, which is defined as follows:

$$F_{(x,y)} = f\left([K * X]_{(x,y)} + b\right), \tag{1}$$

where $F_{(x,y)}$ is the output feature at position (x, y), K is the filter, b is the bias, and f is the activation function. In this case, the activation function is sigmoid. The size of the filter doubles with each layer, as seen in Figure 1. In this study, UNet and RUNet have the same number of layers in the encoder. However, RUNet contains additional skip connections between the encoder layers, which are not present in UNet.

3.1.2 Squeeze-and-Excitation: The extracted features from the decoder are passed to an SE layer. For any given transformation F_{tr} maps the input feature map $\tilde{\mathbf{Y}}_{\mathbf{c}} \in \mathbb{R}^{m \times n}$ of a particular band c to the descriptor z_c. The Squeeze procedure, denoted $F_{sq}(.)$, uses global average pooling, which converts $\tilde{\mathbf{Y}}$ to a column vector of size $1 \times 1 \times c$. The squeeze function is thus defined as:

$$z_c = F_{sq(\tilde{Y}_c)} = \frac{1}{m \times n} \sum_{i=1}^{m} \sum_{j=1}^{n} \tilde{Y}_c(i, j) \tag{2}$$

The excitation procedure is used to automatically determine the significance of each feature, amplifying those that have a bigger impact on understanding the details of the image while suppressing insignificant features. The excitation function can be expressed as:

$$s = F_{ex(z,W)} = \sigma(g(z, W)) = \sigma(W_2 R(W_1 z)) \tag{3}$$

where σ is the Sigmoid activation function, R is the ReLU activation function, $W_1 \in \mathbb{R}^{\frac{c}{r} \times c}$ and $W_2 \in \mathbb{R}^{c \times \frac{c}{r}}$ are the two fully connected layers, W_1 is the dimensionality reduction layer with a dimensionality reduction ratio of r. The Sigmoid function suppresses the final output of the excitation process to a value between zero and one.

3.1.3 Decoder: The output from SE is passed through a decoder network, which upsamples the feature maps and generates a segmentation map that has the same dimensions as the input. The decoder consists of several Transpose Convolution (TC) layers. While convolution layers in the encoder reduce the input to feature maps, TC layers have the opposite effect by expanding the features as follows:

$$F_{(x,y)} = f\left([s * G]_{(x,y)} + b\right). \tag{4}$$

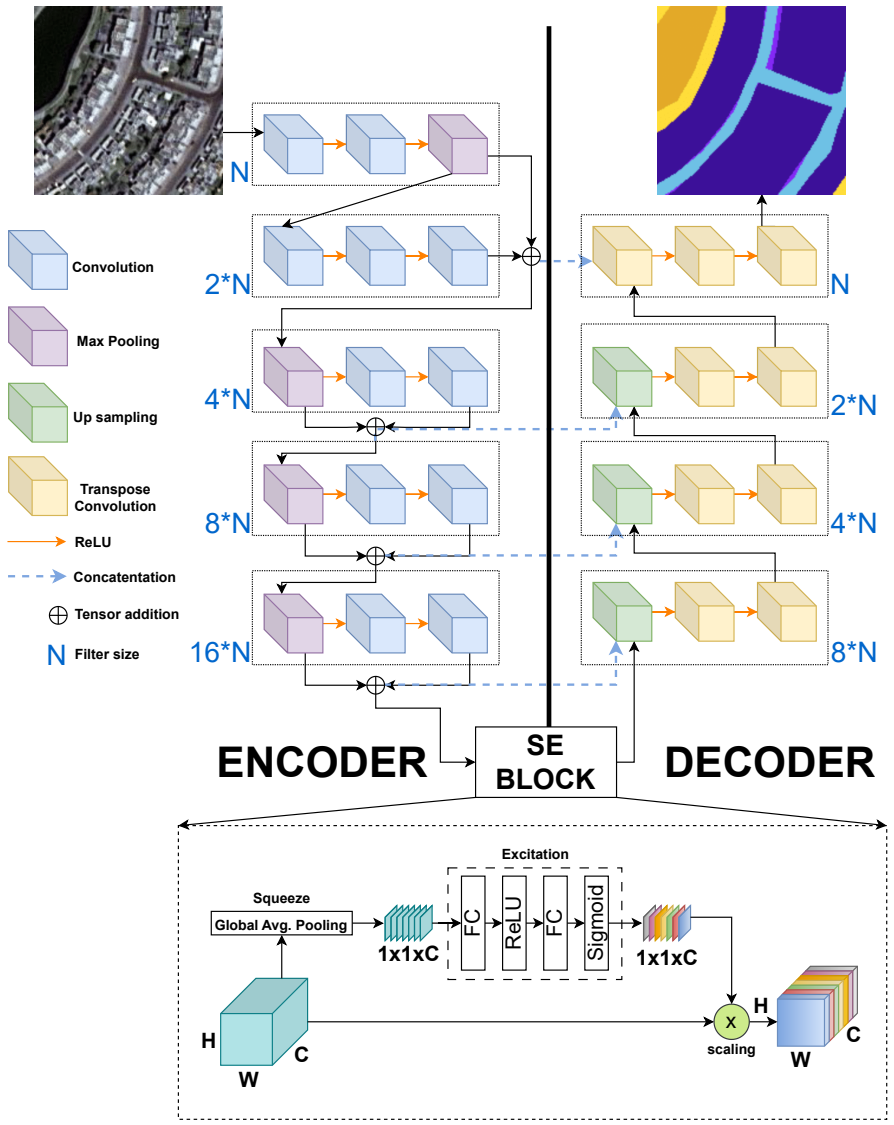


Figure 1. SE-RUNet architecture. The encoder and decoder have external and internal skip connections, and they are connected by an SE block. The number of filters N = 16 in this study. SE-UNet architecture differs by missing the tensor addition.

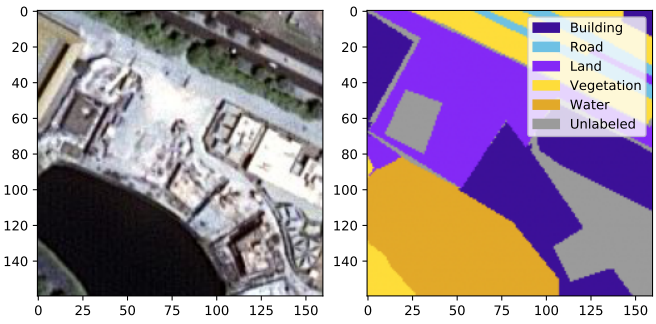
Here, the output from SE layer is convolved with a grid of a larger height and width, such that the feature map gets larger depending on the grid size as it propagates through the decoder. The decoder network also includes skip connections that concatenate the corresponding feature maps from the encoder network, allowing the decoder to make use of both low-level and high-level features for accurate segmentation. It is worth mentioning that the encoder and decoder parts are symmetrical, as seen in Figure 1. The decoder outputs the final segmentation map, which consists of 6 channels; each channel corresponds to a class map. Similar to the encoder side, both UNet and RUNet have the same number of layers in the decoder. Figure 1 shows the overall architecture of SE-RUNet. SE-UNet architecture is similar, but it is missing the tensor addition between the layers.

3.2 Dataset

The dataset used in this research work is an open access dataset called "Semantic Segmentation of Aerial Imagery" (in the Loop, 2020), which was created by Humans in the Loop as a collaborative work with the Mohammed Bin Rashid Space Centre (MBRSC). This dataset is considered significantly challenging due to shadows, occlusions, complex building structures in the captured scenery for Dubai's city. Dubai is the second largest Emirate in the UAE and one of the most rapidly developing cities in the world. This dataset consists of RGB satellite images captured by DubaiSat-2 satellite, which has a spatial resolution of 1-meter, along with their corresponding segmentation masks that represent six classes; building, water, vegetation, road, land (unpaved area), and background. It includes 72 Images of varying sizes, which are divided into patches of 160×160 in order to ease the training procedure of the proposed model. The total number of images after patching is 3483. These images are divided randomly as training, validation, and testing images with compositions of 70%, 15% and 15%, respectively. Samples of the dataset can be seen in Figure 2.

4. RESULTS AND DISCUSSION

UNet, RUNet, SE-UNet, and SE-RUNet, the four pivotal architectures examined in this research, have undergone rigorous training and testing procedures employing the meticulously



(a) (b)

Figure 2. A sample from "Semantic Segmentation of Aerial Imagery" dataset dataset that shows (a) the RGB image and its corresponding (b) segmentation mask.

curated dataset described in Section 3.2. To ensure a fair and unbiased comparison, all networks were trained and experimented within the same well-defined environment, leveraging Python's powerful Tensorflow library. The training parameters were fixed across all networks to maintain consistency, thus minimizing any potential confounding factors.

For the optimization process, the widely-used Adam algorithm served as the optimization function, coupled with categorical cross-entropy as the chosen loss function. The learning rate was empirically set at 10^{-3}, and each network was trained for 100 epochs to ensure ample opportunity for convergence and optimal model performance.

To gauge the efficacy and performance of the trained models, a set of objective quantitative evaluation metrics was employed. These metrics, namely Accuracy, Precision, Recall, F-score, and mIoU, are defined in Equations 5 to 9. Each metric offers unique insights into the model's classification performance and the accuracy of its segmentation maps.

Accuracy measures the ratio of correctly classified pixels to the total number of pixels in the dataset. Precision, on the other hand, quantifies the ability of the model to correctly classify pixels as belonging to the target class, while Recall evaluates its capacity to correctly identify all relevant pixels from the target class. F-score, which balances the trade-off between Precision and Recall, provides a holistic assessment of the model's segmentation performance.

The mIoU metric is of paramount importance, which is an invaluable indicator of the segmentation accuracy, measuring the average degree of overlap between the predicted segmentation and the ground truth across all classes. This metric becomes particularly useful when dealing with datasets that encompass multiple classes. Equation 9 incorporates the parameter N, representing the total number of classes present in the dataset. This ensures that the mIoU metric accurately captures the segmentation performance across all classes, making it an indispensable tool for comprehensive model evaluation.

In the context of these evaluation metrics, the importance of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) becomes apparent. TP represents the number of pixels that are accurately classified as belonging to the target class, while TN signifies the pixels correctly identified as not belonging to the target class. Conversely, FP denotes the pixels that are erroneously classified as belonging to the target class, and FN accounts for the pixels mistakenly identified as not belonging to the target class.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

$$Fscore = 2 \times \frac{Precision * Recall}{Precision + Recall} \tag{8}$$

$$\begin{aligned} IoU &= \frac{TP}{TP + FP + FN} \\ mIoU &= \frac{1}{N} \times \sum_{i=1}^{N} IoU_i \end{aligned} \tag{9}$$

Table 1 summarizes all the quantitative evaluation results for all networks. The findings unveil that UNet emerges as the dominant performer, surpassing RUNet across all evaluation metrics, with the exception of mIoU. This intriguing observation highlights the fact that while RUNet exhibits superiority in the domain of SISR, it does not necessarily carry that advantage into the realm of semantic segmentation.

Delving deeper into the results, SE-RUNet takes center stage, outperforming its RUNet counterpart across all evaluation metrics, with only a marginal decline in mIoU by 0.0061. This slight concession is easily outweighed by the substantial improvements achieved in other metrics, reaffirming the efficacy of the SE mechanism in enhancing the segmentation capabilities of the RUNet architecture.

SE-UNet outshines its UNet counterpart, presenting superior results in all evaluation metrics. The most pronounced enhancement is witnessed in the mIoU metric, illustrating the profound impact of the SE mechanism in elevating the segmentation performance of the classic UNet architecture.

The visual analysis, as depicted in Figure 3, echoes the quantitative findings, further bolstering the credibility of the results. SE-UNet undoubtedly exhibits the best segmentation map across

Table 1. Results summary of all models in terms of Accuracy, Precision, Recall, F-score, and mIoU. SE-UNet shows the best performance among all networks. Additionally, all SE-infused networks perform better than their counterparts without SE.

Model	Acuracy	Recall	Precision	F-score	mIoU
UNet	0.8806	0.8769	0.8854	0.8812	0.4559
RUNet	0.8668	0.8642	0.8705	0.8673	0.4842
SE-UNet	**0.8849**	**0.8819**	**0.8891**	**0.8855**	**0.4900**
SE-RUNet	0.8706	0.8685	0.8736	0.8711	0.4781

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-1-W1-2023-1015-2023 | © Author(s) 2023. CC BY 4.0 License.

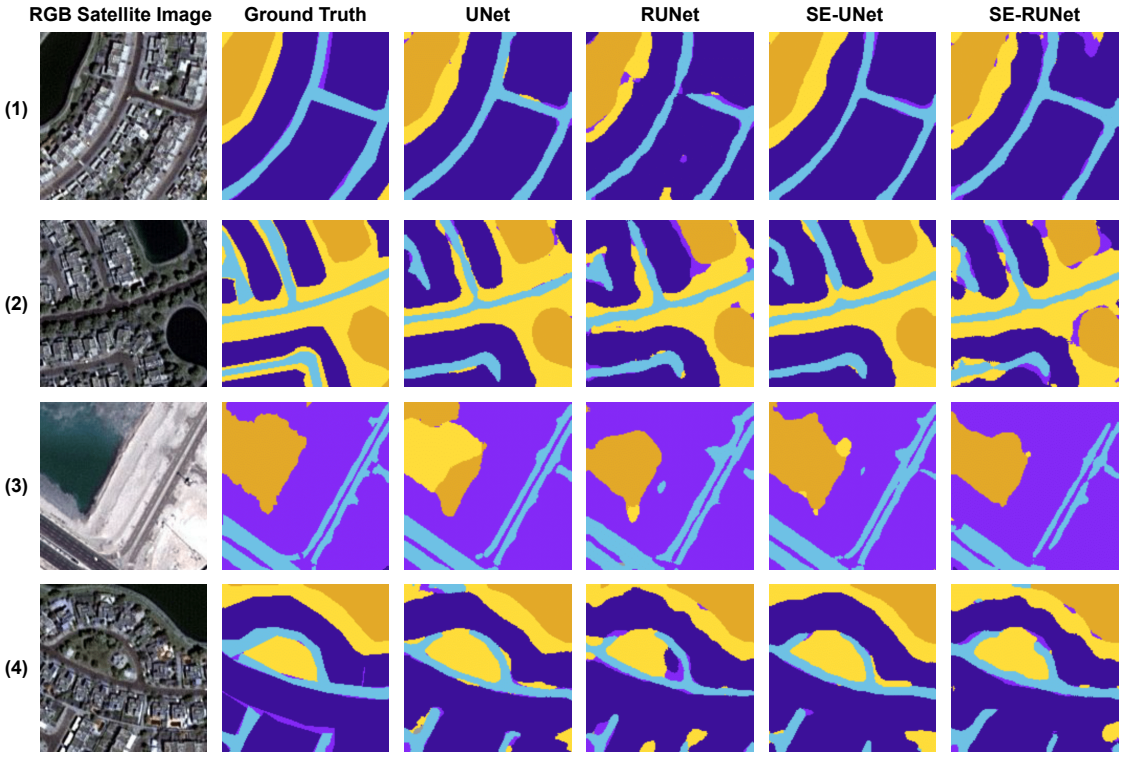


Figure 3. Visual results of segmentation maps produced by UNet, RUNet. SE-UNet, and RUNet. Both SE-UNet and SE-RUNet achieve better than their counterparts. SE-UNet performs better than SE-RUNet, with the exception of sample 3.

all four samples, demonstrating remarkable visual similarity to the ground truth despite the varied complexity of objects, shapes, and clutter in the imagery. Particularly, the second sample highlights SE-UNet's impressive resilience in accurately distinguishing objects, evident from the lack of errors around the edges of each object, in stark contrast to the other networks. This visual consistency holds true for samples 1 and 3 as well, where SE-UNet showcases superior performance compared to its counterparts. Even in the case of sample 4, featuring relatively less clutter but three distinct objects, SE-UNet shines by producing fewer errors, especially in delineating precise boundaries and shapes. The challenge posed by less apparent edges is adeptly handled by SE-UNet, reinforcing its robustness in diverse and complex scenarios.

The quantitative and visual analyses jointly reaffirm the success of the SE mechanism in elevating the performance of both the RUNet and UNet architectures. SE-UNet emerges as the undisputed leader among the networks, impressively outperforming all counterparts in both objective metrics and visual fidelity. These compelling results not only contribute to advancing the field of semantic segmentation in remote sensing but also underscore the transformative potential of attention mechanisms in deep learning architectures. As the pursuit of accurate and efficient remote sensing semantic segmentation continues, the insights gleaned from this study are poised to drive further research and innovation, paving the way for enhanced understanding and decision-making in the realm of remote sensing applications.

5. CONCLUSION

In this research, RUNet architecture that was initially utilized for SISR has been repurposed to achieve semantic segmentation task. Attention mechanism is introduced to the architecture by utilizing SE block to enhance feature extraction and construction of segmentation maps. SE is also tested with the classic UNet to further verify its effectiveness. "Semantic Segmentation of Aerial Imagery" dataset was used to test the proposed approach. Comparisons against the original UNet and and RU-Net are conducted in terms of the Accuracy, Precision, Recall, F-score, and mIoU. Experiments reveal that all SE-infused networks perform better than their counterparts with no SE. SE-UNet shows the best performance compared to all networks. Thus, SE has successfully elevated the performance of both semantic segmentation networks.

REFERENCES

Aburaed, N., Alkhatib, M. Q., Marshall, S., Zabalza, J., Ahmad, H. A., 2022. Sisr of hyperspectral remote sensing imagery using 3d encoder-decoder runet architecture. *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, 1516–1519.

Aburaed, N., Panthakkan, A., Mukhtar, H., Mansoor, W., Almansoori, S., Ahmad, H. A., 2018. Autonomous building detection using region properties and pca. *2018 International Conference on Signal Processing and Information Security (ICSPIS)*, 1–4.

Al Saad, M., Aburaed, N., Al Mansoori, S., Al Ahmad, H., Marshall, S., 2020a. Semantic segmentation of roads in high-resolution satellite imagery. *Proceedings of the International Astronautical Congress, IAC*.

Al Saad, M., Aburaed, N., Al Mansoori, S., Al Ahmad, H., Marshall, S., 2020b. Semantic segmentation of roads in high-resolution satellite imagery. *71st International Astronautical Congress (IAC)*.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12), 2481-2495.

Cavallaro, G., Dalla Mura, M., Carlinet, E., Géraud, T., Falco, N., Benediktsson, J. A., 2016. Region-based classification of remote sensing images with the morphological tree of shapes. *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, IEEE, 5087–5090.

Chen, H., Lu, S., 2019. Building extraction from remote sensing images using segnet. *2019 IEEE 4th International Conference on Image, Vision and Computing (ICIVC)*, 227–230.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L., 2018. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4), 834-848.

El Rai, M. C., Aburaed, N., Al-Saad, M., Al-Ahmad, H., Al Mansoori, S., Marshall, S., 2020. Integrating deep learning with active contour models in remote sensing image segmentation. *2020 27th IEEE International Conference on Electronics, Circuits and Systems (ICECS)*, 1–4.

Fan, X., Yan, C., Fan, J., Wang, N., 2022. Improved U-Net Remote Sensing Classification Algorithm Fusing Attention and Multiscale Features. *Remote Sensing*, 14(15). https://www.mdpi.com/2072-4292/14/15/3591.

Guo, Y., Li, Z., Chen, E., Zhang, X., Zhao, L., Xu, E., Hou, Y., Liu, L., 2021. A Deep Fusion uNet for Mapping Forests at Tree Species Levels with Multi-Temporal High Spatial Resolution Satellite Imagery. *Remote Sensing*, 13(18). https://www.mdpi.com/2072-4292/13/18/3613.

He, K., Gkioxari, G., Dollar, P., Girshick, R., 2017. Mask r-cnn. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Hu, X., Naiel, M. A., Wong, A., Lamm, M., Fieguth, P., 2019. Runet: A robust unet architecture for image super-resolution. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 505–507.

in the Loop, H., 2020. Semantic segmentation of aerial imagery. `https://www.kaggle.com/datasets/humansintheloop/semantic-segmentation-of-aerial-imagery`. accessed: April 27, 2023.

Kurbatova, E., Laylina, V., 2019. Detection of roads from images based on edge segmentation and morphological operations. *2019 8th Mediterranean Conference on Embedded Computing (MECO)*, IEEE, 1–4.

Liu, Z., Chen, B., Zhang, A., 2020. Building segmentation from satellite imagery using u-net with resnet encoder. *2020 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE)*, 1967–1971.

Ma, H., Cheng, X., Wang, X., Yuan, J., 2013. Road information extraction from high resolution remote sensing images based on threshold segmentation and mathematical morphology. *2013 6th International Congress on Image and Signal Processing (CISP)*, 2, IEEE, 626–630.

Matrone, F., Grilli, E., Martini, M., Paolanti, M., Pierdicca, R., Remondino, F., 2020. Comparing Machine and Deep Learning Methods for Large 3D Heritage Semantic Segmentation. *ISPRS International Journal of Geo-Information*, 9(9). https://www.mdpi.com/2220-9964/9/9/535.

Mboga, N., Grippa, T., Georganos, S., Vanhuysse, S., Smets, B., Dewitte, O., Wolff, E., Lennert, M., 2020. Fully convolutional networks for land cover classification from historical panchromatic aerial photographs. *ISPRS Journal of Photogrammetry and Remote Sensing*, 167, 385-395. https://www.sciencedirect.com/science/article/pii/S0924271620301921.

Panda, D. P., Rosenfeld, A., 1978. Image segmentation by pixel classification in (gray level, edge value) space. *IEEE transactions on computers*, 27(09), 875–879.

Shelhamer, E., Long, J., Darrell, T., 2017. Fully Convolutional Networks for Semantic Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4), 640-651.

Talal, M., Panthakkan, A., Mukhtar, H., Mansoor, W., Almansoori, S., Al Ahmad, H., 2018. Detection of water-bodies using semantic segmentation. *2018 International Conference on Signal Processing and Information Security (ICSPIS)*, IEEE, 1–4.

Wang, Z., Fan, B., Tu, Z., Li, H., Chen, D., 2022. Cloud and Snow Identification Based on DeepLab V3+ and CRF Combined Model for GF-1 WFV Images. *Remote Sensing*, 14(19). https://www.mdpi.com/2072-4292/14/19/4880.

Wu, Q., Feng, D., Cao, C., Zeng, X., Feng, Z., Wu, J., Huang, Z., 2021. Improved Mask R-CNN for Aircraft Detection in Remote Sensing Images. *Sensors*, 21(8). https://www.mdpi.com/1424-8220/21/8/2618.

Xianyang, N., Yinbao, C., Zhongyu, W., 2019. Remote sensing semantic segmentation with convolution neural network using attention mechanism. *2019 14th IEEE International Conference on Electronic Measurement & Instruments (ICEMI)*, IEEE, 608–613.

Zhang, H., Tang, X., Han, X., Ma, J., Zhang, X., Jiao, L., 2021. High-resolution remote sensing images change detection with siamese holistically-guided fcn. *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, 4340–4343.

Zhang, Y., Chi, M., 2020. Mask-R-FCN: A Deep Fusion Network for Semantic Segmentation. *IEEE Access*, 8, 155753-155765.