COMPARISON OF DEEP LEARNING ARCHITECTURES FOR THE SEMANTIC SEGMENTATION OF SLUM AREAS FROM SATELLITE IMAGES

Y.A. Lumban-Gaol[a, *], A. Rizaldy[b], A. Murtiyoso[c]

[a] Research Organization for Earth Sciences and Maritime, National Research and Innovation Agency, Indonesia - yust012@brin.go.id
[b] Helmholtz Institute Freiberg for Resource Technology, Helmholtz-Zentrum Dresden-Rossendorf, Germany - a.rizaldy@hzdr.de
[c] Forest Resources Management Group, Institute of Terrestrial Ecosystems, Department of Environmental Systems Science,
ETH Zurich, Switzerland -arnadi.murtiyoso@usys.ethz.ch

Commission III WG III/5

KEY WORDS: deep learning, semantic segmentation, slums, remote sensing

ABSTRACT:

The mapping of slum areas is an important task when considering the necessity for an inclusive, safe and resilient cities. While many methods exist in this regard, the use of machine learning and more specifically deep learning has gained traction in recent years. In this paper, we present a systematic comparison of existing deep learning architectures and backbones. The experiments in the paper investigate the question of which architecture and backbone combination and which configuration of dataset preparation is best for use in slum mapping. In another experiment we implemented the trained model to predict slums in existing open data. The experiments in the paper used public open data provided by Helber et al. (2018). Results show that `FPN` with `vgg16` backbone showed the most potential in this particular application. The results of the semantic segmentation also shows promise, although the discrepancy in slum characteristic still hinders a proper generalization of its use.

1. INTRODUCTION

The 11th Sustainable Development Goal (SDG) calls for sustainable cities and inclusive, safe, and resilient settlements by 2030. The proportion of the urban population living in slums, informal settlements, or inadequate housing is one of the indicators for this agenda. According to UN-Habitat (2022), 20% of the world's population, approximately around 1.6 billion people, live in substandard housing, with one billion living in slums and informal settlements. The Millennium Development Goals (MDGs) defined a slum as an area that combines several characteristics such as a lack of access to clean water and sanitation, as well as poor housing quality, overcrowding, and insecure residential status to varying degrees (UN-Habitat, 2018, 2007). The ability to locate and monitor these areas would support both governmental and non-governmental organisations in taking necessary decisions in reducing slums.

Kuffer et al. (2016) demonstrated the possibility of using very-high-resolution (VHR) remote sensing data to map slums over large areas in a repeatable manner. Several other studies employed machine learning approach, such as Random Forest (RF) (de Mattos et al., 2021; Owusu et al., 2021) and Support Vector Machine (SVM) (Duque et al., 2017; Prabhu and Parvathavarthini, 2022). These methods requires preliminary feature extraction that describes the physical, morphological, and contextual characteristics of slum areas in order to recognise its pattern both geometrically and spatially. Convolutional Neural Networks (CNNs) on the other hand, have the ability to learn high-level spatial features automatically.

The use of AI (artificial intelligence) in classifying satellite images has seen a recent surge in interest. In AI parlance, the act of classifying pixels is analogous to semantic segmentation (Murtiyoso et al., 2022), although it may also involve instance segmentation and eventually panoptic segmentation (Kirillov et al., 2019). Slum area classification using deep learning approach had been conducted by Mboga et al. (2017); Persello and Stein (2017); Gram-Hansen et al. (2019); Liu et al. (2019); Fisher et al. (2022). Mboga et al. (2017) shows that a pixel-wise CNN technique for slum classification in Tanzania resulted in higher accuracy than SVM using texture features. Persello and Stein (2017) assessed deep fully convolutional networks (FCN) to detect informal settlements over the same area as Mboga et al. (2017) and pointed out that FCN perform better than conventional convolutional networks. Liu et al. (2019) also used FCN to study slum mapping, but from a temporal dynamics perspective on temporary slums in Bangalore, India. Gram-Hansen et al. (2019) provides high-resolution images and annotation data pairs over slum areas in different countries and trains the data with canonical correlation forests and the `DeepLabv3+` model. Fisher et al. (2022) used Sentinel-2 images in Mumbai, India, to map slum areas with uncertainty quantification at the pixel level by incorporating a Monte Carlo dropout in the `U-Net` model.

In this study, we used publicly open data provided by Helber et al. (2018) and investigated the performance of three different deep learning architectures for semantic segmentation of slum areas from VHR satellite images, including U-Net (Ronneberger et al., 2015), FPN (Lin et al., 2017), and Linknet (Chaurasia and Culurciello, 2018). Performance will be assessed based on several parameters, including the intersection-over-union value and overall elapsed time.

2. MATERIALS AND METHODS

2.1 Data

Helber et al. (2018) provided several image-target pairs from VHR satellite data for informal settlements in Medellin (Colombia), Kibera (Kenya), Makoko (Nigeria), and El Daein and El Geneina (Sudan). The data pairs include three natural spectral Red-Green-Blue (RGB) images with binary annotation images of slum and

*Corresponding author.

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-1-W2-2023-1439-2023 | © Author(s) 2023. CC BY 4.0 License.

non-slum pixels. The VHR images were provided by DigitalGlobe through the Satellite Applications Catapult with a 30-50 cm spatial resolution (Gram-Hansen et al., 2019). This paper only used datasets in Kibera (Figure 1a) and Makoko (Figure 1b) due to several justifiable reasons. Firstly, the area coverage in Medellin is too small thus provided insufficient training data for our purposes. Secondly, although El Daein and El Geneina cover larger informal settlements than others, their ground truth data proved to be unreliable when setting up the initial experiments. A notable problem observed is the existence of similar spectral objects that appear in different classes.

2.2 Method

This study used three different semantic segmentation architectures, including `U-Net` (Ronneberger et al., 2015), `FPN` (Lin et al., 2017), and `Linknet` (Chaurasia and Culurciello, 2018). The name `U-Net` comes from its characteristic U-shaped architecture which consists of an encoder that captures features at different scales and a decoder that restores the spatial resolution using skip connections. `FPN`, Feature Pyramid Network, addresses the issue of handling objects at different scales by creating a feature pyramid. It enables the network to capture both fine-grained details and high-level semantic information. `Linknet` is a lightweight and efficient architecture that utilizes an encoder-decoder structure with skip connections to capture both local and global context. These skip connections preserve fine-grained details and enable efficient information propagation.

These networks are arguably the three most popular models for semantic segmentation based on literature review. Although initially developed for computer vision and biomedical tasks, these networks have been used lately for remote sensing-related work, such as detecting generic land cover classes, e.g. road, building, vegetation, and water classes. Generally, these networks are similar by having the encoder-decoder mechanism. The differences are in the small details when the networks link the encoder and the decoder for the upsampling purpose to result in high-resolution prediction. In this paper, we adopted the `segmentation_models` library developed by Iakubovskii (2019) to run those networks combined with 32 different backbones. The codes implemented in this paper are available in `https://github.com/yustisiardhitasari/slum_orei` (last accessed 3 July 2023).

The experiments were conducted in three consecutive stages:

Experiment 1 Aims to assess the model performances to select a network-backbone pair which will be used for the next experiment.

Experiment 2 The second stage focuses on investigating model performances with different configurations.

Experiment 3 Finally, a new location is introduced to evaluate the stability of semantic segmentation networks to classify slum pixels in different locations.

The initial experiment used the Kibera data set (Figure 1a). The data preparation is described in Figure 2. Patches were generated by cropping the image into 512 x 512 pixel crops after resampling it into a 1 m resolution to avoid homogeneous classes in the individual patches. In the end, 189 patches were thus generated and used to train the networks. They were then split into 80% training and 20% validation sets and trained with 50 epochs and a batch size of four. We trained this data using three networks with 32 different backbones, resulting in 96 training models. The computational process was conducted using a workstation equipped by an NVIDIA A100 GPU with 40GB VRAM.

The experiment in the second stage still used the Kibera data set but employed different configurations of the resolution, patch size, and overlapping pixels in the patches. The objective is to assess the effect of area coverage per patch and the number of training data to the model performance. Meanwhile, the last experiment applied Kibera and Makoko data which were prepared using the same configurations. The second and third experiments were conducted using a laptop with an NVIDIA GeForce RTX 3060 6GB GPU. All experiments were analyzed quantitatively based on the IoU values from each training model. In addition, the computational efficiency in the initial experiment was also determined based on the duration of each training.

3. RESULTS AND DISCUSSION

3.1 Experiment 1: Model performances

Preliminary observations depicted in Figure 3(a) suggest that the pretrained `senet154` backbone was able to achieve comparable results across networks, with IoU scores of 80.84%, 80.49%, and 80.45% for `FPN`, `Linknet`, and `U-Net`, respectively. The optimum IoU score was generated when using the `seresnet152` backbone on `FPN`, with an IoU score of 80.85%, similar to the `senet154` backbone but with a faster training process. Meanwhile, the same backbone resulted in IoU scores of 80.41% on `FPN` and 80.18% on `U-Net`. However, compared to the `seresnet152` backbone, the training process using `senet154` took roughly six times longer in all networks. Note that a more detailed numerical result can be found in Figure 6 in the appendix section.

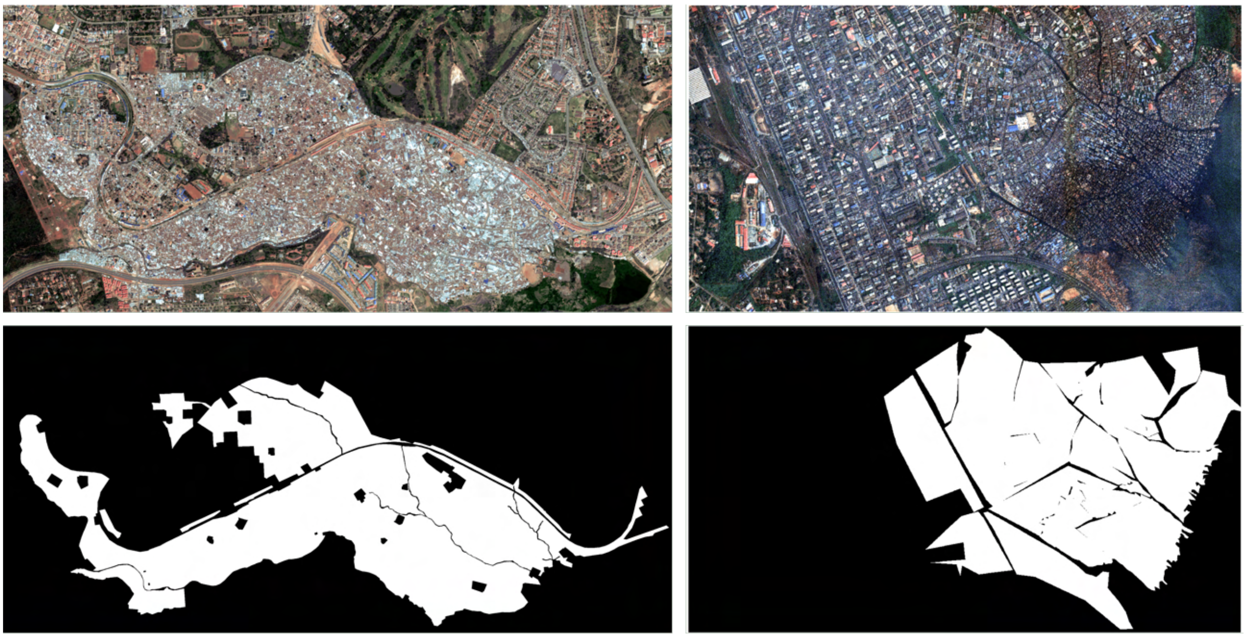
Overall, using the `FPN` architecture yielded an average IoU score of 80.14%, with `Linknet` giving 77.40% and `U-Net` 79.35%. Based on this performance, `FPN` was chosen as the deep learning model to be investigated further for slum mapping. Figure 3(b) shows a graph of `FPN` IoU results plotted against training duration for the different backbones. Based on this figure, `vgg16` scored a good balance between IoU (third best) and training duration, even when compared to `seresnet152`. The combination of `FPN` and `vgg16` was therefore chosen to be investigated further in Experiment 2.

3.2 Experiment 2: Effect of training configurations

Based on the IoU and computational time during training in the initial experiments and considering the processing unit capacity, the `FPN` network with the `vgg16` backbone was selected to reprocess the data with different configurations. Table 1 shows the different set-ups over the same dataset. The first row of the table describes the configurations and the IoU value of the initial experiment. Using the 1 m resolution data and a patch size of 512 with many overlapping pixels in the patches produced IoU of about 80%. However, using the 0.5 m resolution images with a reduced percentage of overlapping pixels shows that the IoU decreases dramatically to below 50%. A patch's area coverage is likely to affect model performance since the probability of having a homogeneous class in a single patch is higher when using the 0.5 resolution data than the 1 m, thus hinting at the averse effect of oversampling the original image.

Nevertheless, to understand the effect of training sets, the patch size was further reduced to 256 pixels. Consequently, the same issue rose. This is probably due to the probability of having a homogeneous class inside one patch being still higher whether

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-1-W2-2023-1439-2023 | © Author(s) 2023. CC BY 4.0 License.



(a) (b)

Figure 1: Original dataset of VHR image and slum annotation in Kibera, Kenya with a 30 cm resolution (a) and in Makoko, Nigeria with a 50 cm resolution (b). Source: https://frontierdevelopmentlab.github.io/informal-settlements/, last accessed 3 July 2023.

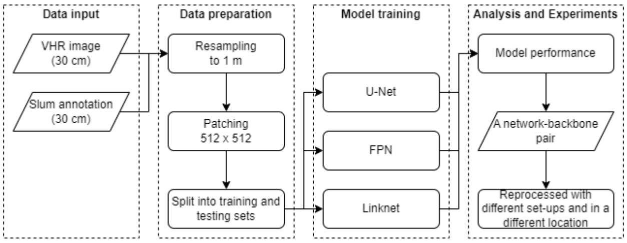


Figure 2: Overview of the methodology conducted in this paper.

Table 1: IoU metric of experimental results using different configurations on the Kibera dataset.

Resolution (m)	Patch size (pixel)	Overlap (%)	Train/Val	IoU (%)
1	512	75	151/38	80.83
1	512	50	60/15	81.11
0.5	512	12.5	86/22	48.82
1	256	50	217/69	54.25
0.5	256	50	927/290	47.44
0.5	256	12.5	374/94	44.27

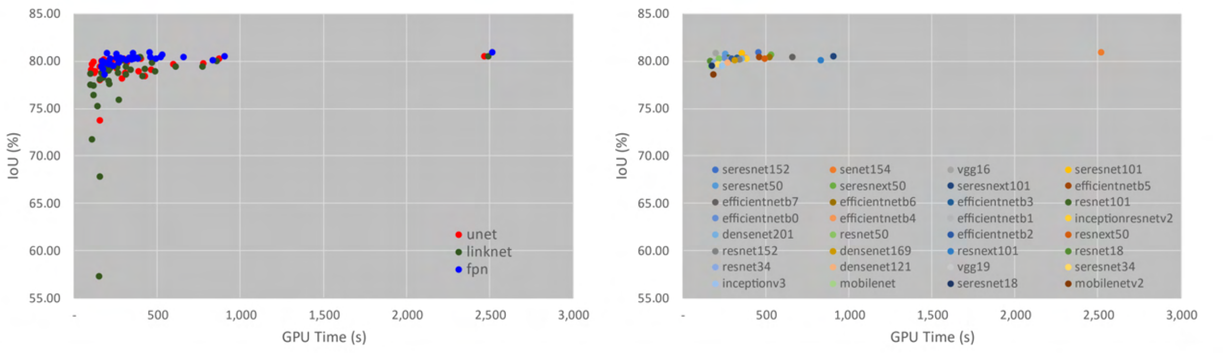
using 256 or 512 pixels, especially considering that the 256 patch size was combined with the lowering of the resolution into 0.5 m. The experimental results using a patch size of 256 with different resolutions and overlapping percentages generated more training data, but the results indicate low IoU values of between 44% and 54%. The 1 m resolution data has better IoU values than the 0.5 resolution image, even though the smaller resolution has training sets three times more than the 1 m resolution with the same number of overlapping pixels. This observation shows that the quality of the training data is as important as the amount fed during the training. In this particular case, a larger patch is recommended to ensure that each patch represents enough diversity of classes.

To better understand the number of training sets' effect on the model performance, we used a second dataset in Makoko, Nigeria. We created two training data sets using the 1 m resolution image of Makoko with a patch size of 256 and 50% overlapping pixels. The first set consists of 136/35 train/val numbers, while the second has smaller sets, which are 108/35. The larger sets resulted in an IoU of 57.06%, approximately 20% higher than the smaller ones, which only yielded an IoU of 38.61%. These results suggest that more training data can increase the model performance.

3.3 Experiment 3: Semantic segmentation for slum mapping

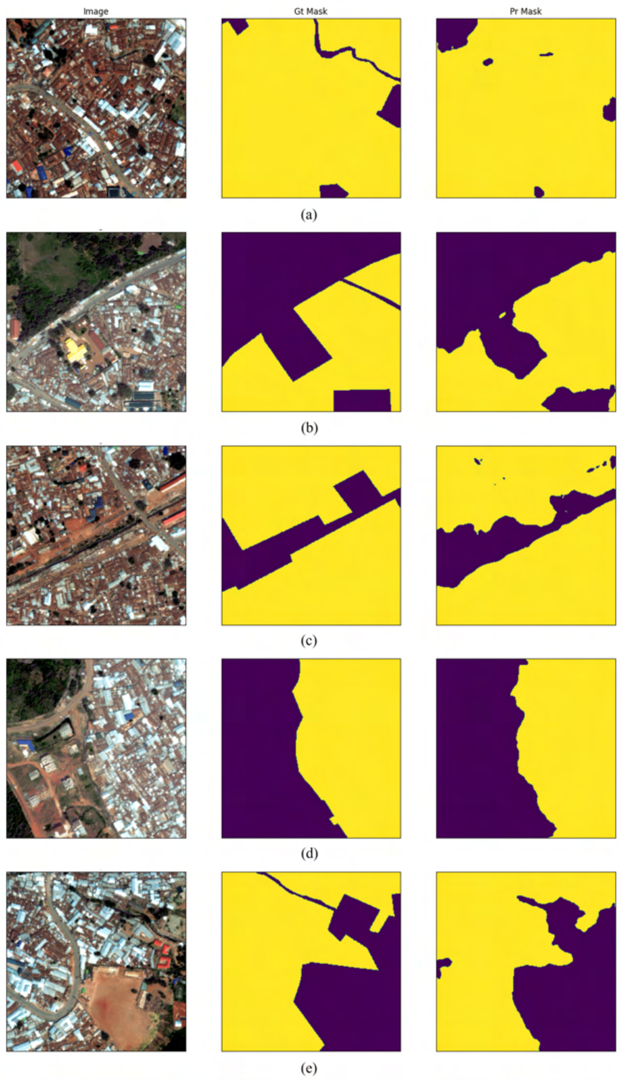
To bring the semantic segmentation model performance into the context of slum mapping, the model was tested on patches not included in the training process, therefore presenting an independent check. The 1 m resolution data with 50% overlapping pixels and a patch size of 256 was used since it produces more datasets for splitting. Using this configuration, 55 patches to test on the Kibera site and 28 on Makoko were established. The computed IoU values when testing using the pretrained model in each area are 64.52 and 51.80 for Kibera and Makoko, respectively.

Figure 4 and Figure 5 illustrate results of the semantic segmentation on the test data. Although the Kibera pretrained model was able to detect vegetation and open bare as non-slums quite well (Figure 4b, Figure 4d, and Figure 4e), its capability to separate slums and other built up objects is still limited (cf. Figure 4a, Figure 4c, and Figure 4e). The RGB images also show that some built-up objects, such as roads, are classified inconsistently. For example, roads in Figure 4b and Figure 4d annotated as non-slum pixels while in Figure 4a and Figure 4e labeled as slums. This inconsistency comes invariably from the the model training process, considering the results depend on the input for training. Note also that the ground truth data was included for validation during training.



(a) (b)

Figure 3: IoU vs training duration colored by architectures (a) and backbones for the `FPN` architecture (b).



(a)

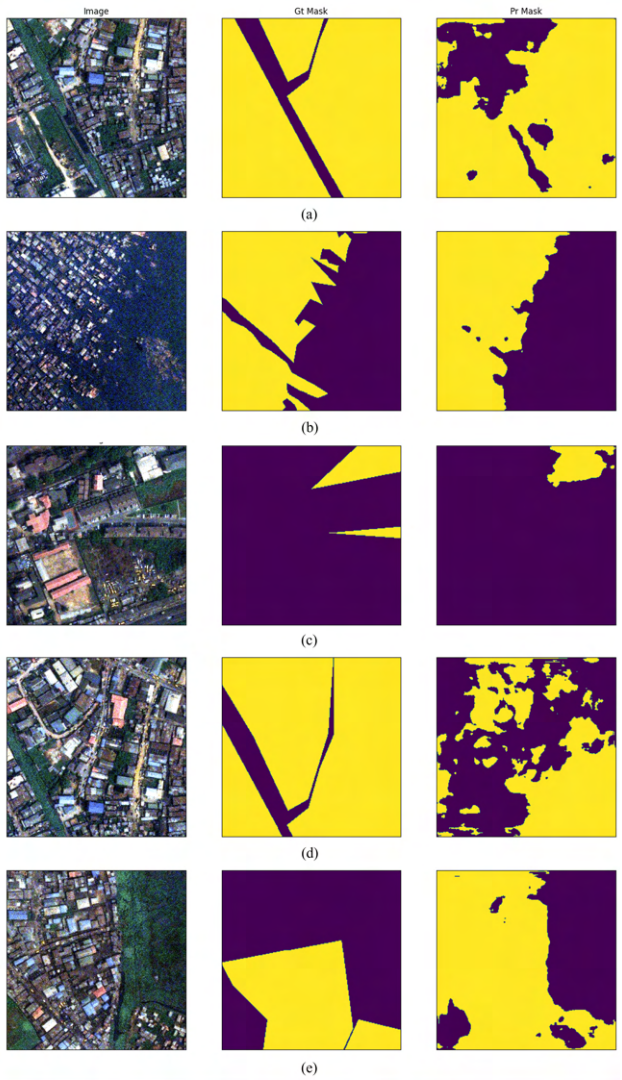
(b)

(c)

(d)

(e)

Figure 4: Results of the prediction for the slum class on the Kibera test dataset.



(a)

(b)

(c)

(d)

(e)

Figure 5: Results of the prediction for the slum class on the Makoko test dataset.

Makoko data also shows similar issues. Figure 5a and Figure 5d show some vegetated area labeled as slum pixels. Figure 5b presents buildings in a coastal area where coastal water pixels are labeled as slums. Moreover, Figure 5c imply that there is an error in the labeling process, while Figure 5e suggests that there should be more slum pixels in the ground truth data, so the prediction is not completely misclassified but was nevertheless labeled as misclassified pixels.

Furthermore, the trained model tends to work only in the area of training. Using the pretrained model in Kibera to do semantic segmentation in Makoko, or vice versa, does not work. The pretrained model on one failed to predict slum pixels in the other area. This illustrates another important challenge on the use of deep learning for slum mapping: the transfer model and model generalization are often hampered by different characteristics of slums in different areas, which depends strongly on both cultural and geographical factors. This in turn showcases the need for a larger and more inclusive training datasets encompassing different characteristics from different parts of the world, which is an immense challenge in its own right.

4. CONCLUSIONS AND FURTHER WORK

This work compared three deep learning architectures for semantic segmentation of slum areas from VHR satellite images: `U-Net`, `FPN` and `Linknet`. Using a dataset from Kibera, Kenya for training we found that `FPN` networks with different backbones showed better IoU scores than others, with backbones having shallow layers resulted in shorter training duration than others having deep layers. The initial experiments demonstrate that the model can reach up to 80% IoU score. However, using a smaller patch size or resolution decreased the IoU, indicating that limited coverage of patches caused more homogeneous classes in patches, which in turn affected training results. This observation was empirically supported by the results of the second experiment.

In the third experiment, semantic segmentation was performed on independent test datasets. The achieved IoU scores of 64.53% (Kibera) and 51.80% (Makoko) were supported by visual inspection; indeed the distinction between certain classes with strong resemblance proved difficult. This is, however, a preidentified challenge when working with slum areas.

Despite the limitations of the current open data set, creating a data set for the slum is not an easy task, especially to create a pixel-wise annotation. This is more so when considering the changes in slum characteristics depending on cultural and geographical factors. However, the ever increasing availability of open data may prove to be the key in implementing this promising solution for slum mapping.

ACKNOWLEDGEMENTS

This paper has received funding from the Artificial Intelligence, Big Data, and Computation Technology for Biodiversity and Remote Sensing Research Grant Program of the National Research and Innovation Agency of the Republic of Indonesia DIPA-124.01. 1.690500/2023 and the Indonesian Endowment Fund for Education (LPDP), Republic of Indonesia.

References

Chaurasia, A. and Culurciello, E., 2018. Linknet: Exploiting encoder representations for efficient semantic segmentation. Vol. 2018-January.

de Mattos, A. C. H., McArdle, G. and Bertolotto, M., 2021. Mapping slums with medium resolution satellite imagery: a comparative analysis of multi-spectral data and grey-level co-occurrence matrix techniques.

Duque, J. C., Patino, J. E. and Betancourt, A., 2017. Exploring the potential of machine learning for automatic slum identification from vhr imagery. Remote Sensing.

Fisher, T., Gibson, H., Liu, Y., Abdar, M., Posa, M., Salimi-Khorshidi, G., Hassaine, A., Cai, Y., Rahimi, K. and Mamouei, M., 2022. Uncertainty-aware interpretable deep learning for slum mapping and monitoring. Remote Sensing.

Gram-Hansen, B. J., Azam, F., Helber, P., Coca-Castro, A., Bilinski, P., Varatharajan, I. and Kopackova, V., 2019. Mapping informal settlements in developing countries using machine learning and low resolution multi-spectral data.

Helber, P., Gram-Hansen, B., Varatharajan, I., Azam, F., Coca-Castro, A., Kopackova, V. and Bilinski, P., 2018. Mapping informal settlements in developing countries using machine learning with noisy annotations and multi-resolution multi-spectral data.

Iakubovskii, P., 2019. Segmentation models. `https://github.com/qubvel/segmentation_models`.

Kirillov, A., He, K., Girshick, R., Rother, C. and Dollar, P., 2019. Panoptic segmentation. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition pp. 9396–9405.

Kuffer, M., Pfeffer, K. and Sliuzas, R., 2016. Slums from space-15 years of slum mapping using remote sensing.

Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B. and Belongie, S., 2017. Feature pyramid networks for object detection. Vol. 2017-January.

Liu, R., Kuffer, M. and Persello, C., 2019. The temporal dynamics of slums employing a cnn-based change detection approach. Remote Sensing.

Mboga, N., Persello, C., Bergado, J. R. and Stein, A., 2017. Detection of informal settlements from vhr images using convolutional neural networks. Remote Sensing.

Murtiyoso, A., Pellis, E., Grussenmeyer, P., Landes, T. and Masiero, A., 2022. Towards semantic photogrammetry: Generating semantically rich point clouds from architectural close-range photogrammetry. Sensors 22, pp. 966.

Owusu, M., Kuffer, M., Belgiu, M., Grippa, T., Lennert, M., Georganos, S. and Vanhuysse, S., 2021. Towards user-driven earth observation-based slum mapping. Computers, Environment and Urban Systems.

Persello, C. and Stein, A., 2017. Deep fully convolutional networks for the detection of informal settlements in vhr images. IEEE Geoscience and Remote Sensing Letters.

Prabhu, R. and Parvathavarthini, B., 2022. Morphological slum index for slum extraction from high-resolution remote sensing imagery over urban areas. Geocarto International 0, pp. 1–19.

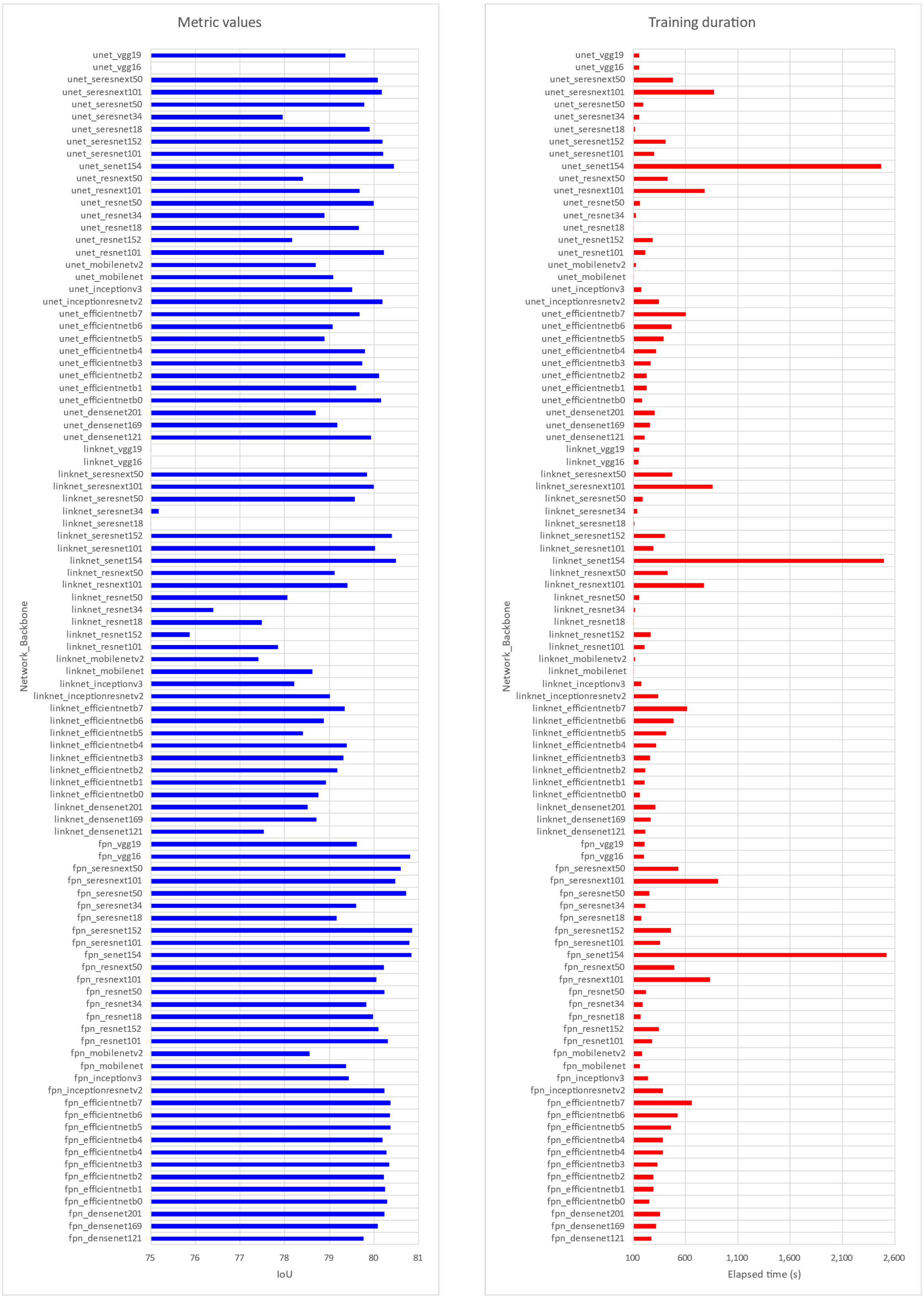
Ronneberger, O., Fischer, P. and Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 9351, pp. 234–241.

UN-Habitat, 2007. What are slums.

UN-Habitat, 2018. Sdg indicator 11.1.1 training module: Adequate housing and slum upgrading.

UN-Habitat, 2022. Envisaging the Future of Cities. Technical report.

APPENDIX



(a) (b)

Figure 6: Model performances based on IoU metric and computational time during training from the initial experimental results on the Kibera dataset.