Semantic Segmentation from Drone Images using efficient-Unet like architecture

Md. Abrar Zahin (Student ID: 0422062540)
Department of Electrical & Electronic Engineering (Bangladesh University of Engineering and Technology), Dhaka, Bangladesh

Abstract—**Due to advancements in deep learning techniques, technological progress, and the increased availability of datasets, significant strides have been made in computer vision. Researchers are currently dedicating considerable efforts to object recognition and image segmentation. Nowadays, drones are gaining popularity in applications such as land surveying, aerial monitoring, search and rescue operations, and surveillance. These tasks generate a substantial number of images, necessitating extensive processing for object detection and segmentation.**

In the quest to develop a robust, effective, secure, and safe autonomous drone control system capable of understanding and interacting with its surroundings, the initial focus is on comprehending nearby environments and obstacles. Modern drones typically employ various obstacle avoidance sensors, such as ultrasonic, infrared sensors, and a 360-degree camera facing inward. However, these sensors pose challenges in terms of complexity in construction, maintenance, and troubleshooting.

In our project, we aimed to perform object segmentation on images sourced from the AI Crowd Drone Image Segmentation Challenge. Despite encountering several setbacks, we ultimately succeeded in segmenting objects using an architecture resembling UNet, where Efficientnetb0 is used as the encoder for optimal performance. Both the AI Crowd Challenge organizers and ourselves chose the mean intersection over union (mIOU) metric for performance evaluation. The mIOU score set for the competition was 0.61. While we endeavored to surpass the baseline, we achieved an mIOU score of 0.51 using the EfficientUNet-like architecture.

Index Terms—**Semantic Segmentation, Deep Learning, UNet, EfficientNet, EfficientUNet**

I. Literature Review

Segmentation plays crucial roles in satellite imaging, medical imaging, autonomous driving system, as in this segmentation performs localization as well as identification of the important key features or objects. At first segmentation was performed by using clustering-based algorithms with edges and contours related information. In the segmentation of satellite images, the clustering was performed based on wavelengths. Here similar pixels were identified and a cluster is formed nearby its location [1]. Resnet based encoder module was used in UNetFormer [2] where the decoder was used by using transformer framework to get details information about the models and semantic levels. Ye George et al. [3] proposed multi-scale dilation net for urban scene segmentation. To solve cross-domain classification problem of satellite images transfer learning based steerable filters were utilized by Yeung et al [4] Segmenting inundated areas in UAV pictures by distinguishing water from structures, foliage, and roadways was performed by Gebrehiwot et al. [5]. The inundated regions and plants in UAV images were separated using a decision fusion-based approach. Using a fusion of leftover U-Net modules, Zhang et al. [6] were able to separate vegetation in UAV pictures. A segmentation method was developed by Dutta et al. [6] where fusion was used with channel based attention for finding river ice. Kerfoot et al. [7] used a U-Net a CNN network based architecture developed using residual blocks for left ventricle segmentation.After that UNet++ was introduced by Zhou et al [8], where densed skip connections and nested layers were introduced to reduce the encoder decoder semantic gaps. Considering the full-scale feature information in mind Huang et al [9] proposed a concept of aggregation of feature map and suggested UNet3+ architecture.

After reviewing the segmentation related research works, UNet like architecture was very popular in the image segmentation specially medical image segmentation but as our main focus was to use the concept in aerial imagery, so we used hybrid UNet like architecture also known as EfficientUNet.

II. Dataset

Our dataset contains two folders. One containing all the input images and another one contained all the masks annotated for each input image. It had 1786 images & 1786 masks. The size of each image was 1550px x 2200px. It contained total 17 classes including the background. They are Water, Asphalt, Grass, Human, Animal, High_vegetation, Ground_vehicle, Facade, Wire, Garden_furniture, Concrete, Roof, Gravel, Soil, Primeair_pattern, Snow, Background. All the input images were gray-scale images.

III. Model and Architecture

We choose Eff-UNet concept from our reference paper [10]. Where this paper used EfficientNetB7 as the encoder and decoder of the network.The encoder relies on a series of convolutional layers to lower the spatial resolution of the input image and simultaneously enhance the number of channels to recognize more complex properties. The process of downsampling enables the neural network to extract high-level features that effectively capture global context and spatial relationships among distinct regions of the image. The decoder employs a sequence of up sampling layers to recover the spatial resolution of the feature maps and produce the ultimate segmentation map. Skip connections are often used to establish connections between corresponding layers in both the encoder and decoder (figure-6). This enables the decoder to locate

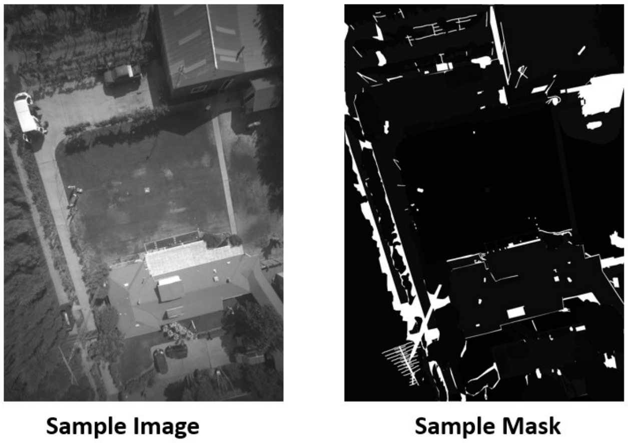


Fig. 1: Sample image and Mask

features at multiple resolutions and scales. UNet is the well suited architecture to use encoder-decoder architecture in CNN so we chose UNet like architecture for our CNN model. Where the other networks rely heavily on increasing the depth for performance. EfficientNet performs well due its Compound Scaling and Mobile Inverted Residual Bottleneck (MBConv) blocks [10]. EfficientNet simultaneously increase resolution, depth and width of the network, such compound scaling plays a big role in achieving higher accuracy with very few parameters and training samples comparing to other models. MBConv block is the fundamental block of EfficientNet. It helps to reduce computational costs while increasing the model capacity simultaneously. Thus it aids to achieve higher accuracy with less data considering other models.The paper suggested to divide efficientnetb7 into seven blocks (figure-6) based on number of channels, stride and filter size.

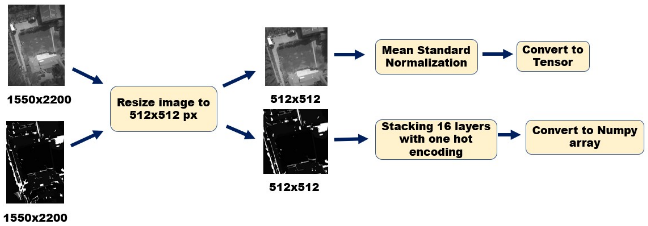
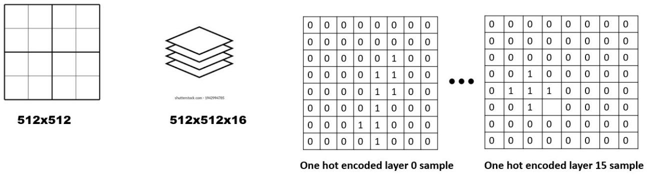

1550x2200
1550x2200
Resize image to 512x512 px
512x512
512x512
Mean Standard Normalization
Convert to Tensor
Stacking 16 layers with one hot encoding
Convert to Numpy array

Fig. 2: Image and Mask Prepossessing



0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	1	0	0
0	0	0	0	1	1	0	0
0	0	0	0	1	1	0	0
0	0	0	0	1	0	0	0
0	0	0	1	1	0	0	0
0	0	0	0	1	0	0	0

One hot encoded layer 0 sample

0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0
0	1	1	1	0	0	0	0
0	0	1		0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0

One hot encoded layer 15 sample

Fig. 3: Mask Layer stacking with one hot-encoding

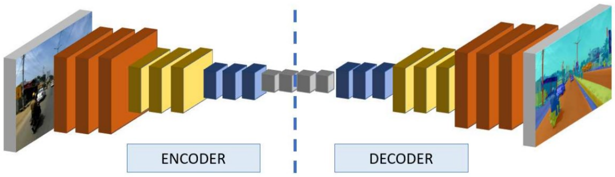


Fig. 4: Encoder Decoder like CNN architecture for semantic segmentation

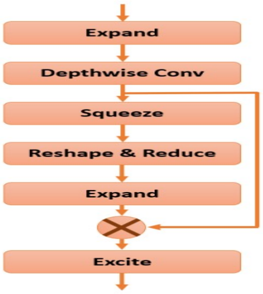


Fig. 5: Mobile Inverted Residual Bottleneck (MBConv) block

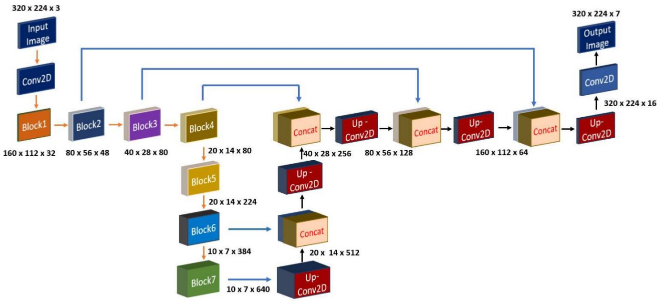


Fig. 6: EfficientNet Architecture for Semantic Segmentation

IV. Description

After reviewing the segmentation related research works, it is found that, UNet like architecture is very effective as well as popular in the deep learning based research works. Though using UNet architecture is very common in medical image segmentation related tasks, but from the above literature review its quite clear that same changed architecture can also be useful in the aerial image segmentation related problem. So we tried to look for effective architecture in terms of higher accuracy. Here we came accross with UNet++ and UNet3+ architecture.

The UNet++ model is an expansion of the initial UNet framework, which utilizes a nested structure to capture multi-scale features. The system includes of several UNet modules, each featuring an independent encoder and decoder network, as well as a sequence of skip connections connecting them together. The UNet3+ model is another variant of the UNet++ framework, designed to enhance its overall performance. The proposed approach employs a nested architecture similar to UNet++, but uses an attention mechanism to enable selective

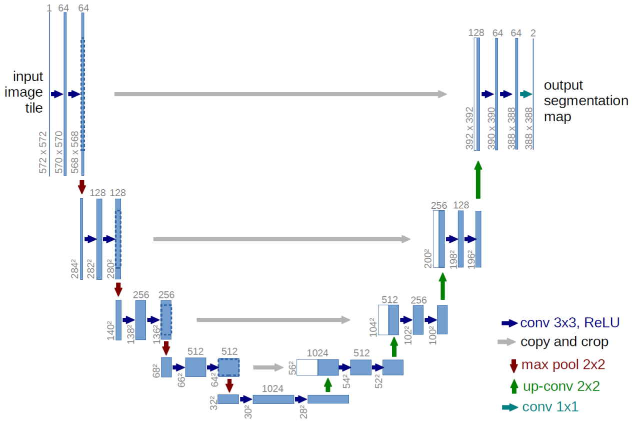


Fig. 7: Architecture of UNet

feature extraction by prioritizing salient features and disregarding irrelevant ones. The UNet3+ model utilises an attention mechanism on the skip connections combining the encoder and decoder networks, which enhances its ability to capture complex details. We tried to implement the segmentation model in both Unet and Unet++ based architecture, but for our segmentation problem UNet gave higher accuracy compared to UNet++.

To perform the semantic segmentation action we used PyTorch, PyTorch Segmentation Models and PyTorch Lightning Module. We divide our tasks in three stages.

- **Stage-1**: Preparing the data set
- **Stage-2**: Train and test the modules
- **Stage-3**: Perform further experiments and update

Initially we experimented with the following models MobileNet, ResNet, EfficientNet for the Encoder. The decoder was automatically chosen based on the encoder we selected from the PyTorch segmentation model library.

At first we unzipped all the files in Google Colab.Then we resized the image to 512x512 size and performed mean scalar normalization (figure-2). After that, we converted the images into tensors. For the mask image we performed another operation known as mask layer stacking with one hot-encoded value (figure-3). Then we created the data-set and Data-loader for the PyTorch for training and testing purpose. we used scikitlearn's test train split library to split our data-set into train, validation and test split. we took 75% data for training 10% for validation and 15% for testing purpose. To initialized the model architecture, We followed the PyTorch official documentation. By using PyTorch Segmentation Library we created our desired encoder and decoder. To train the model we took help from another library called PyTorch Lightning. Here we set it all the required parameters like loss function, optimizer, schedulers. Later after we started the training by using Google Colab's GPU as we are using higher dimension images so for faster calculation of gradients this was the best option to look for.After the training we saved the model's state and calculated mIOU, training accuracy and validation accuracy (the results are illustrated in the Experiments section).Initially we ran the training for 20 epochs then we identified the best model(so far by now UNet with efficentnetb0 model achieved highest score). After that we ran the model(EfficientNetb0) for 40 and 60 epochs the results are illustrate in table-2.

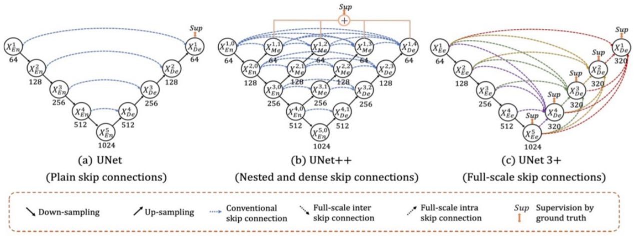

(a) UNet
(Plain skip connections)
(b) UNet++
(Nested and dense skip connections)
(c) UNet 3+
(Full-scale skip connections)
Sup
64
128
256
512
1024
320
Down-sampling
Up-sampling
Conventional skip connection
Full-scale inter skip connection
Full-scale intra skip connection
Sup Supervision by ground truth

Fig. 8: Difference in UNet, UNet++ and UNet3+ architecture

V. Experiments and Results

Here mainly we have experimented with the models and found so far efficientnet b0 gave us the best result with just 20 epochs. However running it for 40 epochs gave slightly better result which was mIOU 0.518 not that high comparing with 20 epochs 0.513. So rather wasting more GPU in Google Colab we decided to stop. We tried to experiment with the loss functions. There are a lot of loss functions were available like Dice Loss, JCard LOss, Focal Loss but we found higher mIOU by using Dice Loss. We choose Dice loss as our base loss function for experiment. We didn't experimented with the optimizers rather chose the default Adam optimizer for this model.

TABLE I: Performance of the Models

Arch.	Encoder	Loss function	Optimizers	Validation mIOU score	Test mIOU score	Epochs
UNet	Mobile Net V3	Dice Loss	Adam	0.458	0.471	20
UNet	ResNet 152	Dice Loss	Adam	0.362	-	20
UNet	Efficient Net B7	Dice Loss	Adam	0.39	-	20
UNet++	Efficient Net B0	Dice Loss	Adam	0.38	-	20
UNet	Efficient Net B0	Dice Loss	Adam	0.517	0.513	20
UNet	Efficient Net B0	Jcard Loss	Adam	0.39	0.39	20

TABLE II: EfficientNet B0 with higher epochs

Epochs	Validation mIOU score	Test mIOU score
20	0.517	0.513
40	0.526	0.518
60	0.508	0.52

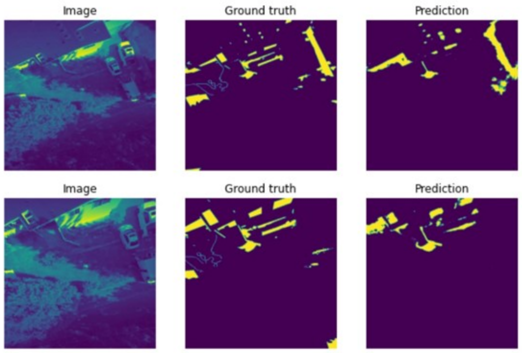


Fig. 9: Ground truth and predicted masks in EfficientNet B0 with 20 epochs

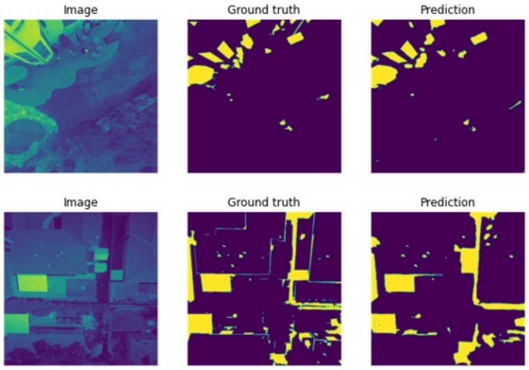


Fig. 10: Ground truth and predicted masks in EfficientNet B0 with 40 epochs

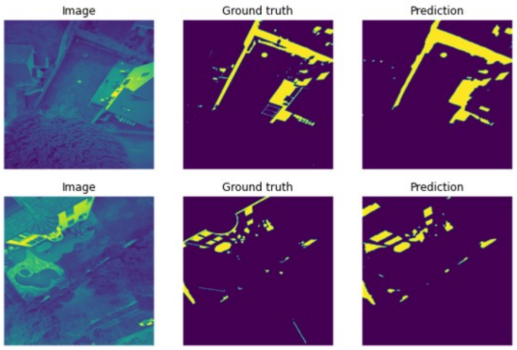


Fig. 11: Ground truth and predicted masks in EfficientNet B0 with 60 epochs

VI. Conclusion

This project was first attempt to learn Deep Learning on real world challenge. The dataset given by the organisers are not well distributed. So images contains highest amount of some classes and there are a few images for other classes. Moreover, there was no guideline and unorganized documentation was suggested by the AI Crowd challenge organizers at the beginning. However as soon as the time passed by we experimented and tried to figure out the gaps and missing points. There is no deny to the fact that we wanted to beat the baseline of 0.61 mIOU score but failed to do, so far with our current approach. It is our plan in the upcoming time to use Transformer framework and GAN (Generative Adversial Network) for this problem,to see if we can achieve more accurate result. In addition to that, we had to resize the images to lower dimension, so we lost a lot of spatial information was missing during training which might be a factor for less performance. In the upcoming time we are looking for to patching these images into smaller dimensions and train the model in it, After that we would stitch the predicted images together to gain the higher dimension masks for higher dimension image. We strongly believe despite of constraints we have come a long way and looking for new challenges to solve in the upcoming time with our gained knowledge and experience.

References

[1] Y. Guo, Y. Liu, T. Georgiou, and M. S. Lew, “A review of semantic segmentation using deep neural networks,” *International journal of multimedia information retrieval*, vol. 7, pp. 87–93, 2018.

[2] L. Wang, R. Li, C. Zhang, S. Fang, C. Duan, X. Meng, and P. M. Atkinson, “Unetformer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 190, pp. 196–214, 2022.

[3] Y. Lyu, G. Vosselman, G. Xia, A. Yilmaz, and M. Y. Yang, “The uavid dataset for video semantic segmentation,” *arXiv preprint arXiv:1810.10438*, vol. 1, 2018.

[4] H. W. F. Yeung, M. Zhou, Y. Y. Chung, G. Moule, W. Thompson, W. Ouyang, W. Cai, and M. Bennamoun, “Deep-learning-based solution for data deficient satellite image segmentation,” *Expert Systems with Applications*, vol. 191, p. 116210, 2022.

[5] A. Gebrehiwot, L. Hashemi-Beni, G. Thompson, P. Kordjamshidi, and T. E. Langan, “Deep convolutional neural network for flood extent mapping using unmanned aerial vehicles data,” *Sensors*, vol. 19, no. 7, p. 1486, 2019.

[6] S. Kumar, A. Kumar, and D.-G. Lee, “Semantic segmentation of uav images based on transformer framework with context information,” *Mathematics*, vol. 10, no. 24, p. 4735, 2022.

[7] E. Kerfoot, J. Clough, I. Oksuz, J. Lee, A. P. King, and J. A. Schnabel, “Left-ventricle quantification using residual u-net,” in *Statistical Atlases and Computational Models of the Heart. Atrial Segmentation and LV Quantification Challenges: 9th International Workshop, STACOM 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers 9*. Springer, 2019, pp. 371–380.

[8] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, “Unet++: Redesigning skip connections to exploit multiscale features in image segmentation,” *IEEE transactions on medical imaging*, vol. 39, no. 6, pp. 1856–1867, 2019.

[9] M. Lv, Y. Zhang, and S. Liu, “Fast forward approximation and multi-task inversion of gravity anomaly based on unet3+,” *Geophysical Journal International*, p. ggad106, 2023.

[10] B. Baheti, S. Innani, S. Gajre, and S. Talbar, “Eff-unet: A novel architecture for semantic segmentation in unstructured environment,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 358–359.