Enhancing House Boundary Prediction from Aerial Imagery: A Hybrid Approach with Ensemble U-Net and Transformer Attention

Muhammad Waleed
22030017@lums.edu.pk

Abdul Rauf
22030015@lums.edu.pk

Faizan Rasool
21030020@lums.edu.pk

Abstract: The use of satellite imagery has revolutionized the field of geospatial analysis, allowing the creation of a wide range of applications that can provide valuable insights into the earth's surface. Prediction of home boundaries using satellite images is one such application. This is a critical issue to resolve because accurate and precise house boundary information is required for a variety of uses ranging from property assessment to urban planning. In this paper, we analyzed different approaches including pre-existing u-net models and proposed the new effective method: A Hybrid Approach with Ensemble U-Net and Transformer Attention to predict the house boundary.

Index Terms—House boundary prediction, attention, satellite imagery

1. INTRODUCTION

Deep learning algorithms for analyzing satellite photos and estimating house borders have gained prominence in recent years. However, this is a challenging task that requires the integration of many data sources, feature extraction, and advanced modeling approaches. The purpose of this State of the Art (SOA) model is to study the most recent research on house border prediction using satellite photos in order to select and implement the most promising strategies.

The use of satellite images has transformed the field of geospatial analysis, allowing the development of a diverse set of applications that can provide useful insights into the earth's surface. One such use is the prediction of home borders using satellite photos. This is a significant issue to answer since exact house boundary information is necessary for a

variety of applications ranging from property assessment to urban planning.

2. LITERATURE REVIEW

Satellite imagery has become an important tool for understanding and analyzing various features of the earth's surface. One area where satellite imagery has found widespread application is in the prediction of house boundaries. With the increasing availability of high-resolution satellite images, there has been a surge in the development of deep learning models for accurately predicting house boundaries. In this literature review, we analyze four papers that have implemented deep learning models for this purpose.

The first paper [1] used the Google Earth dataset and employed a U-Net+VGG16 model for semantic segmentation, achieving a mean IoU of 81.79%. The second paper[2] utilized both the Massachusetts Buildings Dataset and the WHU building dataset, achieving mean IoUs of 79.62% and 83.23% respectively. The third paper [3] conducted an experimental survey of deep learning techniques for understanding satellite imagery, utilizing a nested U-Net model and achieving a mean IoU of 73.52%. Finally, the fourth paper [4] used the Worldview-2 Tianhe District of Guangzhou City dataset and applied a U-Net model, achieving an OA of 87. Overall, these papers demonstrate the effectiveness of various deep learning models in the task of predicting house boundaries through satellite images, using different datasets and achieving high accuracy levels.

The reviewed literature primarily uses four satellite image datasets for house boundary prediction. The most commonly used datasets are the WHU building dataset and the Massachusetts building dataset. These datasets are widely used in research related to building extraction and segmentation. The Worldview-2 dataset from the Tianhe District of Guangzhou City is used in one of the reviewed studies, while Google Earth imagery is used in another. These datasets contain high-resolution satellite images with different spatial and spectral characteristics. By using multiple datasets, researchers are able to test the generalizability of their models and compare their performance on different geographic regions.

Overall, these papers demonstrate the effectiveness of deep learning models in predicting house boundaries from satellite imagery. The results also highlight the importance of using appropriate datasets and architecture choices for achieving accurate predictions.

A. Analysis

Our analysis revealed that deep learning models, such as U-Net, VGG16, and nested U-Net, have been successfully applied for house boundary prediction through satellite images. The use of different datasets and models have resulted in varying levels of accuracy. While some models achieved high accuracy on a specific dataset, they may not perform as well on others. One potential weakness of the literature is the limited diversity of segmentation applied on datasets that are used in the research, which may not represent the full range of real-world scenarios.

3. DATASETS

For this project, we utilized the WHU Building dataset [6] which was provided by The Group of Photogrammetry and Computer Vision (GPCV) at Wuhan University. This dataset comprises over 8,000 images, each with a resolution of 512x512 pixels, as well as the corresponding hot-encoded segmentation mask for houses. To increase the size and diversity of the dataset, we augmented it with rotations, flips, and crops, resulting in a total of

11,000 images. This augmentation process aimed to prevent overfitting and enhance the robustness and generalizability of the model. The WHU Building dataset is widely used for predicting house boundaries and includes high-resolution satellite images of residential areas from various regions of the world.

4. BACKGROUND AND FUNDAMENTALS

A. Ensemble

A hybrid system[8] is a sort of computer system that integrates different technologies, processes or approaches to achieve a specific goal or address a problem. These systems frequently integrate various components or subsystems, each with its own distinct characteristics, in order to maximize their strengths and overcome their limits.

Deep learning has made significant advances in recent years, revolutionizing industries such as computer vision, natural language processing, and speech recognition. The usage of hybrid systems which mix various deep learning models or techniques to obtain greater performance in tackling complicated problems is one of the growing developments in the field of deep learning.

In deep learning a hybrid system is the combination of several deep learning architectures or approaches that leverage their strengths to overcome individual limits and obtain better outcomes. These hybrid systems can be created by combining neural networks of various types such as convolutional neural networks (CNNs) recurrent neural networks (RNNs) and transformer models or by combining deep learning with other machine learning techniques such as decision trees or support vector machines.

Furthermore, hybrid deep learning systems can be utilized for model ensembling which is the process of combining numerous deep learning models to create predictions. Ensemble methods like stacking or bagging, can be used to integrate the outputs of various deep learning models, thereby improving accuracy, resilience, and generalization performance.

B. U-Net

UNet[7] is a prominent convolutional neural network (CNN) architecture used in image segmentation tasks, where the goal is to divide an input image into various regions or items of interest. It was developed by researchers at the University of Freiburg in 2015 and has since become widely employed in a variety of medical image processing jobs.

The UNet architecture is made up of an encoder and decoder network that are linked by a bottleneck layer. The encoder network is made up of convolutional layers that are followed by pooling layers to downsample the input picture, whereas the decoder network is made up of upsampling layers that are followed by convolutional layers that reconstruct the output segmentation mask. The bottleneck layer between the encoder and decoder networks captures and propagates the high-level properties of the input image to the decoder network.

Overall, the UNet architecture has exhibited cutting-edge performance in a variety of picture segmentation tasks making it a popular choice among medical image analysis academics and practitioners.

C. Transformer

The Transformer[9] design, first proposed in a 2017 work by Vaswani et al., is a common deep learning architecture used in natural language processing (NLP) tasks. Since then, the Transformer design has been widely employed in a variety of NLP

applications such as language translation, question answering, and language production.

The Transformer architecture is built on the principle of self-attention, which allows the network to detect correlations between words in the input sequence. The design is made up of an encoder and decoder network, in which the encoder encodes the input sequence into a collection of hidden states and the decoder generates the output sequence based on the encoded information.

The usage of multi-head attention, where the self-attention mechanism is applied numerous times with various sets of weights, is one of the main characteristics of the Transformer architecture. As a result, the network can take note of various elements of the input sequence and integrate them to provide a more detailed representation.

Overall, the Transformer design has shown cutting-edge performance in numerous NLP tasks and has grown to be a favorite among researchers and practitioners in the area.

D. Loss Function

In this project, we utilized the root mean squared error (RMSE) as the loss function for our model. RMSE is a commonly used regression loss function that calculates the square root of the average of the squared differences between the predicted values and the actual values. By using RMSE, we aim to minimize the difference between the predicted and ground truth segmentation masks for houses in satellite images.

The RMSE loss function is well-suited for our project since it penalizes larger errors more than smaller ones. Additionally, it provides a good measure of the overall performance of the model by considering both the magnitude and direction of the errors. By minimizing the RMSE loss during training, we can ensure that our model accurately segments houses in satellite images and generalizes well to new, unseen data.

4. APPROACHES

A. Ensemble U-Net

Initially, we developed an ensemble approach that consists of U-Net models (Fig 1.0). The ensemble model is implemented by integrating 3 U-Net models in order to improve its accuracy. The model takes the satellite image as input and outputs the segmented image of the detected objects. This is done by passing input images through 3 separate identical U-net models and each model generates its outputs depending on its extracted features then output from all 3 U-net models is aggregated in order to get the final output. The results produced by it can be seen in the result section.

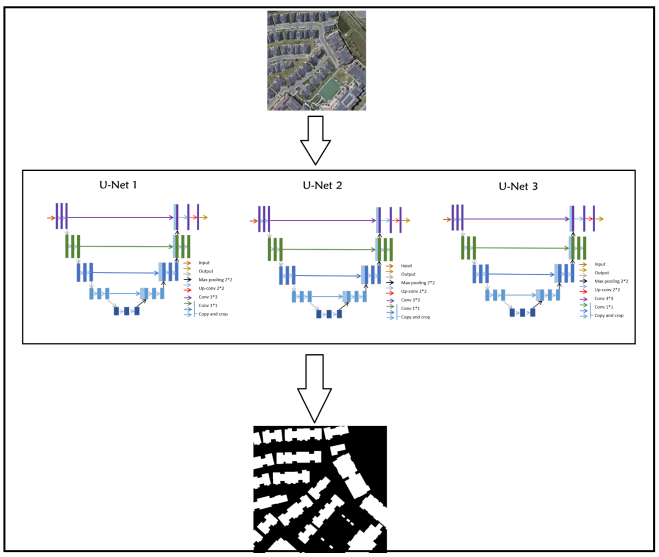

Fig 1.0 Ensemble U-Net Architecture

Input	The Ensemble U-Net model takes a satellite image as its input. The input image is preprocessed to ensure that it has the correct size and format required by the model.
Encoding	The input image is then passed through the encoding layers of each U-Net model. These layers are designed to extract high-level features from the image, which are then used by the model for segmentation and edges.
Decoding	Once the image has been encoded, each U-Net model passes it through its decoding layers. These layers use the features extracted by the encoding layers to produce output maps: a segmentation map.
Ensemble	The output of each U-Net is aggregated to get the final segmentation.
Segmentation	The segmentation map produced by the Ensemble U-Net model indicates which parts of the input image correspond to the objects that need to be segmented. This is achieved by assigning a probability value to each pixel in the segmentation map, with higher values indicating a greater likelihood that the pixel belongs to an object.
Output	The Ensemble U-Net model outputs the segmentation map of objects detected in the input image.

Tabel 1.0: Ensemble U-Net Workflow

B. Ensemble Attention U-Net

Our second approach involved an ensemble model consisting of U-Net models with attention mechanisms. This approach aimed to enhance the model's performance by integrating multiple U-Net models and leveraging their extracted features.

The ensemble model takes satellite images as input and uses three separate attention-based U-Net models to generate outputs. Each U-Net model focuses on different aspects of the image and generates its own output. The outputs from all the U-Net models are then combined or aggregated to produce the final segmentation mask.

Our experimental results demonstrate that this ensemble approach significantly improves the accuracy of object segmentation compared to a single U-Net model. The attention mechanisms used in the U-Net models enable the model to focus on relevant features in the input image and make more accurate predictions. The final outputs produced by the ensemble model are shown in the result section of our research paper.

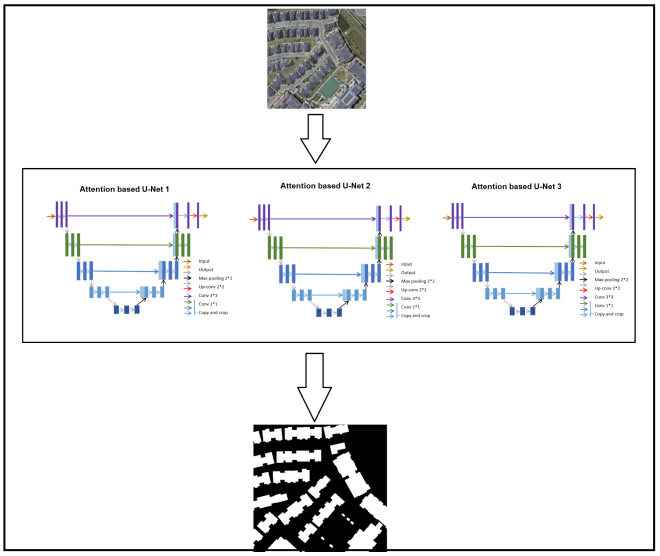

Fig 2.0: Attention based Ensemble U-Net Architecture

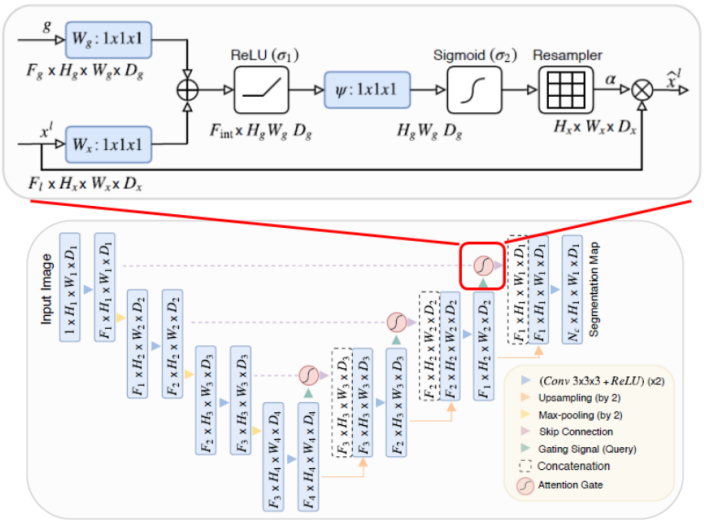

Fig 3.0: Attention based single Unet[10]

Input	The attention based Ensemble U-Net takes an input image as its input. The input image is preprocessed to ensure that it has the correct size and format required by the model.
Decoding	Once the image has been encoded, the 3 U-Net model passes it through the decoding layers. These layers use the features extracted by the encoding layers to produce output maps: a segmentation map.
Ensemble	The output of each U-Net is aggregated to get the final segmentation.
Segmentation	The segmentation map produced by the attention based Ensemble U-Net model indicates which parts of the input image correspond to the objects that need to be segmented. This is achieved by assigning a probability value to each pixel in the segmentation map, with higher values indicating a greater likelihood that the pixel belongs to an object.
Output	The Ensemble U-Net model outputs the segmentation map of objects detected in the input image.

Tabel 2.0: Attention based Ensemble U-Net Workflow

C. Ensemble Transformer U-Net

In our research, we experimented with three different variations of the model to improve the accuracy of house segmentation in satellite images. The third and final model we developed is a combination of U-Net models and Transformer layers, which we integrated into an ensemble model. Each of the three identical U-Net models includes a Transformer layer and a Multihead Attention layer in the bottleneck between the U-Net encoder and decoder to enhance feature extraction and capture long-range dependencies. The model takes satellite images as input, which are then decoded by the U-Net decoder and passed through the Multihead Attention layer. Finally, the model outputs a segmentation mask that accurately identifies the location of houses in the images. Our experimental results show that this ensemble model outperforms the other two variations, demonstrating the effectiveness of incorporating Transformer layers and multi-model ensembles in satellite image segmentation tasks. We believe that this model has the potential for real-world application in various fields such as urban planning, disaster management, and resource allocation.

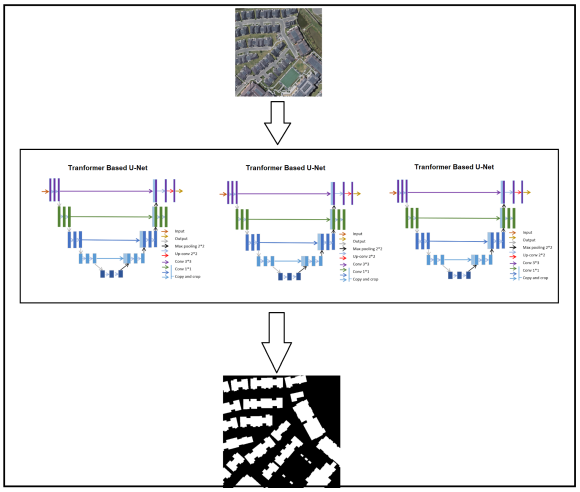

Fig 3.0: Transformer based Ensemble U-Net Architecture

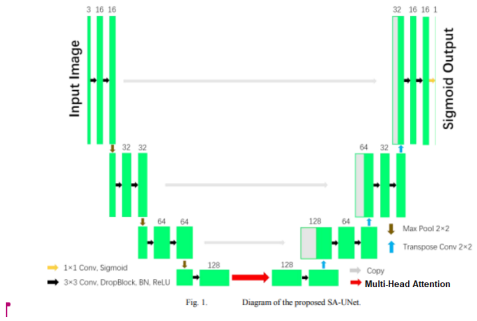

Fig 4.0: Transformer based U-net [11]

6. RESULTS

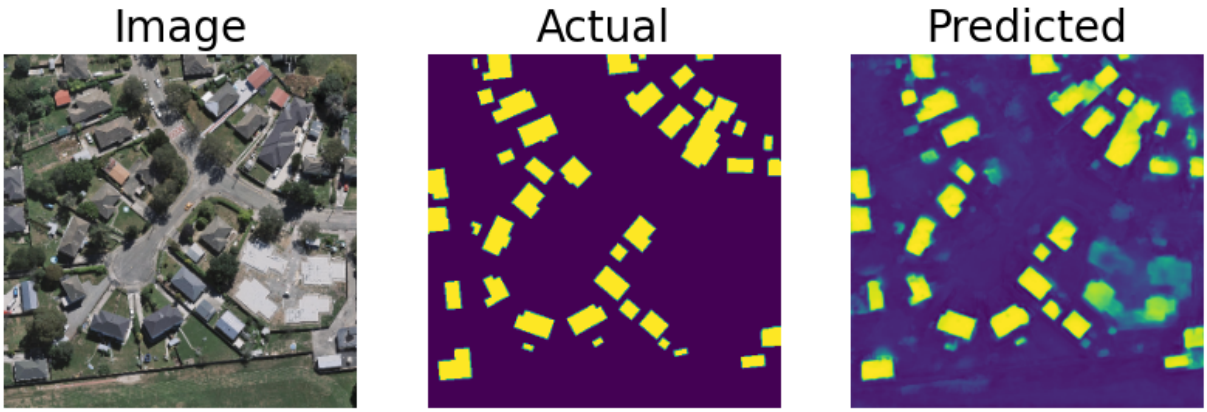

Fig 5.0: First Model (Ensemble U-net**)**

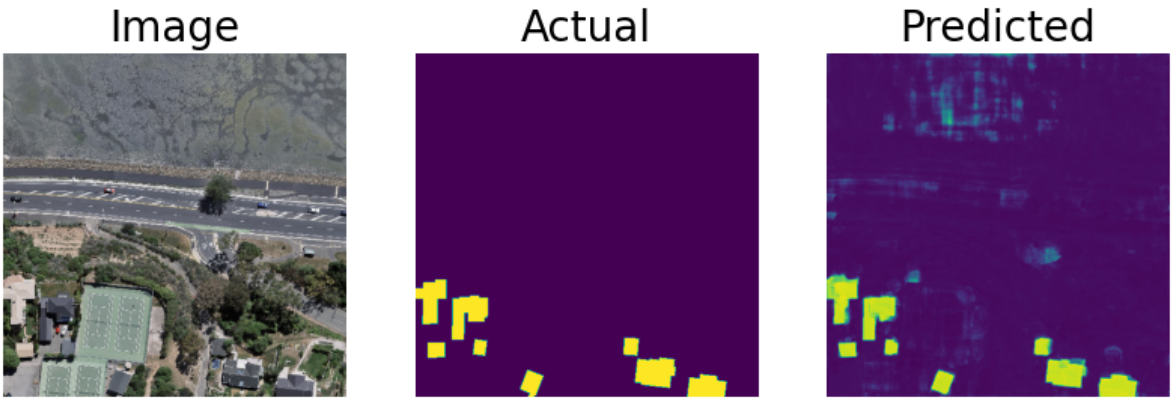

Fig 6.0: Second Model (Ensemble U-net with attention**)**

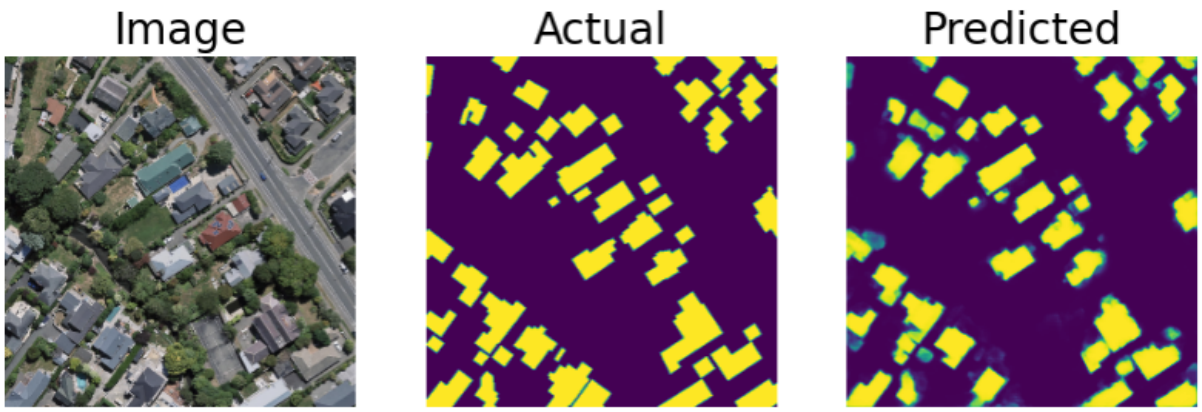

Fig 7.0: Third Model (Ensemble U-net with Transformer**)**

A. Comparative Results

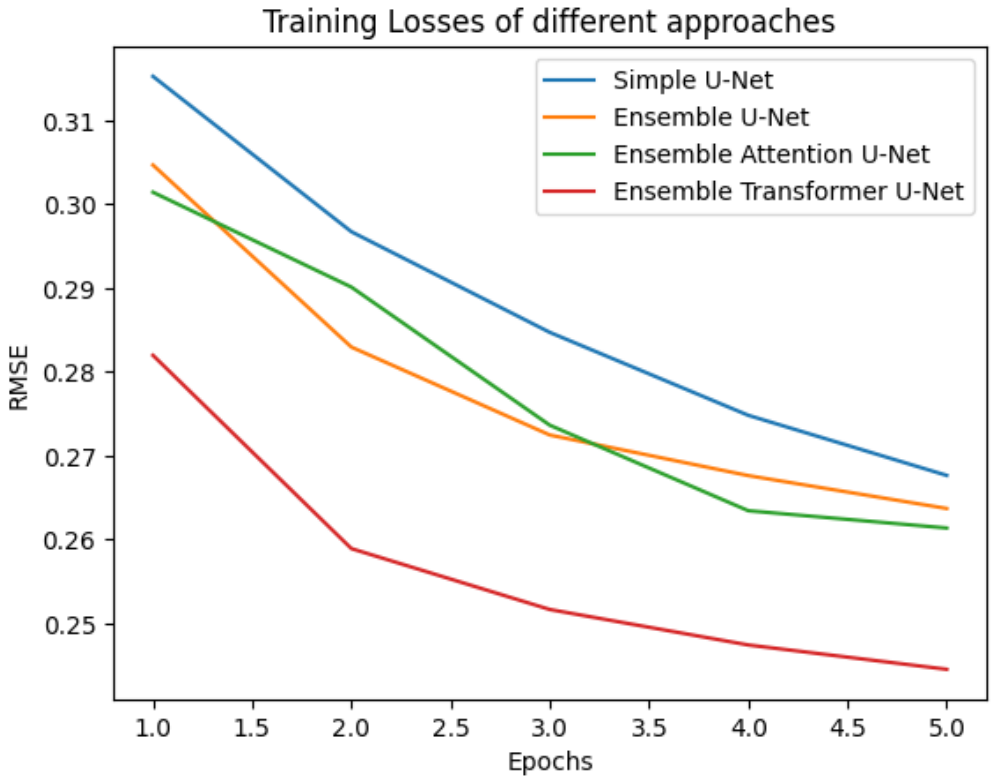

Fig 8.0: Comparative Training Analysis

Model	Epochs	Batch Size	Training Images	Test Images	Test RMSE Loss
U-net	5	5	11,000	4,800	0.219419
Ensemble U-net	5	5	11,000	4,800	0.216946
Ensemble U-net with attention	5	5	11,000	4,800	0.212891
Ensemble U-net with Transformer	5	5	11,000	4,800	0.207431

Tabel 3.0: Testing Loss of different models

7. CONCLUSION

To summarize, the issue of detecting home boundaries via segmentation is a difficult challenge in computer vision. We compared multiple distinct models for this job in this study: U-Net, Ensemble U-Net, Ensemble attention U-Net and Ensemble Transformer U-Net. In terms of accuracy and efficiency, our results show that the Ensemble Transformer U-Net surpasses the other models. The suggested model combines the strengths of the U-Net and Transformer architectures to improve segmentation outcomes. As a result, we can infer that the Ensemble Transformer U-Net is the best model for detecting house boundaries using segmentation, and it has the potential to be used in other segmentation tasks in the future.

8. FUTURE WORK

Moving forward, there are several avenues for future work to further improve the accuracy and robustness of our house segmentation model.

Firstly, we can experiment with more diverse datasets that include images of houses from different parts of the world with varying styles and structures. This would help us create a more generalized model that can accurately detect houses in any location.

Secondly, the quality of the images used in our dataset can be improved, and we can incorporate multiple images of the same area taken at different times to capture temporal changes in the scene. This would make our model more adaptable to changes in the environment and improve its overall performance.

Thirdly, with the availability of more compute resources, we can train our model for more epochs, and experiment with adding more filters in the CNN and more layers to improve its accuracy.

Lastly, we can modify our current model architecture by using a different base segmentation model and then make modifications to it. This would allow us to explore different network architectures and see how they perform on the same tas

REFERENCES

[1] S. Deng et al., "Scattered Mountainous Area Building Extraction From an Open Satellite Imagery Dataset," in IEEE Geoscience and Remote Sensing Letters, vol. 20, pp. 1-5, 2023, Art no. 6003105, doi: 10.1109/LGRS.2023.3247620.

[2] Q. Liu et al., "CFNet: An Eigenvalue Preserved Approach to Multiscale Building Segmentation in High-Resolution Remote Sensing Images," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 16, pp. 2481-2491,
2023, doi: 10.1109/JSTARS.2023.3244336.

[3] Zhongyu Li, Yang Liu, Yin Kuang, Huajun Wang, Cheng Liu, "Remote sensing
images semantic segmentation method based on improved nested UNet," Proc. SPIE 12552, International Conference on Geographic Information and Remote Sensing Technology (GIRST 2022), 1255223 (10 February 2023);
https://doi.org/10.1117/12.2667484

[4] Pan, Z.; Xu, J.; Guo, Y.; Hu, Y.; Wang, G. Deep Learning Segmentation and
Classification for Urban Village Using a Worldview Satellite Image Based on U-Net. Remote Sens. 2020, 12, 1574. https://doi.org/10.3390/rs12101574

[5] paperswithcode.com: Papers With Code. (n.d.). WHU Dataset. Retrieved from
https://paperswithcode.com/dataset/whu

[6] kaggle.com: Xiaoqian970429. (n.d.). WHU Building Dataset. Retrieved from https://www.kaggle.com/datasets/xiaoqian970429/whu-building-dataset

[7] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv preprint arXiv:1505.04597. Retrieved from https://doi.org/10.48550/arXiv.1505.0459

[8] M.A. G., Minghui H, A.K., M. T., P.N. S. Ensemble deep learning: A review
https://doi.org/10.48550/arXiv.2104.02395

[9] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is All You Need. In Advances in Neural Information Processing Systems (NIPS) https://doi.org/10.48550/arXiv.1706.03762

[10] Oktay, O., Schlemper, J., Le Folgoc, L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B., Glocker, B., & Rueckert, D. (2018). Attention U-Net: Learning Where to Look for the Pancreas. arXiv preprint arXiv:1804.03999. https://doi.org/10.48550/arXiv.1804.03999

[11] Guo, C., Szemenyei, M., Yi, Y., Wang, W., Chen, B., & Fan, C. (2020). SA-UNet: Spatial Attention U-Net for Retinal Vessel Segmentation. arXiv preprint arXiv:2004.03696.
https://doi.org/10.48550/arXiv.2004.03696