Enhancing Aerial Image Segmentation: A Comparative Analysis of Machine Learning Approaches

Puneetha Dharmapura Shrirama
Machine Learning and Data Analytics Lab
Friedrich-Alexander-Universitaet Erlangen-Nuernberg (FAU)
Erlangen, Bayern, Germany

Figure 1: A collage of aerial imagery (left) paired with corresponding ground truth segmentation masks (right). The masks highlight key land cover classes such as roads, vegetation, and buildings.

ABSTRACT

Aerial image segmentation is crucial for remote sensing applications such as urban planning, disaster management, environmental monitoring, and land-use classification. Previous methods relying on CNNs achieved high accuracy but struggled to capture long-range spatial dependencies and suffered from domain shifts across datasets. Although transformer-based models like SegFormer improved global feature extraction, they typically required extensive computational resources, limiting practical deployment. Efficiently classifying urban land types (buildings, roads, vegetation) from aerial imagery remains essential, particularly due to challenges like irregular object shapes, occlusions, and mixed land-cover types. This study aims to evaluate the effectiveness of transformer-based segmentation (SegFormer) and zero-shot segmentation (Segment Anything Model, SAM) combined with heuristic rules for accurate and computationally efficient urban land classification from aerial imagery. SegFormer was applied for its capability to model global-local spatial relationships, while SAM provided zero-shot object segmentation enhanced with heuristic classification rules. Performance was evaluated using mean Intersection over Union (IoU), Dice score, precision, recall, F1 score, and pixel accuracy. The SAM + Heuristics model achieved higher mean pixel accuracy (0.5956), demonstrating better overall performance, while SegFormer exhibited strengths specifically in building detection (mean IoU: 0.21552). The study indicates that combining transformer-based global feature extraction with heuristic refinements provides a promising pathway toward scalable and accurate aerial image segmentation. Code and datasets are publicly available at: ***https://github.com/puni-ram48/Aerial_Image_Segmentation***

KEYWORDS

Aerial Image Segmentation, Deep Learning, SegFormer, SAM, Heuristic Classification, Remote Sensing.

1 INTRODUCTION

Aerial image segmentation is a crucial component of remote sensing, supporting applications such as urban planning, disaster management, environmental monitoring, and land-use classification. With the increasing availability of very high-resolution (VHR) aerial imagery, researchers have focused on developing advanced segmentation models capable of accurately classifying complex landscapes. Traditional approaches relied on handcrafted features and classical machine learning algorithms; however, these methods typically struggle to generalize across different datasets due to variations in sensor characteristics, lighting conditions, and geographic differences. The introduction of deep learning architectures, particularly Convolutional Neural Networks (CNNs) and Transformer-based models, significantly improves segmentation accuracy and automation [3, 6].

The emergence of Fully Convolutional Networks (FCNs) marks a major breakthrough in aerial image segmentation. Unlike conventional CNNs, FCNs enable end-to-end pixel-wise classification, eliminating the need for manual feature extraction. Sherrah demonstrated that FCNs achieved state-of-the-art segmentation accuracy, preserving fine-grained details without requiring interpolation [6]. Building on this, Marmanis et al. proposed an ensemble of CNNs incorporating deconvolution layers to enhance spatial resolution and refine object boundaries [3]. Their study, conducted on the International Society for Photogrammetry and Remote Sensing (ISPRS) Vaihingen Dataset, showed that deep learning models significantly outperformed traditional methods, confirming their effectiveness in aerial image segmentation.

Despite these advancements, CNN-based models generally struggle to capture long-range spatial dependencies due to their restricted receptive fields. To address this limitation, researchers introduced Vision Transformers (ViTs), which utilize self-attention mechanisms to model global contextual relationships within images. Wang et al. proposed UNetFormer, a hybrid UNet-like Transformer model designed to improve segmentation accuracy while maintaining computational efficiency [8]. By incorporating ResNet18 as an encoder and employing global-local attention mechanisms, UNetFormer achieved 84.1 % mean Intersection over Union (mIoU) on the ISPRS Vaihingen dataset. However, transformer-based models typically require significant computational resources, limiting their practicality for real-time applications.

Another major challenge in aerial image segmentation is domain shift, where models trained on one dataset (e.g., ISPRS Potsdam) typically perform poorly when applied to another dataset (e.g., ISPRS Vaihingen). This discrepancy arises from differences in sensor modalities, spectral properties, and geographic variations. Benjdira et al. addressed this issue by proposing a Generative Adversarial Network (GAN)-based domain adaptation method, which improved segmentation accuracy from 35 % to 52 % when transitioning between datasets [1]. Their approach transformed source domain images to resemble the target domain, reducing dataset discrepancies and improving generalization. This study demonstrated that unsupervised domain adaptation is a viable solution for aerial image segmentation, particularly when labeled data is scarce.

Although deep learning-based segmentation models have greatly advanced, computational efficiency remains a critical concern. Wang et al. developed an optimized transformer-based segmentation model that achieved an inference speed of 322.4 FPS on an NVIDIA GTX 3090, making it one of the fastest high-resolution segmentation frameworks and demonstrating the trade-off between accuracy and efficiency [8]. Nevertheless, real-time applications still require more efficient segmentation methods without sacrificing accuracy.

Existing aerial image datasets typically focus either on semantic segmentation or object detection, lacking benchmarks specifically for instance segmentation—a task combining object detection and pixel-level segmentation. Instance segmentation in aerial imagery is particularly challenging due to numerous instances per image, large variations in object scales, and many tiny objects. To address this gap, Waqas et al. introduced the first large-scale benchmark dataset designed explicitly for instance segmentation in aerial imagery, termed iSAID (Instance Segmentation in Aerial Images Dataset) [9]. Popular instance segmentation methods originally developed for natural scenes, such as Mask R-CNN (Mask Region-based Convolutional Neural Network) and PANet (Path Aggregation Network), performed poorly on iSAID, highlighting the need for specialized methods tailored to aerial imagery.

To overcome the limitations outlined above, the SegFormer model [10] and Segment Anything Model (SAM) [2] were employed, combined with heuristic methods to classify roads, buildings, and vegetation. SegFormer efficiently captures both local and global image contexts through hierarchical transformer structures, enhancing semantic segmentation performance. SAM offers the advantage of segmenting objects without labeled annotations, enabling effective zero-shot segmentation. Because SAM alone does not provide semantic class labels, heuristic methods were applied to classify segmented objects based on their characteristics. This integration provides a fully automated segmentation and classification pipeline with improved adaptability and robustness, especially in scenarios where labeled datasets are unavailable or limited. This multi-scale feature learning increases scalability.

In conclusion, advancements in CNNs, transformer-based architectures, and domain adaptation techniques have significantly improved aerial image segmentation accuracy and automation. However, ongoing challenges such as domain shift, small-object detection, and computational complexity continue to hinder wider practical adoption. This study aims to address these challenges through a novel combination of transformer-based segmentation (SegFormer) and zero-shot segmentation capabilities (SAM), further improving accuracy, computational efficiency, and generalizability.

2 METHODOLOGY

In this study, two advanced methods were applied for segmentation and classification of aerial images: SegFormer and Segment Anything Model (SAM) combined with heuristic rules. SegFormer was employed for semantic segmentation of large-scale land features, while SAM + Heuristic was used for fine-grained segmentation and classification. The performance of both methods was quantified using standard evaluation metrics, including mean Intersection over Union (mIoU), precision, recall, and F1-score.

2.1 DATASET

The aerial dataset consisted of high resolution images (See **Figure 1**) that included various types of land cover such as roads, buildings, and vegetation. The dataset contained a total of 80 images, all of which were resized to 1024x1024 pixels to ensure consistency and compatibility with the models. Of these, 23 images were manually annotated to serve as ground truth to evaluate the segmentation results. It is important to note that models were not trained but instead pre-trained models were used for segmentation and classification. SegFormer and SAM models were applied to the unlabeled images, and evaluation metrics were calculated based on the manually annotated images.

2.2 SegFormer

SegFormer is a transformer-based architecture designed specifically for semantic segmentation tasks. It used an encoder-decoder structure, where the encoder extracted multi-scale features from the input image, and the decoder uses these features, through multi-layer perceptron (MLP) layers, to generate the final segmentation mask. The encoder captured long-range dependencies between pixels using transformer blocks, which processed the image in overlapping patches. This approach was enhanced by Efficient Self-Attention mechanisms to improve segmentation accuracy [10].

For this study, we used the pre-trained SegFormer-B5 model, which had been fine-tuned on the Cityscapes dataset, a well-known urban scene segmentation dataset. The model was selected for its ability to generalize across urban environments, making it highly suitable for large-scale segmentation tasks like those found in our aerial dataset. Importantly, the architecture was used without modifications or fine-tuning (See **Figure 2**). The input images were resized to meet the model's input requirements and processed into

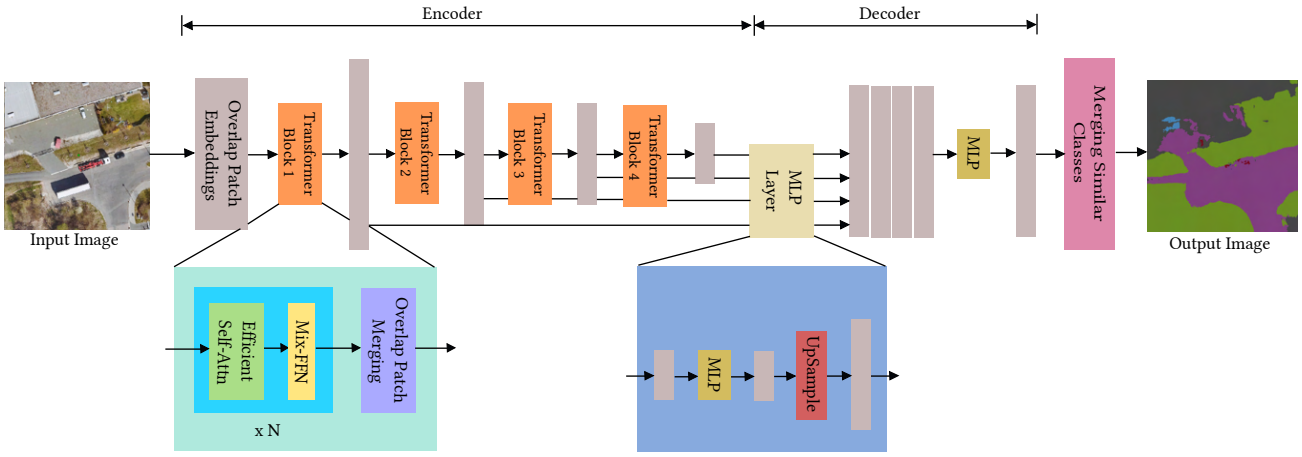

Figure 2: The SegFormer architecture, consisting of a hierarchical Transformer-based encoder and a lightweight All-MLP decoder. The encoder generates overlapping patch embeddings, applies transformer blocks with efficient self-attention and Mix Feed-Forward Networks (Mix-FFN), and performs overlap patch merging to create multi-scale representations. The decoder fuses these multi-level features through MLP layers and upsampling, generating a semantic segmentation mask refined by merging similar classes [10]

tensors using SegFormerImageProcessor, which prepared the data for segmentation.

The model's encoder extracted hierarchical features from the input image, and the decoder combined these features to predict pixel-wise class labels. The model's output was a logit map, which was converted into the segmentation mask by applying the argmax function to select the most probable class for each pixel. During post-processing, related classes were merged to better align with the land cover types in the aerial dataset. For example, sidewalks (class 1), poles (class 5), and traffic signs (class 7) were combined into a single road class (class 0). Similarly, terrain (class 9) was assigned to vegetation (class 8), and walls (class 3) and fences (class 4) were merged into the building class (class 2). To visualize the merged classes, a Cityscapes colormap was applied, providing a distinct color representation for each class in the final segmentation mask.

2.3 SAM + HEURISTIC

SAM is a state-of-the-art segmentation model developed by Meta [2], designed to generate segmentation masks for objects in an image. SAM utilized a vision transformer (ViT) architecture, where the ViT-L image encoder processed the input image to generate an image embedding. The model also incorporated a prompt encoder for embedding user-defined inputs such as clicks or bounding boxes [2]. However, in this study, we focused on SAM's automatic mask generation capability, which generated segmentation masks based on the input image, without the need for user prompts.

For this study, the pre-trained SAM model was directly applied to segment the aerial images. The input images were resized to 1024x1024 pixels to meet the model's input size requirements. The automatic mask generation feature of SAM was used to generate segmentation masks for various regions in the image, such as roads, buildings, and vegetations.

After generating the segmentation masks, a heuristic-based classification approach was employed to refine the segmentation (See **Figure 3**). This classification was based on two main features: HSV (Hue, Saturation, Value) color space values and contour properties. The HSV values of each segmented region were analyzed to determine whether it matched the characteristics of roads, buildings, or vegetations. For instance, roads were identified using specific saturation and value ranges, while vegetation were detected based on hue and saturation. Additionally, contour features such as area and aspect ratio (AR) were used to classify the regions further, with roads exhibiting larger areas and higher aspect ratios, and buildings showing more rectangular shapes. Following the classification, the segmented regions were colorized based on their assigned class (e.g., purple for roads, green for green areas, and gray for buildings

Additionally, the generated masks were merged if there was significant overlap, with the mask merging process identifying and combining overlapping masks above a certain threshold (e.g., 0.5). This merging step ensured that the segmentation results were clean and that any ambiguous regions were consolidated for accurate classification. The final classified image was then saved for further evaluation. The combined use of SAM's automatic mask generation and heuristic classification yielded better segmentation results for land cover types in the aerial dataset.

2.4 EVALUATION METRICS

In the context of aerial image segmentation, various evaluation metrics [4, 7] are used to assess the model performance. Below are the definitions and formulas for key segmentation evaluation metrics [5] .

2.4.1 ***Mean Intersection over Union (Mean IoU) :*** Intersection over Union (IoU), also known as the Jaccard Index, measures the overlap between the predicted segmentation and the ground truth. Mean IoU (mIoU) is the average IoU across all classes [7] .

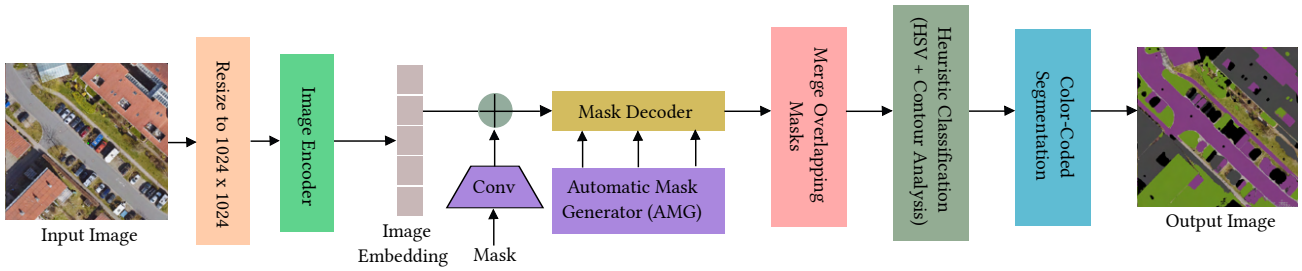

Figure 3: Semantic segmentation pipeline utilizing SAM's Automatic Mask Generator (AMG) combined with heuristic classification. The input image is resized to 1024 × 1024 pixels, encoded into embeddings, and automatically decoded into segmentation masks. Overlapping masks are then merged, followed by heuristic classification (using HSV color space and contour analysis) to categorize segments, ultimately producing a color-coded segmentation map [2]

$$\text{IoU}_c = \frac{|\text{TP}_c|}{|\text{TP}_c| + |\text{FP}_c| + |\text{FN}_c|} \tag{1}$$

$$\text{mIoU} = \frac{1}{N}\sum_{c=1}^{N} \text{IoU}_c \tag{2}$$

where:

- TP_c : True positives for class c
- FP_c : False positives for class c
- FN_c : False negatives for class c
- N : Number of classes

2.4.2 ***Mean Dice Score :*** Dice Score (or F1 Score for segmentation) evaluates the similarity between two sets. It is defined as the harmonic mean of precision and recall. The mean Dice Score averages this metric across all classes [7] .

$$\text{DS}_c = \frac{2|\text{TP}_c|}{2|\text{TP}_c| + |\text{FP}_c| + |\text{FN}_c|} \tag{3}$$

$$\text{Mean Dice Score} = \frac{1}{N}\sum_{c=1}^{N} \text{DS}_c \tag{4}$$

2.4.3 ***Mean Pixel Accuracy :*** Pixel accuracy measures the proportion of correctly classified pixels. Mean Pixel Accuracy averages the accuracy for each class [7] .

$$\text{PA} = \frac{\sum_{c=1}^{N} |\text{TP}_c|}{|T|} \tag{5}$$

$$\text{Mean Pixel Accuracy} = \frac{1}{N}\sum_{c=1}^{N} \frac{|\text{TP}_c|}{|\text{TP}_c| + |\text{FN}_c|} \tag{6}$$

where:

- TP_c: True positives for class c
- T: Total number of pixels in the ground truth

2.4.4 ***Mean Precision :*** Precision measures how many predicted positive pixels are actually correct. Mean Precision averages this metric across all classes [4] .

$$\text{P}_c = \frac{|\text{TP}_c|}{|\text{TP}_c| + |\text{FP}_c|} \tag{7}$$

$$\text{Mean Precision} = \frac{1}{N}\sum_{c=1}^{N} \text{P}_c \tag{8}$$

2.4.5 ***Mean Recall :*** Recall measures how many actual positive pixels were correctly predicted. Mean Recall averages recall across all classes [4] .

$$\text{R}_c = \frac{|\text{TP}_c|}{|\text{TP}_c| + |\text{FN}_c|} \tag{9}$$

$$\text{Mean Recall} = \frac{1}{N}\sum_{c=1}^{N} \text{R}_c \tag{10}$$

2.4.6 ***Mean F1 Score :*** F1 Score is the harmonic mean of precision and recall. Mean F1 Score averages the F1 scores of all classes [4] .

$$\text{F1}_c = \frac{2 \cdot \text{P}_c \cdot \text{R}_c}{\text{P}_c + \text{R}_c} \tag{11}$$

$$\text{Mean F1 Score} = \frac{1}{N}\sum_{c=1}^{N} \text{F1}_c \tag{12}$$

These metrics are widely used in semantic segmentation tasks, including aerial image analysis, to evaluate model performance across different land cover classes. The mean versions provide a comprehensive view of segmentation quality across all categories.

3 RESULTS

A detailed comparison of the segmentation results is presented in **Table 1** . Metrics range from 0 to 1, where higher values indicate better segmentation performance. SAM + Heuristics achieves higher Mean Pixel Accuracy (0.5956) compared to SegFormer (0.5103), indicating superior overall pixel-level accuracy. SAM + Heuristics outperforms SegFormer in most classes, particularly Background, Road, and Vegetation. For example, in the Road class, SAM + Heuristics obtains significantly higher Mean IoU (0.55037) and Dice Score (0.68945), compared to SegFormer's 0.32301 and 0.44890, respectively. Similarly, for Vegetation, SAM + Heuristics achieves better Mean IoU (0.49495 vs. 0.43979). However, SegFormer performs better in the Building class, achieving higher Mean IoU (0.21552 vs. 0.069997) and Mean Recall (0.59458 vs. 0.09454). This indicates that

Table 1: Comparison of Evaluation Metrics for SegFormer and SAM + Heuristics

Metric	Background		Building		Road		Vegetation	
	SegFormer	SAM + Heuristics	SegFormer	SAM + Heuristics	SegFormer	SAM + Heuristics	SegFormer	SAM + Heuristics
Mean IoU	0.00490	**0.10152**	**0.21552**	0.069997	0.32301	**0.55037**	0.43979	**0.49495**
Mean Dice Score	0.00967	**0.17593**	**0.31462**	0.11312	0.44890	**0.68945**	0.57046	**0.61690**
Mean Precision	0.07126	**0.13695**	**0.24970**	0.20669	0.66434	**0.74371**	**0.85275**	0.71920
Mean Recall	0.00643	**0.36650**	**0.59458**	0.09454	0.39790	**0.69952**	0.49748	**0.61711**
Mean F1 Score	0.00967	**0.17593**	**0.31462**	0.11312	0.44890	**0.68945**	0.57046	**0.61690**
Mean Pixel Accuracy: **SegFormer:** 0.5103 | **SAM + Heuristics:** 0.5956								

Figure 4: SegFormer best-case segmentation example. Left to right: input image (resized to 1024×1024), ground truth segmentation, and the SegFormer model output, demonstrating high-quality segmentation closely matching the ground truth.

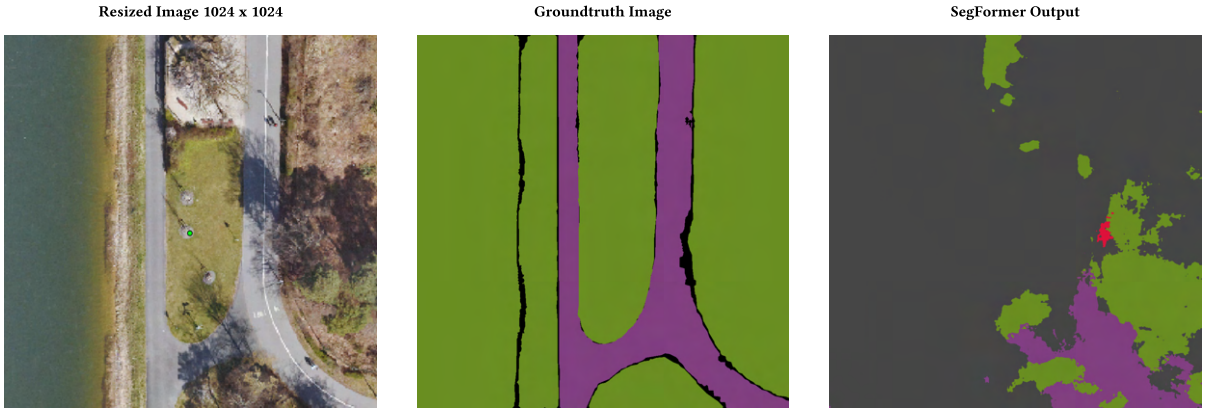

Figure 5: SegFormer challenging segmentation example. Left to right: resized input image, ground truth, and SegFormer output highlighting significant misclassifications and omissions.

SegFormer is more effective at identifying structured objects such as buildings, typically characterized by complex yet uniform features.

Figure 4 and **Figure 5** illustrate SegFormer's best and worst segmentation results. In the best case, it effectively distinguishes roads from vegetation, whereas in the worst case, it struggles with building boundaries, causing misclassification. Similarly, **Figure 6** and **Figure 7** depict SAM + Heuristics' strongest and weakest performances, demonstrating accurate segmentation of roads and green areas but revealing challenges with urban structures, particularly complex rooftops and overlapping shadows. These visual findings confirm the quantitative evaluation, clearly highlighting each method's strengths and limitations.

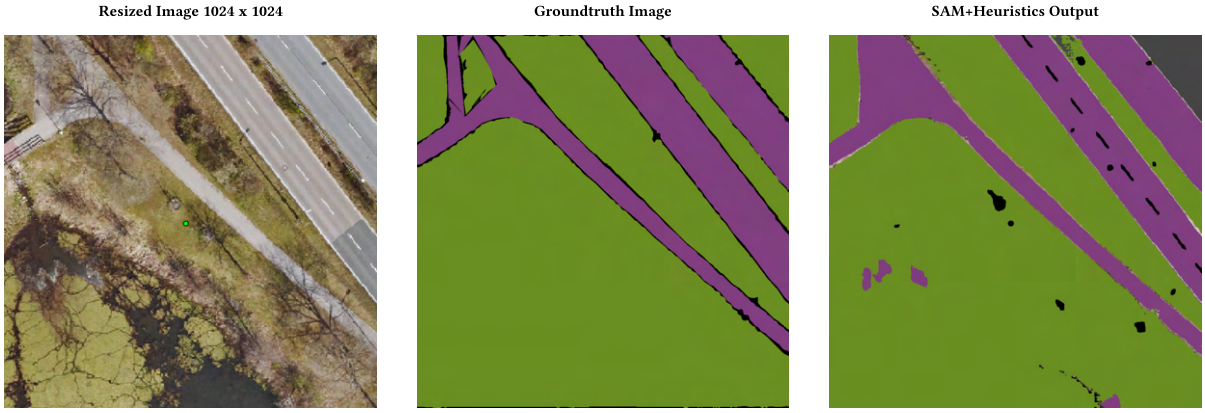

Figure 6: Best-case segmentation example using Segment Anything Model (SAM) with heuristic-based classification. Left to right: resized input image (1024×1024), ground truth segmentation, and SAM+Heuristics output showing accurate segmentation closely matching ground truth.

Figure 7: Challenging segmentation example illustrating limitations of SAM combined with heuristic rules. Left to right: resized input image, ground truth segmentation, and SAM+Heuristics output highlighting notable misclassifications and incomplete regions.

4 DISCUSSION

Accurate aerial image segmentation is essential for urban planning, environmental monitoring, and disaster management, yet remains challenging due to domain-specific complexities such as high object density, varying scales, shadows, and domain shift issues. This study evaluated two state-of-the-art approaches—SegFormer and SAM combined with heuristic methods—on high-resolution aerial images, revealing distinct strengths and weaknesses through quantitative and qualitative analyses.

SegFormer demonstrated superior performance in segmenting structured and complex urban objects, such as buildings. This is due to SegFormer's transformer-based architecture, which captures global contextual relationships effectively through self-attention mechanisms. In the Building class, SegFormer achieved a Mean IoU of 0.21552, significantly outperforming SAM + Heuristics (Mean IoU: 0.069997) **Table 1**. Transformer-based models, such as UNetFormer by Wang et al. [8], have been shown to perform well on structured objects due to their ability to process long-range dependencies in the image.

However, SegFormer's performance in segmenting natural features like Roads and Vegetation was limited, with Mean IoU values of 0.32301 for Road and 0.43979 for Vegetation, which were significantly lower than those of SAM + Heuristics. These discrepancies can be attributed to domain shift, as SegFormer was trained on the Cityscapes dataset, which comprises street-level imagery. Aerial imagery, with its distinct visual patterns, scales, and top-down perspectives, posed significant challenges. Benjdira et al. (2019) [1] also emphasized the impact of domain adaptation, highlighting that models trained on one type of data often underperform when applied to different domains, as seen in SegFormer's performance.

Figure 4 and Figure 5 demonstrate these challenges. **Figure 4** highlights SegFormer's successful segmentation of structured objects like buildings, but it also reveals limitations, particularly in delineating building boundaries, as shown in **Figure 5**. The model struggles with natural features due to its lack of fine-tuning for aerial imagery, resulting in misclassifications, especially in road and vegetation segmentation, where objects appear smaller or obscured from the aerial perspective.

SAM combined with heuristic methods (SAM + Heuristics) excelled in segmenting natural features like Roads and Vegetation, largely due to SAM's zero-shot segmentation capability combined with the heuristic post-processing classification rules. In the Road class, SAM + Heuristics achieved a Mean IoU of 0.55037 and a Dice Score of 0.68945, significantly outperforming SegFormer. SAM's ability to segment objects without requiring labeled data provides a scalable solution, especially in situations with limited annotations.

However, SAM + Heuristics struggled with urban features, particularly Buildings, where it obtained a lower Mean IoU of 0.069997 compared to SegFormer. The reliance on heuristic methods introduces inconsistency in segmentation, especially when dealing with shadows, complex rooftops, and overlapping objects. **Figure 7** exemplifies these issues, showing that SAM + Heuristics struggles with urban structures, particularly misidentifying rooftops due to overlapping shadows. SAM + Heuristics performs best on natural features like Roads and Vegetation, where the reliance on color and intensity-based heuristic rules leads to more consistent segmentation, as shown in **Figure 6**. While SAM + Heuristics accurately segmented roads and vegetation in some areas, the model is prone to errors, particularly in urban settings with varying lighting and shadows, which affect segmentation accuracy.

Both SegFormer and SAM + Heuristics show promising performance but fall short of the state-of-the-art in some areas. Wang et al. [8] achieved a Mean IoU of 84.1 % on the ISPRS Vaihingen dataset with UNetFormer, far surpassing the performance of both SegFormer and SAM + Heuristics. This discrepancy is due to the lack of fine-tuning on aerial-specific datasets, which is crucial for improving model performance. Benjdira et al. [1] also emphasized the need for domain adaptation, which can significantly reduce domain shift, a key limitation in SegFormer's performance on aerial imagery.

In contrast, SAM + Heuristics performed well on Roads and Vegetation, but it struggled with Buildings, primarily due to the reliance on heuristic rules. This aligns with challenges seen in previous research, where models struggled to generalize across complex urban structures [6]. Despite these challenges, SAM + Heuristics is highly scalable due to its zero-shot segmentation capability, and when combined with domain adaptation techniques, could significantly improve segmentation accuracy.

5 CONCLUSION & FUTURE WORK

In this study, two advanced segmentation methods, SegFormer and SAM combined with heuristic rules, were evaluated for segmenting aerial imagery. SAM + Heuristics achieved superior pixel-level accuracy overall, particularly for classes characterized by natural elements like roads and vegetation. In contrast, SegFormer provided better results for structured objects such as buildings. While SAM + Heuristics demonstrated greater flexibility and robustness to segment natural features, its heuristic-based classification showed sensitivity to shadows and complex urban features. Conversely, SegFormer was effective in segmenting structured objects but was limited by domain shifts due to differences between street-level and aerial imagery.

Future research should focus on addressing the limitations identified in this study. Fine-tuning SegFormer on domain-specific aerial datasets will improve its generalization to aerial images and enhance its segmentation of natural elements. Additionally, optimizing heuristic strategies for SAM to better handle variations in lighting, shadows, and complex urban structures will improve its reliability in urban environments. Another avenue for future work includes utilizing larger SAM model checkpoints (e.g., ViT-H optimized for 2048×2048 resolution), which could help exploit SAM's full potential for high-resolution aerial segmentation tasks.

REFERENCES

[1] Bilel Benjdira, Yakoub Bazi, Anis Koubaa, and Kais Ouni. Unsupervised domain adaptation using generative adversarial networks for semantic segmentation of aerial images. *Remote Sensing*, 11(11):1369, 2019.

[2] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[3] Dimitrios Marmanis, Jan D Wegner, Silvano Galliani, Konrad Schindler, Mihai Datcu, and Uwe Stilla. Semantic segmentation of aerial images with an ensemble of cnss. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2016*, 3:473–480, 2016.

[4] David Martin Ward Powers. Evaluation: From precision, recall and f-measure to roc, informedness, markedness & correlation. *Journal of Machine Learning Technologies*, 2(1):37–63, 2011.

[5] Oona Rainio, Jarmo Teuho, and Riku Klén. Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, 14(1):6086, 2024.

[6] Jamie Sherrah. Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery. *arXiv preprint arXiv:1606.02585*, 2016.

[7] Abdel Aziz Taha and Allan Hanbury. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *BMC Medical Imaging*, 15(1):1–28, 2015.

[8] Libo Wang, Rui Li, Ce Zhang, Shenghui Fang, Chenxi Duan, Xiaoliang Meng, and Peter M Atkinson. Unetformer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 190:196–214, 2022.

[9] Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. isaid: A large-scale dataset for instance segmentation in aerial images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 28–37, 2019.

[10] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.