Zero-Shot vs. Few-Shot CLIPSeg: Efficient Urban Feature Segmentation

Yun-Dam Ko
Department of Civil and Environmental Engineering
yundamko@stanford.edu

Abstract

The building sector accounts for nearly 40% of global carbon emissions, highlighting the urgent need for decarbonization and the importance of urban building energy modeling (UBEM). However, conventional UBEM frameworks often neglect surrounding urban systems, lacking scalable methods to quantify and integrate them. To address this, CLIP-based (Contrastive Language–Image Pretraining) segmentation is evaluated for extracting urban features (green spaces, roads, built areas, buildings) from high-resolution imagery with sparse annotations. Zero-shot CLIPSeg is compared to few-shot fine-tuning, which—with augmentation and optimized loss—improves mean Intersection-over-Union (mIoU) from 0.236 (zero-shot) to 0.443 (few-shot), an increase of 87.7%. Notably, the building class mIoU rises from 0.248 to 0.694 (179.8% increase). Qualitative results show sharper representation of urban structures after fine-tuning, though some classes (e.g, road and built area) remain difficult. These results suggest that prompt-driven CLIPSeg provides a viable baseline for urban feature extraction, and that minimal fine-tuning can substantially improve performance in limited-label scenarios relevant to urban features relevant to building energy.

1. Introduction

The building sector is a significant contributor to global carbon emissions, accounting for nearly 40% of the total [13]. In response to this urgent need for decarbonization, urban building energy modeling (UBEM) has emerged as a pivotal methodology for evaluating and enhancing the energy performance of existing building stocks at the urban scale [12]. However, traditional UBEM frameworks often exhibit limitations by neglecting the crucial influence of surrounding urban systems on building energy dynamics. These systems, encompassing diverse elements such as green spaces, asphalt roads, and varying land cover types, exert a substantial impact on local microclimates, thereby influencing heating and cooling demands within buildings. Consequently, the accurate quantification of these urban systems and their subsequent integration as variables within building energy models represent a significant challenge. A key obstacle hindering this integration lies in the scarcity of scalable measurement methods capable of comprehensively characterizing the urban context.

Recent advancements in computer vision technologies offer promising avenues for addressing this methodological gap. Specifically, the capacity of computer vision to automate the extraction of urban features from remotely sensed imagery presents a compelling solution. However, the effective deployment of these technologies is contingent upon the availability of robust machine learning models capable of accurate feature identification. A persistent challenge in this domain is the substantial data requirement for training such models; labeled data, particularly at the scale required for urban analysis, are often scarce or entirely absent in many urban environments. This limited availability of labeled data severely restricts the applicability of conventional supervised learning approaches.

This study aims to address this challenge by investigating the performance of vision models under limited-label scenarios, specifically comparing zero-shot and few-shot segmentation strategies employing Contrastive Language-Image Pre-Training (CLIP)-based models, exemplified by CLIPSeg. As visually represented in Figure1, zero-shot segmentation, leveraging the pre-trained model's inherent knowledge, serves as a readily deployable baseline. Furthermore, this study explores the potential for substantial performance gains through fine-tuning this pre-trained model on a carefully curated, yet limited, dataset of fewer than 20 images. The efficacy of the proposed methodologies is evaluated based on their ability to accurately identify key urban features – including green spaces, roads and buildings – from high-resolution satellite imagery. The study focuses specifically on the applicability of these techniques to enable scalable and accurate urban context extraction for integration within UBEM applications, ultimately contributing to a more comprehensive and realistic understanding of building energy performance at the urban scale.

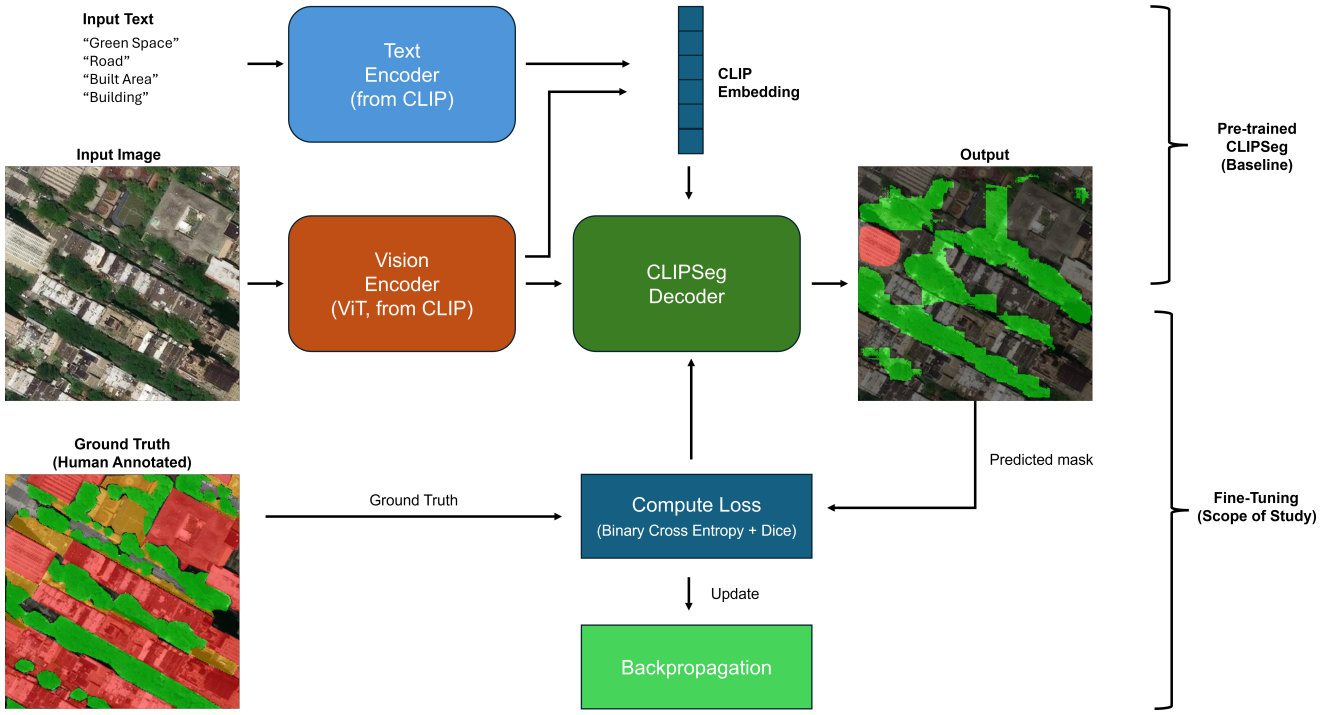

Figure 1. Architecture and Training Pipeline of the CLIPSeg-based Segmentation Model

2. Related Work

Recent advances in computer vision have been instrumental in extracting energy-relevant features from urban imagery. Early approaches heavily relied on convolutional neural networks (CNNs), which, while strong in classification and segmentation, often require extensive manual annotation and struggle to capture long-range contextual relationships in complex urban scenes.

Several CNN-based approaches have been proposed to capture energy-relevant urban features. For example, Mayer et al. [10] predicted building energy efficiency by extracting visual cues from street-view and aerial imagery, using a pretrained Inception-v3 encoder to focus on façades and rooftops and integrating auxiliary data via a multilayer perceptron. While their method achieved high classification accuracy, it also exposed challenges related to limited-label learning and cross-regional generalization. In a similar vein, Boccalatte and Chanussot [3] performed semantic segmentation of rooftop obstructions to enhance solar potential estimates; they trained a U-Net with a ResNet-152 backbone on weakly supervised vector labels and attained moderate IoU scores despite annotation scarcity. The 3D-PV Locator project [11] further demonstrated the scalability of CNN classifiers under sparse labeling by projecting roof surfaces into multiple views and classifying rooftop solar panels using a ResNet-based network. Collectively, these studies illustrate that fine-tuning pre-trained CNNs can successfully identify objects of interest; however, they tend to concentrate on specific targets (e.g., solar panels) and require relatively large datasets (on the order of approximately 30,000 to 100,000 samples), which, while necessary for achieving robust performance, may nonetheless be appropriate for scalable measurement of various urban features.

CNNs have also been employed for building segmentation and detection in urban environments. Zhang et al. [18] proposed a hybrid CNN-transformer framework for extracting building footprints from remote sensing imagery. Their method filters irrelevant tiles, enhances image resolution, and applies instance segmentation to improve boundary precision and detection accuracy. Behera et al. [2] proposed a two-stage CNN framework that incorporates superpixel-based preprocessing for segmenting urban features such as roads, vegetation, and buildings from UAV (Unmanned Aerial Vehicle) imagery, achieving high segmentation accuracy. Huang et al. [5] evaluated instance segmentation of individual building rooftops with fine-grained roof-type labels using Mask R-CNN, Cascade Mask R-CNN, and SOLOv2. Their analysis revealed that cascade architectures performed best for distinguishing roof types but struggled with small or densely clustered structures. While these studies propose methodologies that effectively identify buildings, they often suffer from limitations in scalability, such as reliance on aerial imagery or the utilization of large-scale datasets for training. Furthermore, some are limited by fo-

cusing solely on a single building class, thereby restricting their applicability.

Although CNN-based techniques perform well on specific tasks, they often struggle to capture the complex interactions among urban components—such as trees, roads, built-up areas, and buildings—that influence building energy performance, particularly in regions where annotated imagery is scarce. Material variations and spatial context necessitate models that recognize both local appearance and long-range dependencies, limitations that fixed-kernel CNNs cannot fully address.

The transition to Transformer-based models addresses the critical need for capturing long-range dependencies and contextual relationships prevalent in urban scenes and beyond. CNNs, while efficient at extracting local features, are constrained by limited receptive fields, hindering the modeling of complex spatial arrangements and interdependencies. In contrast, Transformer architectures leverage self-attention mechanisms, enabling each element to attend to all others regardless of distance, thereby facilitating the learning of global contextual information. Key advantages over CNNs include superior modeling of long-range dependencies, enhanced flexibility with variable-length inputs, and the capacity to learn intricate feature representations without explicit feature engineering, making them particularly apt for semantic segmentation and object detection in complex urban environments.

Transformer-based architectures offer distinct advantages in modeling complex spatial dependencies. Liu and Abbasabadi [7] classified façade materials from street-view images using a Vision Transformer, achieving high accuracy and contributing to improved UBEM simulations. Li et al. [6] proposed a hybrid few-shot segmentation framework for high-resolution land-cover mapping, combining multiple base learners with a Projection onto Orthogonal Prototypes (POP) network in a two-stage process. Their model effectively recognizes both base classes (e.g., tree, building) and novel classes (e.g., bridge, river, vehicle) using only a few labeled examples, with a final fusion step to enhance prediction accuracy. Yi et al. [17] introduced UAVFormer, a composite transformer network tailored for urban scene segmentation in UAV imagery. It incorporates adaptive feature fusion, aggregation window multi-head self-attention, and a position attention module, and was evaluated on eight urban classes, including building, road, tree, static car, and clutter. Wang et al. [15] presented UNetFormer, a lightweight architecture for segmenting urban scenes in remote sensing images. The model combines a CNN encoder with a transformer-based decoder, achieving high performance on classes such as buildings, roads, trees, vegetation, and cars, while maintaining real-time efficiency.

While transformer-based methods have shown strong performance in urban feature segmentation, they largely fall outside the few-shot learning paradigm, typically relying on large datasets and full supervision. Despite their effectiveness, many still require substantial manual annotation or complex architectural tuning, limiting their scalability in data-scarce urban settings.

Building on these developments, recent studies illustrate a clear trajectory from conventional CNN-centric models toward hybrid and transformer-based frameworks capable of modeling complex interdependencies across urban surfaces. This evolution underscores the growing demand for scalable, low-label methods that can jointly segment multiple feature types critical to urban energy modeling and climate-responsive planning.

3. Dataset

A satellite image dataset focusing on the central area of Manhattan, New York City, was constructed for this study. The imagery was obtained using the Esri (Environmental Systems Research Institute) World Imagery basemap [4]. Esri provides a widely used collection of global satellite imagery frequently utilized in geographic information systems (GIS). The satellite tiles are served in the Web Mercator projection.

To encompass the urban core of Manhattan Island, characterized by its dense mixture of buildings, roads, parks, and waterfront areas, the study utilized three 7168×7168 pixel satellite images. This image dataset corresponds to an approximate ground area of $25.5km \times 8.5km$ at zoom level 18, with a resolution of approximately 0.6 meters per pixel. For subsequent segmentation and analysis, the images were tiled into 600 individual patches of 512×512 pixels.

For the development of fine-tuning and test datasets, 23 images were manually annotated using Labelme [14]. Of these, 18 images were designated for fine-tuning and 5 for testing. The annotation included pixel-level masks for the following classes: green space, road, built area, and building. The ground truth masks are binary, indicating the presence (1) or absence (0) of each class at each pixel.

To enhance model robustness under limited data, each of the 18 original training images was augmented into 10 variants, resulting in a total of 180 training samples. The augmentation pipeline applied a mix of geometric and photometric transformations. Geometric augmentations included horizontal and vertical flips (each with a 50% probability), random cropping to 448×448 pixels, and moderate shift, scale, and rotation adjustments. Photometric augmentations involved random changes in brightness and contrast, hue-saturation-value shifts, RGB channel perturbations, as well as the addition of Gaussian noise and blur. All transformations were applied with moderate intensity and probability, and the same operations were consistently applied to the corresponding binary masks to preserve label accuracy.

The augmented dataset was used to fine-tune the CLIPSeg model, while the original 5 manually labeled images were reserved for testing.

4. Methodology

4.1. CLIPSeg Model

CLIPSeg [9] is a transformer-based vision–language model built on CLIP's pretrained image and text encoders. CLIP itself is trained on large-scale image–text pairs to learn joint embeddings, and CLIPSeg extends this by adding a lightweight segmentation head for dense, pixel-wise mask prediction. Given a natural language prompt and an input image, CLIPSeg encodes both using CLIP's ViT-based image encoder and its text encoder, computes a similarity map between visual tokens and the prompt embedding, and upsamples the resulting map to produce a segmentation score for each pixel. This enables zero-shot segmentation—i.e., mask generation without additional task-specific training. In this project, each 512×512 aerial tile $I_{\text{orig}} \in \mathbb{R}^{3\times512\times512}$ is first resized to

$$I \;=\; \text{Resize}(I_{\text{orig}}), \quad I \in \mathbb{R}^{3\times352\times352},$$

to match the pretrained model's expected resolution. A textual prompt

$$c \;\in\; \{\text{“green space”, “road”, “built area”, “building”}\}$$

is tokenized by the text encoder to produce

$$\tau(c) \;=\; \text{TextEncoder}(c), \quad \tau(c) \in \mathbb{R}^{D},$$

where D is the embedding dimension (512 in our case), while the visual encoder extracts feature maps

$$\phi(I) \;=\; \text{VisualEncoder}(I), \quad \phi(I) \in \mathbb{R}^{D\times H\times W},$$

where H and W are the height and width of the visual feature map. A small multi-layer perceptron (MLP) head f then fuses $\phi(I)_{i,j}$ and $\tau(c)$ at each spatial location (i, j) to compute logits:

$$z_{i,j} = f\big(\phi(I)_{i,j},\, \tau(c)\big), \\ (i,j) \in \{1,\dots,H\} \times \{1,\dots,W\}.$$

Applying a sigmoid activation σ yields pixel probabilities:

$$p_{i,j} \;=\; \sigma\big(z_{i,j}\big),$$

and the binary prediction mask $\hat{M}^c \in \{0,1\}^{352\times352}$ is obtained by thresholding:

$$\hat{M}^c(i,j) \;=\; \big[p_{i,j} > 0.5\big].$$

Finally, $\hat{M}^c$ may be upsampled as needed to match downstream resolutions.

Implementation is performed with Hugging Face Transformers[16] and the output is a logit map $\{z_{i,j}\}$. During inference, each resized 352×352 tile is processed with a given prompt to produce $\hat{M}^c$, which is then compared against ground-truth annotations (e.g., via IoU and accuracy).

4.2. Fine-tuning CLIPSeg

To tailor the pre-trained CLIPSeg model to the task of segmenting key urban elements in Manhattan satellite imagery, a fine-tuning process was implemented. A critical component of this process was optimizing the fine-tuning procedure itself via a hyperparameter search, conducted using Optuna [1] and its Tree-structured Parzen Estimator (TPE). The hyperparameters explored were:

- Learning rate (lr): sampled from a log-uniform distribution between 10^{-6} and 10^{-4}.
- Weight decay (weight_decay): sampled from a log-uniform distribution between 10^{-3} and 10^{-1}.
- Dice loss weight (dice_weight): sampled from a uniform distribution between 0.1 and 1.0. This parameter controls the contribution of the Dice loss to the overall loss function, in addition to the binary cross-entropy (BCE) loss.

The objective function was defined as the average validation loss over a fixed number of epochs (10 in this case). The validation loss was computed as a weighted sum of the BCE loss and the Dice loss, according to the following equation:

$$Loss = BCE + \text{dice\_weight} \cdot \text{DiceLoss} \tag{1}$$

The dataset was split into training and validation sets, with 20% of the images reserved for validation. All images were resized to 352×352 pixels, and text prompts corresponding to each semantic class (green space, road, built area, building) were used as input to the CLIPSeg model. To mitigate overfitting, several regularization strategies were employed. The AdamW optimizer [8] was used, incorporating the sampled learning rate and weight decay. A learning-rate scheduler reduced the learning rate by a factor of 0.5 when the validation loss plateaued for two consecutive epochs. Early stopping was also applied with a patience of three epochs: if the validation loss did not improve for three consecutive epochs, training was terminated.

The objective function was defined as the mean validation loss over 10 epochs, where each trial was trained for a maximum of 10 epochs. The validation loss was computed

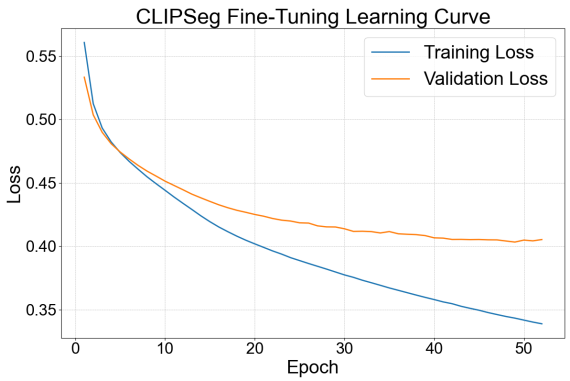

Figure 2. Learning Curve for CLIPSeg Fine-tuning

as a weighted sum of the binary cross-entropy (BCE) loss and the Dice loss, according to the following equation:

$$\text{Loss} = \text{BCE} + \text{dice\_weight} \cdot \text{DiceLoss} \tag{2}$$

The hyperparameter search was conducted over 30 trials with a total timeout of 3 hours. To conserve computational resources, pruning was enabled to terminate unpromising trials early. The optimal hyperparameters—a learning rate of 1.575e-05, weight decay of 0.002, and Dice loss weight of 0.240—were selected based on the lowest average validation loss of 0.451, computed across 10 epochs during the tuning process. These best-performing hyperparameters were then used to fine-tune the CLIPSeg model on the augmented training dataset, aiming to improve segmentation performance. Fine-tuning yielded a best validation loss of 0.403, and early stopping was triggered at epoch 52. The learning curve for this fine-tuning process is shown in Figure 2.

5. Results

5.1. Quantitative Results

Table 1 presents a detailed comparison of the Intersection over Union (IoU) scores between the pre-trained CLIPSeg model and our fine-tuned CLIPSeg model. The IoU metric, defined as the ratio of the area of overlap between the predicted segmentation and the ground truth to the area of their union, was used to quantitatively evaluate the segmentation performance of both models across various classes and test images. The IoU is calculated as follows:

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}} = \frac{|A \cap B|}{|A \cup B|} \tag{3}$$

where A represents the predicted segmentation mask and B represents the ground truth mask.

Table 1. Per-image and overall IoU (%) comparison between the pre-trained and fine-tuned CLIPSeg models. The higher value in each row is **bolded**.

Image	Class	Pre-trained	Fine-tuned
Test Image 01	green space	0.453	**0.714**
	road	0.401	**0.491**
	built area	0.000	**0.012**
	building	0.118	**0.828**
	Mean	0.243	**0.511**
Test Image 02	green space	0.672	**0.692**
	road	0.100	**0.272**
	built area	0.000	**0.306**
	building	0.660	**0.794**
	Mean	0.358	**0.516**
Test Image 03	green space	0.405	**0.548**
	road	0.022	**0.099**
	built area	0.000	**0.127**
	building	0.065	**0.652**
	Mean	0.122	**0.356**
Test Image 04	green space	0.568	**0.746**
	road	0.000	**0.179**
	built area	0.000	**0.029**
	building	0.050	**0.692**
	Mean	0.154	**0.412**
Test Image 05	green space	**0.612**	0.557
	road	0.251	**0.564**
	built area	0.000	**0.055**
	building	0.347	**0.504**
	Mean	0.309	**0.420**
Overall Mean	green space	0.542	**0.651**
	road	0.155	**0.321**
	built area	0.000	**0.106**
	building	0.248	**0.694**
	Mean	0.236	**0.443**

As shown in Table 1, the fine-tuned CLIPSeg model demonstrates a significant improvement in segmentation accuracy compared to the pre-trained model. Specifically, the overall mean IoU across all classes increased from 0.236 to 0.443, representing an 87.7% improvement after fine-tuning. This indicates a substantial enhancement in the model's ability to accurately segment different objects in the images.

The fine-tuned model shows particularly notable improvements in the segmentation performance of the *building* and *road* classes. This is especially significant, as accurate building segmentation is essential for the downstream application of this study—Urban Building Energy Modeling (UBEM)—and asphalt roads are known to be major contributors to the urban heat island effect. For instance, the mean IoU for the *building* class increased markedly from 0.248 to 0.694 after fine-tuning, representing a 179.8% im-

provement. These results demonstrate that effective few-shot learning, even on a relatively small dataset, can substantially enhance performance for critical urban applications.

Similarly, the 'road' class saw an improvement from 0.155 to 0.321 (a 107.1% increase). The 'built area' class also improved, though from 0.000 to 0.106, suggesting that the pre-trained model struggled with this more ambiguous category and, despite fine-tuning, performance remained limited. These results indicate that the fine-tuning process effectively adapted the model to better recognize and delineate the 'building' and 'road' categories, while highlighting the continued challenges in accurately segmenting the more ambiguous 'built area' class.

While the fine-tuned model generally surpasses the pre-trained model's performance, the pre-trained model achieves a marginally higher IoU for the 'green space' class in Test Image 05 (0.616 vs. 0.561). This may be attributed to 'green space' being the most readily identifiable class among the four, as evidenced by its highest mean IoU in the pre-trained model. This suggests that, in specific instances, the pre-trained model retains some advantages, and that few-shot fine-tuning is most effective for classes or prompts that pre-trained models struggle to accurately capture.

In summary, the quantitative results presented in Table 1 demonstrate the effectiveness of few-shot fine-tuning of the CLIPSeg model in improving image segmentation accuracy. The substantial increase in the overall mean IoU, along with significant improvements in key urban features such as buildings and roads, underscores the advantages of adapting a pre-trained model to a domain-specific, few-shot dataset.

5.2. Qualitative Results

Figure 3 presents CLIPSeg overlay visualizations for the five test images alongside their corresponding annotated ground truth, providing a qualitative assessment of segmentation performance.

The segmentation results for the *building* class vary considerably across test images. In particular, Test Images 01, 03, and 04 show clear failures in building detection by the pre-trained model, whereas Test Images 02 and 05 exhibit relatively accurate segmentation. This inconsistency is a key factor contributing to the performance gap between the pre-trained and fine-tuned models, indicating that few-shot fine-tuning improves robustness and consistency in urban feature segmentation.

While the quantitative results indicate improved performance in *road* segmentation with the fine-tuned model, qualitative analysis reveals more nuanced challenges. For example, in Test Image 01 (Figure 3), the fine-tuned model fails to segment a clearly visible road (upper left side, lined with trees) that the pre-trained model correctly identified. However, the pre-trained model struggled with accurately segmenting adjacent tree regions. This trade-off illustrates the ongoing difficulty in balancing the segmentation of visually similar or adjacent classes.

The pre-trained model also exhibits coarse and imprecise boundaries, particularly in distinguishing *built area* (e.g., park paths) from *green space* (e.g., parks), and in misclassifying regions between buildings as either *road* or *built area*. As seen in Test Images 01, 03, and 04, the model struggles to extract buildings from complex 2D imagery. In Test Image 03, for instance, buildings partially occluded by trees are missed, and a prominent structure with a dark gray roof is misclassified as *road*. This misclassification may be due to roof-mounted white structures that visually resemble cars on asphalt, highlighting a fundamental limitation of the pre-trained model in cluttered urban contexts. These findings underscore the importance of few-shot adaptation for urban scene segmentation.

Shadows cast by buildings frequently pose classification challenges, particularly near object boundaries. These shadowed regions are often misclassified as *built area*, which typically includes sidewalks. Additionally, in Test Image 03, an urban stream running diagonally from the center toward the lower-left corner is missegmented as *road*, likely due to its long and narrow shape and its location between building blocks. This highlights the importance of incorporating a broader set of semantic classes to enhance segmentation fidelity, as features such as streams and roads have distinct visual characteristics and differing impacts on the urban microclimate.

In summary, the qualitative evaluation highlights the limitations of the pre-trained CLIPSeg model in handling complex urban imagery, including challenges with occlusions, shadows, and semantically ambiguous regions. Few-shot fine-tuning substantially improves segmentation consistency and accuracy across key classes such as buildings and roads. While the results demonstrate clear gains, some challenging cases remain, suggesting that there is still room for refinement. These findings underscore both the value and the ongoing need for domain adaptation in urban scene understanding, particularly under limited supervision and diverse visual conditions.

6. Conclusion

This study investigated the potential of applying few-shot learning to urban systems segmentation using the CLIPSeg model, by comparing its performance to that of zero-shot inference using the pretrained model. The experimental results demonstrated a substantial performance gain through fine-tuning. Specifically, few-shot fine-tuning led to an increase in overall mean IoU from 0.236 to 0.443, representing an 87.7% improvement.

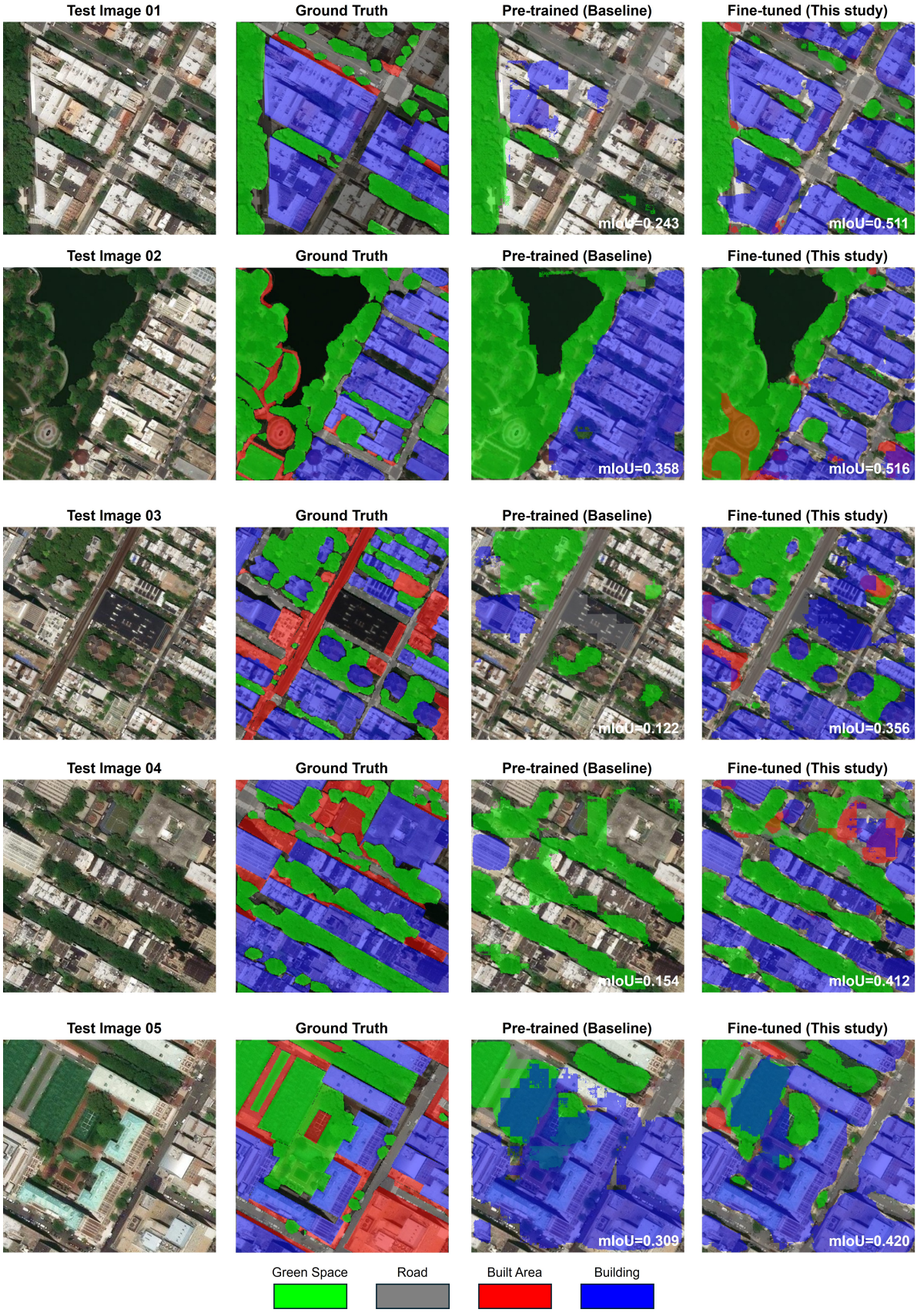

Figure 3. CLIPSeg overlay visualization: Pre-trained and fine-tuned segmentation compared to ground truth.

The most significant gain was observed in the segmentation of buildings—critical features for the downstream application of Urban Building Energy Modeling (UBEM)—with mean IoU rising from 0.248 to 0.694, a 179.8% improvement. These findings underscore the effectiveness and necessity of few-shot learning in enhancing segmentation accuracy for urban-scale applications.

Accurately segmenting urban systems remains inherently challenging, due not only to the limitations of 2D imagery but also to ambiguities in class definitions and potential errors in manually annotated ground truth masks. For instance, the *built area* class—intended to encompass various artificial surface coverings—proved difficult to delineate precisely, leading to segmentation ambiguity and inconsistency in human labeling.

The reliance on static 2D images further complicates the task. Differentiating between building rooftops and cement-covered open spaces, for example, can be difficult even for human annotators, particularly in unfamiliar urban contexts. Similarly, distinguishing parking lots from roads, or determining whether trees are located along streets, in parks, or on rooftops, is often non-trivial without additional spatial cues.

To address these limitations, future work could refine class definitions in alignment with downstream objectives—for example, distinguishing between surface materials such as asphalt, water, grass, and sand—to better support UBEM. Additionally, incorporating street-view imagery alongside satellite data could provide 3D contextual cues that improve the accuracy of urban feature quantification for energy performance analysis.

7. Contributions

The author conceived the study, curated the dataset, implemented the model, conducted all training and evaluation experiments, and performed the analysis. No other students or collaborators were involved in this project. This work was conducted independently and is not part of any other research project or class assignment.

References

[1] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2623–2631, 2019.

[2] T. K. Behera, S. Bakshi, M. Nappi, and P. K. Sa. Superpixel-based multiscale cnn approach toward multiclass object segmentation from uav-captured aerial images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16:1771–1784, 2023.

[3] A. Boccalatte and J. Chanussot. Quantifying urban solar potential losses from rooftop superstructures via aerial imagery and convolutional neural networks. *Renewable Energy*, 249:123088, 2025.

[4] Esri. World imagery. `https://www.arcgis.com/home/group.html?id=702026e41f6641fb85da88efe79dc166`, 2024. Accessed: June 4th, 2025.

[5] X. Huang, L. Ren, C. Liu, Y. Wang, H. Yu, M. Schmitt, R. Hänsch, X. Sun, H. Huang, and H. Mayer. Urban building classification (ubc) – a dataset for individual building detection and classification from satellite imagery. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1412–1420, 2022.

[6] Z. Li, F. Lu, J. Zou, L. Hu, and H. Zhang. Generalized few-shot meets remote sensing: Discovering novel classes in land cover mapping via hybrid semantic segmentation framework, 2024.

[7] Y. Liu and N. Abbasabadi. Enhancing urban building energy models with vision transformers: A case study in material classification from google street view. *Energy and Buildings*, 333:115457, 2025.

[8] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

[9] T. Lüddecke and A. S. Ecker. Prompt-based multi-modal image segmentation. *CoRR*, abs/2112.10003, 2021.

[10] K. Mayer, L. Haas, T. Huang, J. Bernabé-Moreno, R. Rajagopal, and M. Fischer. Estimating building energy efficiency from street view imagery, aerial imagery, and land surface temperature data. *Applied Energy*, 333:120542, 2023.

[11] K. Mayer, B. Rausch, M.-L. Arlt, G. Gust, Z. Wang, D. Neumann, and R. Rajagopal. 3d-pv-locator: Large-scale detection of rooftop-mounted photovoltaic systems in 3d. *Applied Energy*, 310:118469, 2022.

[12] C. F. Reinhart and C. Cerezo Davila. Urban building energy modeling – a review of a nascent field. *Building and Environment*, 97:196–202, 2016.

[13] United Nations Environment Programme. 2021 global status report for buildings and construction: Towards a zero-emission, efficient and resilient buildings and construction sector. `https://globalabc.org/resources/publications/2021-global-status-report-buildings-and-construction` 2021. Accessed: 2025-04-09.

[14] K. Wada. Labelme: Image polygonal annotation with python, 2022.

[15] 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*, 190:196–214, Aug. 2022.

[16] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2020.

[17] S. Yi, X. Liu, J. Li, and L. Chen. Uavformer: A composite transformer network for urban scene segmentation of uav images. *Pattern Recognition*, 133:109019, 2023.

[18] H. Zhang, H. Dou, Z. Miao, N. Zheng, M. Hao, and W. Shi. Extracting building footprint from remote sensing images by an enhanced vision transformer network. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024.