INTERACTIVE FEW-SHOT ONLINE ADAPTATION FOR USER-GUIDED SEGMENTATION IN AERIAL IMAGES

Yeongsu Kim
Department of Economics
Jeonbuk National University
gsh05136@jbnu.ac.kr

Haeyun Lee
School of Computer Science and Engineering
Korea University of Technology & Education
haeyun.lee@koreatech.ac.kr

Kyungsu Lee
Department of Computer Science & Artificial Intelligence[*]
Jeonbuk National University
ksl@jbnu.ac.kr

ABSTRACT

Over the past few decades, geospatial objects have been extensively recognized as significant components in remote sensing applications, including environmental monitoring, urban planning, and defense. Particularly, accurate segmentation of objects has aimed at meaningful observations from aerial imagery, leading to the necessity of deep learning-based methodologies. However, conventional deep learning-based segmentation methodologies exhibit limited generalization capabilities across diverse geographical domains due to inherent variations in regional characteristics and data distribution shifts. Furthermore, most existing approaches strongly rely on static, pre-trained models lacking the adaptability to handle previously unseen data. To alleviate these limitations, we propose a novel Few-shot Semi-Online Adaptation framework incorporating interactive user feedback to iteratively refine segmentation outputs. By leveraging online learning and test-time adaptation, our approach enables models to continuously be accurate based on minimal user corrections, ensuring flexibility and adaptability to new environments. Experimental results demonstrate that our method effectively enhances the segmentation accuracy with minimal user intervention, bridging the gap between automated segmentation and domain-specific expertise. Our research contributes to the development of interactive, user-adaptive segmentation models to facilitate geospatial analysis more efficiently and reliably.

1 INTRODUCTION

Geospatial objects, particularly buildings, are structurally significant features in various remote sensing applications, including urban planning, disaster response, and environmental monitoring (Kaiser et al., 2017; Yi et al., 2019; Guo et al., 2019; Wang et al., 2021). Over the past decades, deep learning (DL)-based approaches have significantly improved segmentation accuracy by leveraging large-scale annotated aerial imagery. The strong improvements have facilitated automation in building segmentation, enabling large-scale geospatial analysis (Zhao et al., 2018; Liu et al., 2020; She, 2022; Lee et al., 2024). Despite these advancements, however, building segmentation remains challenging due to various factors inherent to aerial imagery (Liu et al., 2024; Memar et al., 2024; Wu et al., 2024). One of the primary challenges in DL-based building segmentation is the ambiguity in

[*] Corresponding Author

defining building boundaries, due to low spatial resolution, occlusions, and variations in architectural structures (Kim et al., 2018; Ye et al., 2021; Zhang et al., 2024b;a). Compared to natural objects with well-defined edges, buildings exhibit indistinct boundaries, resulting segmentation outputs highly subjective and dependent on the specific application. Furthermore, since the individuals recognize the differences in perception of defining building area, segmentation methodologies adapting to diverse environments and requirements have been necessitated (Benjdira et al., 2019; Wittich & Rottensteiner, 2021; Lee et al., 2021). Another significant limitation in existing DL-based segmentation methods is the vulnerability to domain shifts, due to variations in data acquisition conditions, sensor types, and geographical regions (Wang et al., 2021; Lee et al., 2021). Traditional DL models, typically trained offline (i.e., pre-trained) on specific datasets, experience significant performance degradation when applied to unseen domains (Lee et al., 2021; Min et al., 2023; Chen et al., 2023).

Recently, to address these challenges, TTA has been developed as robust technique, allowing pre-trained models to adapt dynamically using unlabeled test data before predictions or inference. TTA techniques, such as entropy minimization and self-training, have been widely studied in image classification (Liang et al., 2024; Wang et al., 2020), and recent works have demonstrated their effectiveness in segmentation tasks as well (Prabhudesai et al., 2023; Ma, 2024; Chen et al., 2024). However, most existing TTA approaches have still struggled with domain shifts in segmentation due to the reliance on pre-defined adaptation strategies, leading to limited generalization to diverse test scenarios. Most recently, beyond TTA, Online Learning has been explored as a methodologies of enabling DL models to incrementally update based on sequential input data. Unlike traditional batch learning, which requires full dataset re-training, online learning efficiently integrates new information into models, reducing computational costs and improving adaptability to real-time data streams (Hoi et al., 2021). Moreover, online learning has been successfully applied to segmentation tasks, demonstrating its potential for real-time adaptation and improved performance in dynamic environments (Zhao et al., 2017; Volpi et al., 2022). However, conventional online learning techniques fail to account for the subjective nature of building delineation as perceived by individual users, resulting in limited adaptability and suboptimal refinement of segmentation outputs in practical applications.

To address these challenges, we propose an interactive segmentation framework that integrates TTA and Semi-Online Learning, enabling user-driven refinement of segmentation predictions. Unlike conventional DL models that remain static after deployment, our approach incorporates an online learning mechanism where user corrections are encoded as feature embeddings and integrated into the processing pipeline of transformer. The iterative refinement process ensures that segmentation outputs align closely with human-defined criteria. Particularly, our framework exhibits the contribution of a hybrid architecture leveraging a transformer-based feature extractor with a variational autoencoder (VAE)-driven fusion module, facilitating adaptive segmentation refinement.Rather than focusing on mitigating occlusions or low-resolution imagery, our approach dynamically adjusts segmentation predictions in response to user feedback, continuously improving based on interactive corrections. By prioritizing human-guided refinement over rigid feature extraction strategies, our framework establishes a more adaptive and user-aligned segmentation paradigm.

2 Methods

2.1 Overall Architecture

The proposed framework integrates Test-Time Adaptation (TTA) and Semi-Online Learning, enabling iterative refinement of building segmentation predictions via user interactions. Unlike conventional models that remain static post-deployment, our approach dynamically updates segmentation outputs by incorporating user corrections as feature embeddings within a transformer-based processing pipeline. Given an input aerial image $\mathbf{I} \in \mathbb{R}^{H \times W \times C}$, the baseline segmentation network generates an initial prediction $\mathbf{S}_b = f_{\theta_b}(\mathbf{I})$, where f_{θ_b} represents the baseline segmentation model parameterized by θ_b. The segmentation logits are then converted into a softmax probability map as $\mathbf{P}_b = \text{softmax}(\mathbf{S}_b)$. To incorporate user-driven refinements, a transformer encoder extracts spatial features from both the input image and the initial segmentation output, producing a refined tokenized representation $\mathbf{T} = f_{\theta_t}(\mathbf{I}, \mathbf{P}_b)$, where f_{θ_t} represents the transformer-based feature extraction process. The final segmentation output is obtained by integrating $\mathbf{T}$ with VAE-generated latent features $\mathbf{Z}$, resulting in $\mathbf{S}_f = g_{\theta_v}(\mathbf{T}, \mathbf{Z})$, where g_{θ_v} denotes the VAE-based fusion module. Instead of automatically adjusting to domain shifts or low-resolution imagery, this fusion mechanism ensures that segmentation updates are directly influenced by human corrections, allowing the model to refine predictions in an interactive and user-guided manner.

2.2 Training and Inference Pipeline

The proposed framework follows a two-stage pipeline comprising offline pre-training and TTA-based adaptive learning, followed by real-time inference with user-driven refinements. In the offline pre-training phase, the baseline segmentation network is trained using a supervised learning approach on a large-scale aerial imagery dataset. To enhance local feature representations, we employ a patch-based supervision loss L_1, defined as:

$$L_1 = \sum_p \left(\mathcal{L}_{\text{CE}}(\mathbf{P}_b^p, \mathbf{Y}^p) + \lambda_{\text{dice}} \mathcal{L}_{\text{Dice}}(\mathbf{P}_b^p, \mathbf{Y}^p) \right), \tag{1}$$

where p indexes a patch in the image, $\mathcal{L}_{\text{CE}}$ represents the cross-entropy loss, and $\mathcal{L}_{\text{Dice}}$ denotes the Dice loss. The weight term λ_{dice} balances the contribution of Dice loss, ensuring improved segmentation precision at object boundaries. In the TTA-based adaptive learning phase, domain shift augmentation is applied by perturbing input images with contrast transformations, spatial distortions, and noise injections. The model adapts to these variations via contrastive learning techniques that reinforce feature alignment across different imaging conditions. Additionally, a modified triplet-inspired loss function L_2 is introduced to integrate user modifications effectively.

During inference, the framework employs real-time segmentation refinement via user interactions and adaptive learning mechanisms. Given an input aerial image, the baseline segmentation network produces an initial softmax probability map, which is subsequently tokenized and processed via the transformer-based encoder. The transformer enhances feature representations by capturing contextual cues from surrounding regions, leading to more precise object delineation. User corrections are incorporated via an interactive online learning mechanism. When a user refines the predicted segmentation mask, their modifications are encoded as additional feature embeddings and integrated into the transformer's processing pipeline. These user-driven embeddings dynamically influence subsequent segmentation predictions, ensuring that the model continuously aligns with human-defined segmentation criteria. To reinforce this adaptation, the triplet-inspired loss function L_2 is defined as:

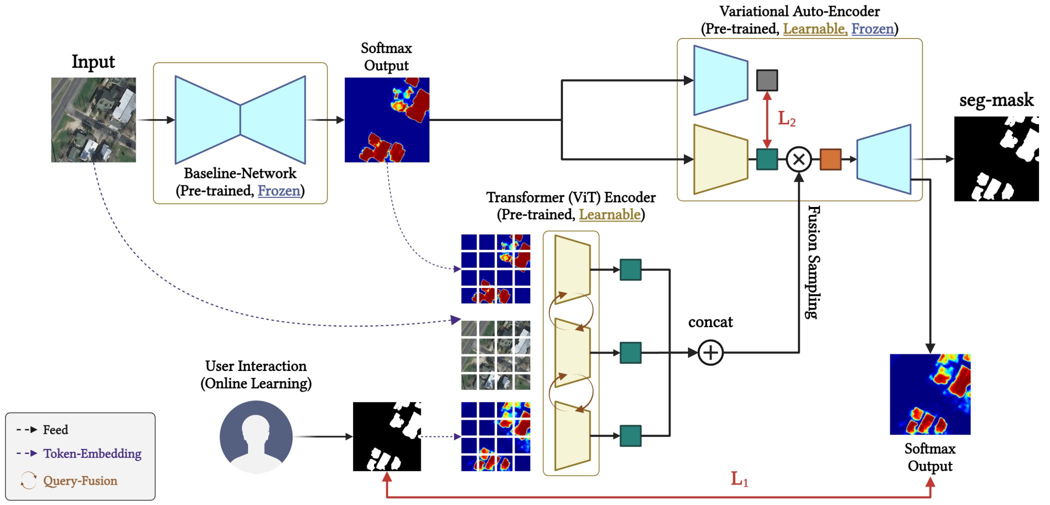

Figure 1: Proposed interactive segmentation framework, where a baseline network generates initial outputs refined by a transformer encoder with user corrections. A VAE-driven fusion module further enhances segmentation, optimized via losses L_1 and L_2.

$$L_2 = \sum_p \left(d(f_{\theta_v}(\mathbf{T}^p), \mathbf{Y}^p) - d(f_{\theta_v}(\mathbf{T}^p), \mathbf{Y}_u^p) + \alpha\right)_+ , \tag{2}$$

where $d(\cdot,\cdot)$ denotes a distance metric such as cosine similarity or Euclidean distance, and α is a margin hyperparameter. This loss function enforces that the refined segmentation $f_{\theta_v}(\mathbf{T}^p)$ is closer to the user-modified mask $\mathbf{Y}_u^p$ than to the original mask $\mathbf{Y}^p$, thereby incorporating user preferences into segmentation refinement. The final segmentation prediction is generated by passing the refined feature tokens via the VAE-driven fusion module, where a learned probabilistic sampling strategy produces robust segmentation masks. This process allows the model to adapt to occlusions, low-resolution imagery, and complex urban structures. A hierarchical attention mechanism further enhances segmentation performance by capturing multi-scale dependencies, enabling fine-grained segmentation in densely cluttered environments. To ensure long-term adaptability, the system incorporates a feedback loop that updates model parameters based on accumulated user corrections and environmental variations. By leveraging semi-online learning and test-time adaptation, our framework establishes a scalable and flexible solution for aerial building segmentation.

3 Experiments

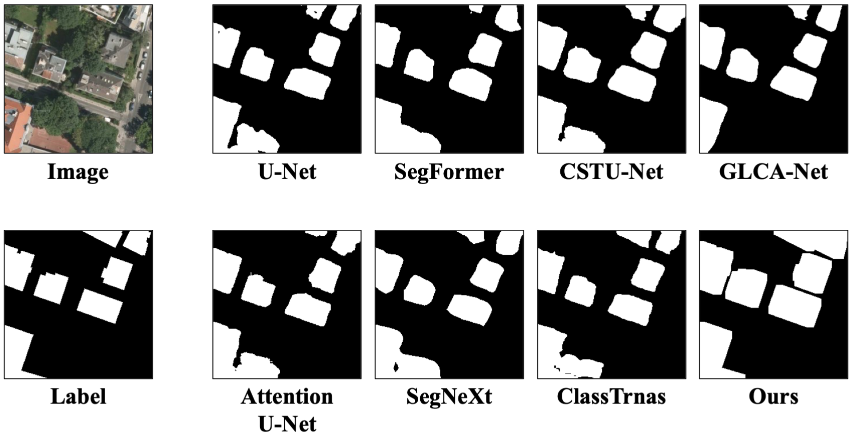

Figure 2: Representative results of our experiments compared to other models.

To evaluate the effectiveness of our interactive segmentation framework, we compare its performance against various baseline models. Table 1 presents the Intersection over Union (IoU) scores with 95% confidence intervals for different models, demonstrating a consistent improvement when incorporating our method. Notably, our approach achieves a significant performance boost across all architectures, with improvements ranging from approximately 5.4% (GLCANet) to 5.9% (ClassTrans) in IoU. Table 2 further analyzes the IoU, Precision, and Recall metrics under different model configurations. Here, Ours-N indicates the number of modified buildings with uer-interactions. As the number of user interactions increases, our framework progressively refines segmentation accuracy, with Ours-100 achieving the highest IoU of 88.87. Compared to TTA models, our method maintains higher recall and precision, highlighting the impact of user-guided refinements in improving segmentation quality. These results validate that incorporating user interactions effectively enhances segmentation accuracy beyond static deep learning models.

4 Conclusion

In this work, we introduced an interactive segmentation framework that incorporates TTA and Semi-Online Learning to iteratively refine building segmentation based on user interactions. Unlike conventional models that remain static after deployment, our approach dynamically updates segmentation outputs by integrating user corrections as additional feature embeddings within a transformer-based processing pipeline. The hybrid training pipeline employs a patch-based supervision loss to

Model	Baseline	Ours
U-Net	71.30 (71.00, 71.60)	77.28 (76.81, 77.75)
AttentionU-Net	68.75 (68.44, 69.06)	74.59 (74.10, 75.08)
SegFormer	72.91 (72.63, 73.19)	78.40 (77.99, 78.80)
SegNeXt	71.51 (71.22, 71.80)	77.91 (77.59, 78.23)
ClassTrans	79.38 (79.10, 79.66)	84.77 (84.39, 85.15)
CSTU-Net	79.25 (79.00, 79.50)	84.66 (84.27, 85.06)
GLCANet	71.58 (71.29, 71.87)	76.91 (76.48, 77.34)

Table 1: Comparison of IoU performance between baseline and our method with 95% C.I.

Model	IoU	Precision	Recall
Ours-1	76.92	84.10	90.01
Ours-10	83.34	87.48	94.63
Ours-100	88.87	92.57	95.69
TTA-1	82.42	86.20	94.94
TTA-2	80.67	87.66	91.01
TTA-3	80.06	84.41	93.94

Table 2: IoU, Precision, and Recall comparison for different model settings.

enhance feature extraction and a modified triplet loss to effectively incorporate user modifications, ensuring that segmentation predictions align with human-defined criteria. Rather than autonomously adapting to domain shifts or low-resolution imagery, our method prioritizes user-driven refinements to iteratively improve segmentation accuracy. Experimental results validate the effectiveness of our approach in refining segmentation outputs via interactive corrections. Future work will explore further optimizing the feedback mechanism and extending the framework to support multi-modal user interactions for broader remote sensing applications.

5 Acknowledgement

This research was supported by the MSIT(Ministry of Science, ICT), Korea, under the National Program for Excellence in SW), supervised by the IITP(Institute of Information & communications Technology Planing & Evaluation) in 2025 (2022-0-01067)

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pp. 265–283, 2016.

Mathilde Bateson, Herve Lombaert, and Ismail Ben Ayed. Test-time adaptation with shape moments for image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 736–745. Springer, 2022.

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.

Jie Chen, Peien He, Jingru Zhu, Ya Guo, Geng Sun, Min Deng, and Haifeng Li. Memory-contrastive unsupervised domain adaptation for building extraction of high-resolution remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023.

Ziyang Chen, Yongsheng Pan, Yiwen Ye, Mengkang Lu, and Yong Xia. Each test image deserves a specific prompt: Continual test-time adaptation for 2d medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11184–11193, 2024.

Lili Fan, Yu Zhou, Hongmei Liu, Yunjie Li, and Dongpu Cao. Combining swin transformer with unet for remote sensing image semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

Meng-Hao Guo, Cheng-Ze Lu, Qibin Hou, Zhengning Liu, Ming-Ming Cheng, and Shi-Min Hu. Segnext: Rethinking convolutional attention design for semantic segmentation. *Advances in Neural Information Processing Systems*, 35:1140–1156, 2022.

Zhiling Guo, Guangming Wu, Xiaoya Song, Wei Yuan, Qi Chen, Haoran Zhang, Xiaodan Shi, Mingzhou Xu, Yongwei Xu, Ryosuke Shibasaki, et al. Super-resolution integrated building semantic segmentation for multi-source remote sensing imagery. *IEEE Access*, 7:99381–99397, 2019.

Steven CH Hoi, Doyen Sahoo, Jing Lu, and Peilin Zhao. Online learning: A comprehensive survey. *Neurocomputing*, 459:249–289, 2021.

Jian Hu, Jiayi Lin, Shaogang Gong, and Weitong Cai. Relax image-specific prompt requirement in sam: A single generic prompt for segmenting camouflaged objects. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 12511–12518, 2024.

Minhao Hu, Tao Song, Yujun Gu, Xiangde Luo, Jieneng Chen, Yinan Chen, Ya Zhang, and Shaoting Zhang. Fully test-time adaptation for image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, pp. 251–260. Springer, 2021.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

Pascal Kaiser, Jan Dirk Wegner, Aurélien Lucchi, Martin Jaggi, Thomas Hofmann, and Konrad Schindler. Learning aerial image segmentation from online maps. *IEEE Transactions on Geoscience and Remote Sensing*, 55(11):6054–6068, 2017.

P Murphy Kevin. Machine learning: a probabilistic perspective, 2012.

Jun Hee Kim, Haeyun Lee, Seonghwan J Hong, Sewoong Kim, Juhum Park, Jae Youn Hwang, and Jihwan P Choi. Objects segmentation from high-resolution aerial images using u-net with pyramid pooling layers. *IEEE Geoscience and Remote Sensing Letters*, 16(1):115–119, 2018.

Kyungsu Lee, Haeyun Lee, and Jae Youn Hwang. Self-mutating network for domain adaptive segmentation in aerial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7068–7077, 2021.

Kyungsu Lee, Haeyun Lee, Juhum Park, and Jae Youn Hwang. Fine-grained binary segmentation for geospatial objects in remote sensing imagery via path-selective test-time adaptation. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

Jian Liang, Ran He, and Tieniu Tan. A comprehensive survey on test-time adaptation under distribution shifts. *International Journal of Computer Vision*, pp. 1–34, 2024.

Guangjie Liu, Kuo Diao, Jinlong Zhu, Qi Wang, and Meng Li. Stransu2net: Transformer based hybrid model for building segmentation in detailed satellite imagery. *PloS one*, 19(9):e0299732, 2024.

Zhongwei Liu, Baisong Chen, and Ao Zhang. Building segmentation from satellite imagery using u-net with resnet encoder. In *2020 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE)*, pp. 1967–1971. IEEE, 2020.

I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Jiahao Lu, Jiacheng Deng, and Tianzhu Zhang. Beyond the final layer: Hierarchical query fusion transformer with agent-interpolation initialization for 3d instance segmentation. *arXiv preprint arXiv:2502.04139*, 2025.

Jing Ma. Improved self-training for test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23701–23710, 2024.

Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 2017.

Babak Memar, Luigi Russo, and Silvia Liberata Ullo. A u-net architecture for building segmentation through very high resolution cosmo-skymed imagery. In *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*, pp. 4653–4657. IEEE, 2024.

Jeongho Min, Yejun Lee, Dongyoung Kim, and Jaejun Yoo. Bridging the domain gap: A simple domain matching method for reference-based image super-resolution in remote sensing. *IEEE Geoscience and Remote Sensing Letters*, 2023.

Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

Mihir Prabhudesai, Anirudh Goyal, Sujoy Paul, Sjoerd Van Steenkiste, Mehdi SM Sajjadi, Gaurav Aggarwal, Thomas Kipf, Deepak Pathak, and Katerina Fragkiadaki. Test-time adaptation with slot-centric models. In *International Conference on Machine Learning*, pp. 28151–28166. PMLR, 2023.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pp. 234–241. Springer, 2015.

Yongyang She. Building instance segmentation in high-resolution remote sensing images based on multi-task learning. In *2022 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA)*, pp. 1517–1520. IEEE, 2022.

Riccardo Volpi, Pau De Jorge, Diane Larlus, and Gabriela Csurka. On the road to online adaptation for semantic image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19184–19195, 2022.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.

Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv preprint arXiv:2110.08733*, 2021.

Shihong Wang, Ruixun Liu, Kaiyu Li, Jiawei Jiang, and Xiangyong Cao. Class similarity transition: Decoupling class similarities and imbalance from generalized few-shot segmentation. *arXiv preprint arXiv:2404.05111*, 2024.

Dennis Wittich and Franz Rottensteiner. Appearance based deep domain adaptation for the classification of aerial images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 180:82–102, 2021.

Yingbin Wu, Peng Zhao, Fubo Wang, Mingquan Zhou, Shengling Geng, and Dan Zhang. A prior-guided dual branch multi-feature fusion network for building segmentation in remote sensing images. *Buildings*, 14(7):2006, 2024.

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.

Huanran Ye, Sheng Liu, Kun Jin, and Haohao Cheng. Ct-unet: An improved neural network based on u-net for building segmentation in remote sensing images. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 166–172. IEEE, 2021.

Yaning Yi, Zhijie Zhang, Wanchang Zhang, Chuanrong Zhang, Weidong Li, and Tian Zhao. Semantic segmentation of urban buildings from vhr remote sensing imagery using a deep convolutional neural network. *Remote sensing*, 11(15):1774, 2019.

Jinglin Zhang, Yuxia Li, Zhonggui Tong, Lei He, Mingheng Zhang, Zhenye Niu, and Haiping He. Glcanet: Global–local context aggregation network for cropland segmentation from multi-source remote sensing images. *Remote Sensing*, 16(24):4627, 2024a.

Zipeng Zhang, Wei Chen, Weiwei Guo, Yiming Liu, Jianhua Yang, and Houguang Liu. Cst-unet: Cross swin transformer enhanced u-net with masked bottleneck for single-channel speech enhancement. *Circuits, Systems, and Signal Processing*, pp. 1–22, 2024b.

Kang Zhao, Jungwon Kang, Jaewook Jung, and Gunho Sohn. Building extraction from satellite images using mask r-cnn with building boundary regularization. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 247–251, 2018.

Ying Zhao, Zhiwei Luo, and Changqin Quan. Unsupervised online learning for fine-grained hand segmentation in egocentric video. In *2017 14th Conference on Computer and Robot Vision (CRV)*, pp. 248–255. IEEE, 2017.

A Related Works

A.1 Test-time Adaptation

Machine learning aims to train robust models capable of generalizing effectively to test samples, even in the presence of distribution shifts. However, these models often experience performance degradation due to uncertainties in test distributions. TTA enables pre-trained models to adapt using unlabeled test data before making predictions, thereby mitigating such degradation (Liang et al., 2024). Previous research on TTA has primarily focused on image classification tasks, employing techniques such as entropy minimization, self-training, and batch normalization statistics adaptation to enhance test-time performance Wang et al. (2020). Test-Time Adaptation has also demonstrated its effectiveness in segmentation tasks. (Prabhudesai et al., 2023) introduced slot-TTA, which leverages slow inference to significantly improve segmentation performance in out-of-distribution (OOD) scenarios. Ma (2024) proposed the Improved Self-Training (IST) method, achieving performance enhancements not only in classification but also in detection and segmentation tasks. Furthermore, (Chen et al., 2024) presented a novel approach that utilizes prompts tailored to individual test images, demonstrating performance improvements in continual test-time adaptation settings for segmentation. These studies indicate that TTA can substantially enhance segmentation models, particularly in adapting to domain shifts and novel environments.

A.2 Online Learning

Traditional machine learning paradigms predominantly operate in a batch learning or offline learning manner, particularly in supervised learning. These approaches involve training models using an entire dataset at once, followed by deployment for inference without further updates. However, this paradigm entails high retraining costs when processing new data, limiting scalability in real-world applications. As batch learning becomes increasingly restrictive, adapting machine learning models to continuously evolving data streams has emerged as a critical challenge in the field of artificial intelligence. Unlike traditional machine learning, online learning is a subfield that enables models to incrementally learn from sequentially arriving data. This approach overcomes the limitations of batch learning by allowing efficient and immediate model updates upon receiving new training samples Hoi et al. (2021). Online learning has been extensively studied in various machine learning domains, including segmentation tasks. (Zhao et al., 2017) proposed an online learning-based hand segmentation method, while (Volpi et al., 2022) introduced a new protocol for continuously learning semantic segmentation from image sequences on a frame-by-frame basis. These studies demonstrate the integration of online learning with segmentation, highlighting its potential for real-time adaptation, efficient learning, and improved performance in dynamic environments.

A.3 Query-Fusion

Query-fusion has been widely explored in segmentation tasks, particularly in transformer-based architectures, where it plays a crucial role in feature aggregation and attention-based refinement. Existing approaches commonly utilize query embeddings to capture contextual dependencies across spatial and semantic features, enhancing segmentation accuracy. Particularly, the attention-based query mechanisms in transformer models facilitated that query tokens interact with key-value pairs to refine feature representations dynamically.

Recent works have demonstrated the effectiveness of query-based feature fusion in deep learning models for segmentation (Lu et al., 2025). Transformers leverage self-attention to propagate information across different spatial regions, allowing for better contextual understanding. In particular, query-based fusion mechanisms enable models to aggregate multi-scale features and refine object boundaries by dynamically selecting relevant spatial information. These methods have been widely applied in semantic and instance segmentation tasks, where refining segmentation masks based on attention-weighted query embeddings has shown significant improvements in performance.

B Motivation

Accurate building segmentation in aerial imagery is a fundamental task in remote sensing and urban analysis. While deep learning models have significantly improved segmentation performance, existing methods often struggle to generalize across diverse geographic regions and imaging conditions. A major challenge should be in the inherent variability in human perception when defining building boundaries, leading to inconsistencies in automated segmentation. This paper addresses these challenges by introducing an interactive segmentation framework that refines predictions based on user-defined criteria via a query-fusion mechanism. Traditional deep learning-based segmentation models typically rely on static feature extraction pipelines trained on large-scale datasets. However, the conventional models face significant limitations when applied to real-world scenarios, where domain shifts, occlusions, and variations in imaging conditions degrade segmentation performance. Previous methods have attempted to mitigate the significant issues via domain adaptation and test-time augmentation, but such approaches often fail to fully align with user-defined segmentation requirements. Furthermore, conventional models lack the ability to incorporate human corrections post-deployment, making them less effective in practical applications where subjective interpretation plays a role in defining building structures.

B.1 Variations of Cognitive Recognition in Defining Buildings

Defining building boundaries is inherently subjective, as different users may perceive and annotate structures differently based on contextual or cognitive biases. Factors such as occlusions, roof patterns, and shadow effects introduce further ambiguity in aerial imagery segmentation. Existing segmentation models rely on fixed ground-truth annotations, which may not always reflect the most contextually accurate representations of buildings. This variability in human cognition necessitates a more flexible segmentation approach that can adapt to individual user preferences, enabling more precise and interpretable segmentation outputs.

B.2 Query-Fusion

To address these challenges, our method extends traditional segmentation approaches by integrating user interactions into the query-fusion process. Unlike conventional query-based segmentation, which primarily relies on learned priors and spatial relationships, our framework dynamically encodes user corrections as query embeddings within a transformer-based processing pipeline. This interactive refinement process allows real-time modifications to influence segmentation predictions, ensuring that updates align with human-defined segmentation criteria rather than relying solely on pre-trained feature extraction. By incorporating user-driven query-fusion, our approach bridges the gap between static deep learning models and interactive segmentation frameworks. Unlike existing query-based attention mechanisms, which are typically optimized for autonomous feature selection, our method directly integrates human feedback, making it more adaptable to dynamically evolving segmentation tasks. This ensures that the segmentation model remains flexible and responsive to user-defined modifications, ultimately improving segmentation accuracy and interpretability in practical applications.

C Methods

C.1 Training Pipeline

Our proposed framework follows a structured training pipeline that aligns with the architecture depicted in Fig. 2. The training process ensures consistency across all components while introducing a mechanism to simulate user interactions without explicit manual annotations. To achieve this, we employ a data augmentation strategy where the ground truth segmentation masks are randomly perturbed using morphological transformations. These modifications mimic potential user corrections, enabling the model to learn from dynamically adjusted segmentation masks. Instead of direct user interactions, these perturbed masks serve as interactive refinements during training.

Initially, an input aerial image is processed via a frozen pre-trained baseline segmentation network, generating an initial softmax probability output. This output, along with the perturbed ground truth mask, is then tokenized and passed via a transformer-based encoder, which extracts spatial and contextual features. The transformer embeddings are concatenated and fed into a variational autoencoder (VAE)-driven fusion module, refining the final segmentation prediction. The optimization process leverages three key loss functions: (1) a patch-based supervision loss L_1, which ensures local feature consistency; (2) a triplet-inspired loss L_2, designed to align the refined segmentation output with the modified ground truth while preserving structural accuracy; (3) supervision between final output segmentaiton mask and augmented ground truth. This pipeline enables the model to adapt to variations in ground truth annotations while maintaining consistency across training iterations.

C.2 Inference Pipeline

During inference, the framework efficiently refines segmentation predictions without requiring retraining of the baseline segmentation model. Instead of relying on additional model parameters, the refinement process is guided by user-provided corrections, which are incorporated as query embeddings in the transformer processing pipeline. The pre-trained baseline model remains frozen, ensuring computational efficiency by eliminating redundant re-training. Given an input image, the initial segmentation mask is generated and tokenized, followed by feature extraction via the transformer encoder. User interactions, represented as manual segmentation corrections, are integrated into the query-fusion process. These corrections dynamically adjust the transformer embeddings, allowing the model to refine predictions iteratively. Finally, the refined segmentation output is generated via the VAE-based fusion module, ensuring that the final segmentation aligns with user-defined criteria. By leveraging a lightweight adaptation mechanism without modifying the baseline model's parameters, our method maintains efficiency while providing flexible, user-driven segmentation refinement.

D Experiments

D.1 Datasets

To evaluate the segmentation performance of the proposed method, we utilized two publicly available datasets: the Inria Aerial Image Labeling Dataset (Maggiori et al., 2017) and the LoveDA Dataset (Wang et al., 2021). The Inria dataset consists of high-resolution aerial images captured over urban and suburban areas, providing detailed building annotations for segmentation tasks. LoveDA offers a diverse set of satellite imagery covering urban and rural landscapes, making it suitable for assessing model generalization across varying environmental conditions. Each dataset was split into five subsets following a $k = 5$ cross-validation scheme. Three subsets were used for training, one for validation, and one for testing. User interactions were simulated by introducing manual corrections to segmentation masks, allowing evaluation of the adaptability of the models to human-defined refinements.

D.2 Experimental Environment

The proposed framework was implemented using TensorFlow (Abadi et al., 2016) (ver. 2.12.0) and PyTorch (ver. 2.4.0), maintaining consistency in experimental parameters across all models. The mini-batch size was set to 8, and input images were resized to 256×256. Model training was conducted using the AdamW optimizer (Loshchilov, 2017) with batch normalization (Ioffe & Szegedy, 2015) applied to stabilize convergence. To ensure robust evaluation, we applied a k-fold cross-validation approach with $k = 5$ (Kevin, 2012). The datasets used include the Inria dataset and the LoveDA dataset. Each dataset was split into five subsets, where three splits were used for training, one for validation, and one for testing. For real-time application deployment, the front-end was implemented using React (ver. 18.3.1), and the back-end utilized Flask (ver. 3.0.3). The framework was developed using Python (ver. 3.8.19) and Node.js (ver. 20.18.0), with specific version details omitted for brevity.

All models were evaluated under the same experimental conditions using the Inria and LoveDA datasets for cross-dataset validation. The models trained on the Inria dataset were validated on LoveDA, and vice versa, ensuring that segmentation performance was tested across diverse geographic regions and imaging conditions. Performance was assessed using IoU, precision, and recall metrics, ensuring a fair comparison between interactive segmentation refinement and non-interactive adaptation approaches. Unlike existing methods, which focus solely on automatic adaptation, our framework integrates user-driven refinements, enabling segmentation updates based on human-defined criteria without requiring extensive retraining. By embedding multiple segmentation architectures, systematically analyzing the impact of user interactions, and benchmarking against state-of-the-art adaptation techniques, our experimental setup provides a rigorous evaluation of the effectiveness and adaptability of our interactive query-fusion framework. To evaluate the effectiveness of our interactive segmentation framework, we conducted extensive ablation studies and comparative analyses against SotA TTA models.

D.3 Experimental Settings

To evaluate the generalization capability of our proposed method, we conducted cross-dataset validation using the Inria and LoveDA datasets. The Inria dataset consists of high-resolution aerial images primarily covering urban environments, while the LoveDA dataset contains diverse scenes, including both urban and rural landscapes with varying spatial resolutions and domain characteristics. To assess the robustness of our method to domain shifts, we trained the model on the Inria dataset and validated it on LoveDA, and vice versa. This approach allows us to analyze how well the segmentation framework adapts to unseen spatial distributions and different environmental contexts. The model trained on Inria was tested on LoveDA, where it encountered more diverse geographic structures, while the model trained on LoveDA was evaluated on Inria to determine its ability to generalize to high-resolution urban imagery.

To facilitate interactive segmentation refinement, we implemented a web-based map application using React for the front-end and Flask for the back-end. The application enables users to visualize segmentation outputs overlaid on aerial imagery and provides multiple interaction modes for refining segmentation masks. Users can modify the segmentation results via two primary mechanisms: (1) a *polygon selection* tool for outlining misclassified regions, and (3) a *click-based refinement* method where users indicate incorrect segmentation regions for automated correction. These user corrections are incorporated into the query-fusion mechanism, dynamically adjusting the segmentation predictions based on real-time human feedback. During inference, the system processes user inputs by embedding corrections into the transformer-based processing pipeline. Instead of retraining the entire segmentation model, only the query-fusion module is updated, ensuring computational efficiency while enabling interactive refinement. The web application records user interactions, allowing for quantitative evaluation of segmentation performance across iterative refinements. The front-end, built with React (ver. 18.3.1), communicates with the back-end Flask (ver. 3.0.3) server, which processes segmentation updates in real-time as depicted in Fig. A-1.

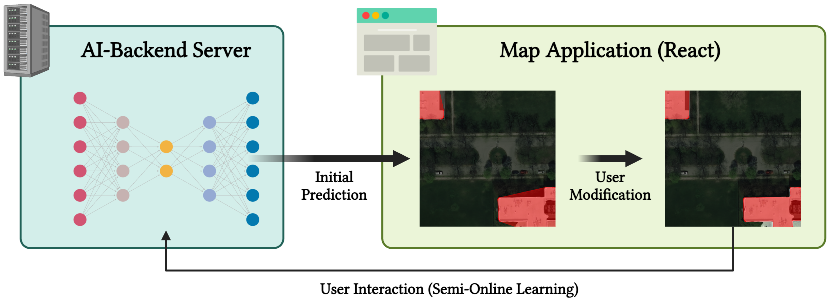

Figure A-1: Pipeline of user interaction with online map applications.

By integrating cross-dataset evaluation with real-time user interaction, our experimental setup validates the effectiveness of query-fusion for dynamic segmentation refinement. The results demonstrate that incorporating human feedback allows for improved segmentation accuracy without requiring extensive retraining, highlighting the advantages of our interactive adaptation approach.

D.4 Ablation Study

Our implementation integrates multiple baseline segmentation architectures, systematically examines the impact of user interactions, and benchmarks our approach under controlled experimental conditions. The proposed framework embeds a diverse set of baseline segmentation models, including convolutional and transformer-based architectures. Specifically, we implemented U-Net, AttentionU-Net, SegFormer, SegNeXt, ClassTrans, CSTU-Net, and GLCANet as backbone networks to assess the compatibility of our query-fusion mechanism with different model architectures. U-Net (Ronneberger et al., 2015) is a widely adopted convolutional model featuring an encoder-decoder structure with skip connections, facilitating precise boundary delineation. Attention U-Net (Oktay et al., 2018) enhances this design by incorporating attention mechanisms, improving feature selection in complex scenes. However, both models struggle with long-range dependencies in aerial imagery. Transformer-based architectures, such as SegFormer (Xie et al., 2021) and SegNeXt (Guo et al., 2022), leverage self-attention mechanisms to capture global spatial relationships, yielding improved segmentation robustness across varying scales. ClassTrans (Wang et al., 2024) introduces class-aware tokens to enhance segmentation consistency, while CSTU-Net (Fan et al., 2023) combines convolutional and transformer components for hybrid feature extraction. GLCANet (Zhang et al., 2024a) further refines segmentation by aggregating global-local contextual information.

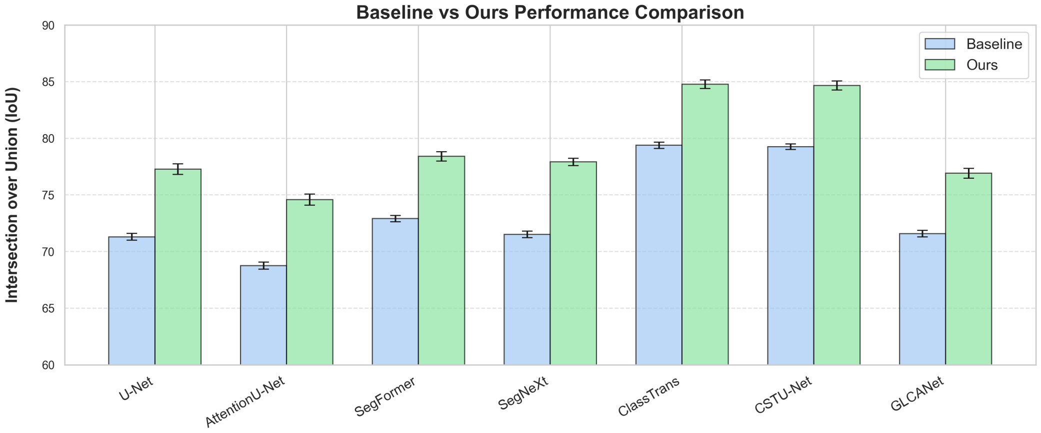

Figure A-2: IoU values of our framework compared to baseline models.

Table 1 and appendix Fig. A-2 present the results of IoU values, demonstrating our method consistently improves segmentation accuracy across all baseline models. The results indicate that applying query-fusion for user-guided refinement leads to a significant IoU improvement. On average, our method achieves a performance gain of approximately 5.57% across all models, demonstrating the effectiveness of integrating interactive segmentation refinements. Notably, transformer-based models such as ClassTrans and CSTU-Net exhibit the highest improvements, with IoU increases of 5.39% and 5.41%, respectively. This suggests that self-attention mechanisms effectively capture refined spatial dependencies when integrated with user-driven corrections. Traditional convolutional models like U-Net and Attention U-Net also benefit from the refinement process, with IoU improvements of 5.98% and 5.84%, respectively. Despite their reliance on local receptive fields, incorporating user modifications via query-fusion enhances their ability to refine segmentation boundaries. Additionally, hybrid architectures such as SegFormer and SegNeXt achieve substantial performance boosts of 5.49% and 6.40%, indicating that multi-scale feature aggregation synergizes well with interactive corrections. The lowest performance improvement is observed in GLCANet (5.33%), which suggests that its global-local feature aggregation approach already captures significant spatial relationships. However, even in this case, user-guided refinements still provide measurable benefits in segmentation accuracy.

Overall, these findings validate that our method effectively enhances segmentation accuracy across diverse architectural paradigms. By integrating interactive corrections, query-fusion dynamically refines segmentation outputs, surpassing static deep learning models that rely solely on pre-trained feature extraction. The consistent performance gains across different architectures highlight the robustness and adaptability of our approach in improving segmentation precision.

D.5 Comparative Analysis

To assess the contribution of interactive learning, we conducted an ablation study by varying the level of user interaction during segmentation refinement. We simulated user corrections by modifying segmentation masks for different numbers of buildings, the number of buildings from no modification (0-buildings) to full user correction (100-buildings). These perturbations were applied via controlled morphological transformations to emulate real-world user adjustments. We evaluated the effectiveness of interactive refinements by measuring segmentation performance at different levels of user interaction, quantifying the impact of query-fusion in dynamically refining segmentation outputs. Additionally, we explored different embedding strategies for integrating user corrections within the transformer pipeline, ensuring that the most effective approach was selected for final deployment. For comparative analysis, we compare it against three state-of-the-art test-time adaptation and segmentation methods, each addressing domain adaptation, shape-guided refinement, and prompt-based segmentation. Hu et al. (2021) mitigates domain shift effects by adapting a pre-trained model using unlabeled target data. It updates only batch normalization layers while employing Regional Nuclear-norm (RN) and Contour Regularization (CR) losses to improve segmentation consistency. This method has demonstrated effectiveness in pancreas and liver segmentation tasks. Bateson et al. (2022) introduces a shape-guided entropy minimization approach that optimizes batch normalization parameters during inference. Without requiring target domain training, it ensures segmentation alignment with structural priors, achieving superior performance in MRI-to-CT cardiac segmentation and cross-site prostate segmentation. Generalizable SAM (GenSAM) (Hu et al., 2024) extends prompt-based segmentation by eliminating manual prompts via Cross-modal Chains of Thought Prompting (CCTP). By leveraging vision-language models, it generates automatic visual prompts for camouflaged object detection, refining segmentation iteratively via Progressive Mask Generation (PMG).

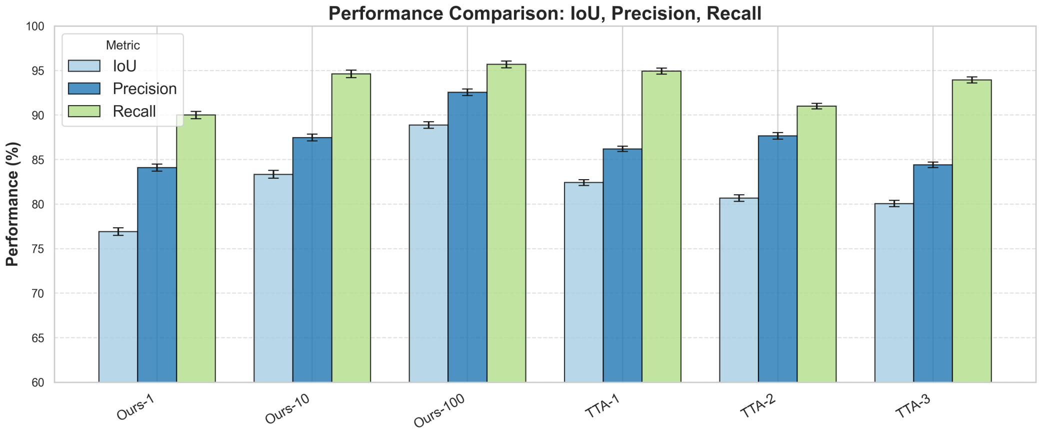

Figure A-3: Evaluation metric values of our framework compared to SotA models.

Our method consistently outperforms TTA models in terms of Intersection over Union (IoU), Precision, and Recall, with performance improving as the number of user interactions increases as illustrated in Fig. A-3. Ours-1, which incorporates minimal user feedback, achieves an IoU of 76.92, which is lower than TTA-1 (82.42) but demonstrates comparable precision (84.10) and recall (90.01). As user interactions increase, segmentation accuracy improves significantly, with Ours-10 achieving an IoU of 83.34, surpassing all TTA models. The most refined segmentation output, Ours-100, achieves the highest IoU (88.87), Precision (92.57), and Recall (95.69), demonstrating the benefit of iterative user-driven refinements. In contrast, TTA models exhibit varying levels of adaptation effectiveness. TTA-1 performs the best among TTA approaches, achieving an IoU of 82.42, but its recall (94.94) suggests that it over-segments certain regions, leading to a potential increase in false positives. TTA-2 and TTA-3 show slightly lower IoU scores (80.67 and 80.06, respectively), indicating that automatic test-time adaptation methods struggle to generalize optimally across all segmentation scenarios without explicit user guidance.

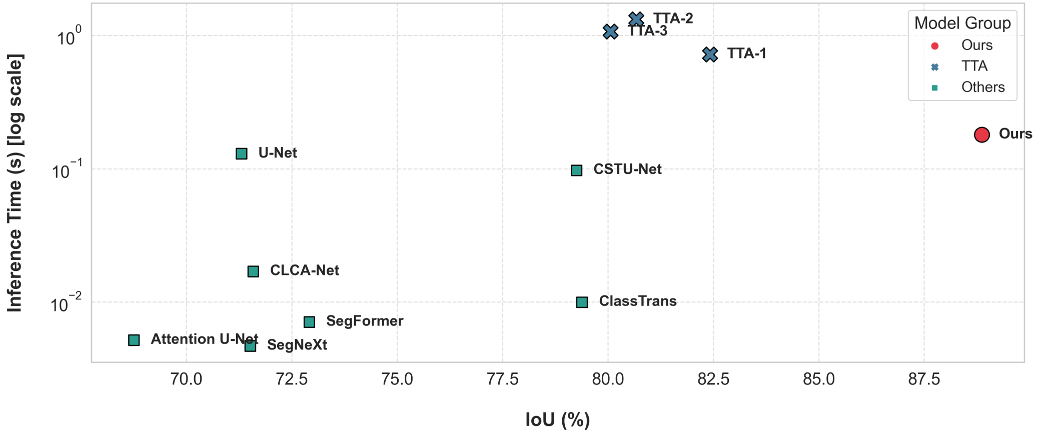

Figure A-4: Evaluation metric values of our framework compared to SotA models.

To further assess the effectiveness of our interactive segmentation framework, we compared its performance against SotA segmentation models and TTA methods in terms of IoU performance and inference time, as shown in Fig. A-4. The x-axis represents IoU (%), measuring segmentation accuracy, while the y-axis, in logarithmic scale, represents inference time (s), indicating computational efficiency. The results clearly demonstrate that our proposed approach achieves the highest IoU among all models while maintaining an efficient inference time. Compared to TTA methods (TTA-1, TTA-2, TTA-3), which show longer inference times due to their adaptation processes, our method significantly reduces computational overhead while improving segmentation accuracy. TTA models, though improving IoU over static models, require substantial inference time, making them less suitable for real-time applications. Among the baseline segmentation models, transformer-based architectures such as ClassTrans and CSTU-Net achieve relatively high IoU scores but exhibit slower inference times compared to lighter convolutional models like SegFormer and SegNeXt. CNN-based models, including U-Net and AttentionU-Net, have lower inference times but struggle to maintain high segmentation accuracy. Our method, positioned in the upper-right region of the graph, demonstrates a strong balance between accuracy and efficiency, outperforming SotA models in segmentation quality while avoiding the high computational costs associated with TTA-based approaches.

The results clearly demonstrate that incorporating user interactions significantly enhances segmentation accuracy beyond what is achievable via automated adaptation methods alone. While TTA approaches provide moderate improvements via batch normalization tuning and entropy minimization, they lack the ability to iteratively refine segmentation outputs based on human-defined criteria. Our method, by leveraging user-driven query-fusion, continuously refines segmentation boundaries, leading to higher accuracy and better precision-recall balance. Overall, these findings validate the advantage of interactive refinement over static adaptation, reinforcing the importance of integrating user feedback into segmentation frameworks. The ability to dynamically adjust segmentation predictions based on human corrections enables our approach to achieve superior generalization and precision across diverse imaging conditions.