Semi-Automated Generation of Object Mask Annotations from Bounding Box Labels Using SAM with Domain-Aware Geometric Constraints

SungWon Moon
Content Research Division
Electronics and Telecommunications Research Institute(ETRI)
Daejeon, Republic of Korea
moonstarry@etri.re.kr

Jiwon Lee
Content Research Division
Electronics and Telecommunications Research Institute(ETRI)
Daejeon, Republic of Korea
ez1005@etri.re.kr

Jungsoo Lee
Content Research Division
Electronics and Telecommunications Research Institute(ETRI)
Daejeon, Republic of Korea
jslee2365@etri.re.kr

Dowon Nam
Content Research Division
Electronics and Telecommunications Research Institute(ETRI)
Daejeon, Republic of Korea
dwnam@etri.re.kr

Seungjae Lee
Content Research Division
Electronics and Telecommunications Research Institute(ETRI)
Daejeon, Republic of Korea
seungjlee@etri.re.kr

***Abstract*— Accurate object segmentation is vital for training and evaluating computer vision models, yet manual mask annotation remains a major bottleneck, especially for domain-specific datasets like traditional paintings. This paper proposes a semi-automated framework that transforms bounding box labels into object masks using MobileSAM, a lightweight prompt-based segmentation model. Full-image candidate masks are generated and refined through geometric constraints, including bounding box containment, relative size thresholds, and spatial alignment. We evaluate candidate masks using Intersection over Union (IoU) metrics and apply domain-aware filtering heuristics to select the most accurate masks. Experiments on traditional painting datasets show that the framework yields high-quality segmentation masks with minimal human intervention, offering a scalable and efficient solution for enhancing annotation pipelines in specialized visual domains.**

Keywords*— *Weakly Supervised Segmentation, Domain-Aware Annotation, Automatic Mask Generation

I. INTRODUCTION

High-quality object masks are essential for a wide range of computer vision applications [1]-[5]. These masks provide fine-grained supervision for training deep learning models for tasks such as object detection, instance segmentation, and semantic scene understanding. Unlike coarse annotations, such as bounding boxes, pixel-level masks allow models to learn the precise shapes, boundaries, and spatial structures of objects. In addition to training models, segmentation masks play a critical role in downstream applications, including visual saliency analysis, interactive editing, and domain-specific object tracking, where precise localization is necessary.

Various weakly supervised methods have been proposed to reduce the cost and effort of manual annotation and to generate object masks. For example, BoxInst [1] demonstrates that instance segmentation can be learned from bounding box annotations by incorporating projection-based and pairwise color affinity losses. Recently, foundation models such as the Segment Anything Model (SAM) [6] have enabled prompt-based segmentation with minimal supervision, offering broad applicability across diverse visual domains. MobileSAM [7], a lightweight variant of SAM, further improves efficiency by significantly reducing model size and inference latency. These advances present promising solutions for scalable mask generation and automatic dataset construction.

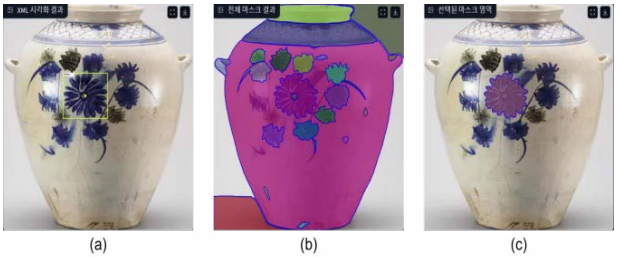

Fig. 1. Example results of (a) bounding box annotations, (b) segmentation masks generated using MobileSAM, and (c) final object masks selected and refined through geometric filtering

Nevertheless, applying these approaches to traditional paintings presents unique challenges. Traditional artworks vary greatly in artistic style, abstraction, and color schemes. They are also often affected by image degradation resulting from aging or poor preservation. These factors violate key assumptions of conventional segmentation models, such as the presence of consistent textures or well-defined object boundaries. Additionally, the symbolic and stylized nature of painted objects complicates the use of low-level visual cues, such as color affinity and contour continuity. Consequently, domain-aware segmentation techniques are necessary to address the visual ambiguity and stylistic diversity inherent in traditional paintings. In this work, we introduce a technique for automatically generating object masks for specific visual elements, as shown in Fig. 1. The method uses bounding box annotations embedded in cultural heritage data and combines them with segmentation masks derived from the entire image. Appropriate filtering is then applied to extract object masks that match the target objects.

The remainder of this paper is organized as follows. Section II reviews related work and Section III introduces the proposed method, which combines bounding box annotations with full-image masks and domain-aware filtering. Section IV presents the experimental results, and Section V concludes with future directions.

Fig. 2. Comparison of mask selection results without IoU-based filtering (left) and with IoU-based filtering (right)

II. Related Work

A. Object Mask Generation from Bounding Boxes

Several recent studies have examined the generation of object masks from bounding box annotations. One of the most prominent approaches is BoxInst [1]. Built upon the CondInst framework, BoxInst eliminates the need for pixel-level supervision by introducing projection and color affinity losses. Although BoxInst performs well on general datasets such as Pascal VOC and COCO, its reliance on color-based affinity can result in suboptimal segmentation in domain-specific settings where object boundaries are ambiguous or have low contrast. For example, in medical imaging, these limitations have been shown to reduce segmentation quality, suggesting the challenge of applying BoxInst to specialized domains. To address this issue, Box2Mask [4] proposed a level set–based contour optimization method that improves boundary accuracy, though it introduces additional computational overhead. In contrast, our approach emphasizes domain adaptability and scalability by leveraging prompt-based segmentation with lightweight architectures.

B. MobileSAM

SAM [6] has recently emerged as a foundation model for promptable image segmentation across a wide range of domains. Leveraging a powerful image encoder and a flexible prompt interface (e.g., points, boxes, or masks), SAM enables zero-shot or few-shot segmentation with high accuracy. However, its substantial computational demands, stemming from its large, ViT-based encoder and decoder architecture, limit its practicality for large-scale annotation tasks or deployment in environments with limited resources.

To address these limitations, MobileSAM [7] was introduced as a lightweight variant of SAM [6]. MobileSAM replaces the original ViT backbone with MobileViT and uses a distillation strategy to transfer knowledge from SAM. This allows it to retain competitive segmentation performance while significantly reducing inference time and model size. These characteristics make MobileSAM particularly well-suited for high-throughput object mask generation pipelines. In this work, we adopt MobileSAM as the core segmentation module to efficiently and scalably convert bounding box annotations into pixel-level masks.

III. Methodology

A. Overview

The overall pipeline of the proposed method is illustrated in Fig. 1. Given a cultural heritage image and its associated bounding box annotations, such as those derived from XML labels or object detection outputs, the objective is to automatically generate high-quality segmentation masks for individual objects. The pipeline comprises two main stages: selecting candidate masks generated by MobileSAM [7] and refining them through geometric filtering designed for the characteristics of traditional paintings.

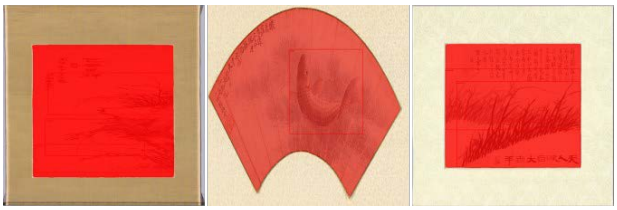

Fig. 3. Example of inaccurate masks filtered out using geometric constraints

In the first stage, MobileSAM [7] generates segmentation masks for the entire image without using bounding box prompts. For each bounding box, the intersection-over-union (IoU) is computed between the box and all candidate masks, and the mask with the highest IoU is selected as the initial object mask. As illustrated in Fig. 2, this step is critical: without proper constraints, MobileSAM may yield masks that cover excessive or irrelevant regions, such as the entire vase instead of a specific target like a flower.

To improve localization accuracy, a second filtering stage applies domain-aware geometric constraints. Fig. 3 shows cases where masks that significantly overflow the bounding box or occupy disproportionate areas are filtered out. Specifically, we discard masks that exceed a predefined overflow threshold or occupy disproportionately large area relative to the bounding box. This two-stage process enables robust and localized object mask generation, even under the stylistic variability and visual ambiguity common in traditional paintings.

B. Mask Generation and IoU-Based Selection

Candidate segmentation masks are first generated using MobileSAM [7]. For each image, a fixed number of binary masks are produced, representing potential object regions. To associate these masks with the given bounding box annotations, we compute the IoU between each box and all candidate masks. The mask with the highest IoU is selected as the initial object mask. This selection ensures that the mask aligns spatially with the annotated object region while preventing common failure cases, such as selecting overly broad or unrelated regions. An example of this issue is shown in Fig. 2.

This issue is particularly important in the domain of traditional paintings, where stylistic abstraction, symbolic representation, and visual degradation often lead to masks that encompass large contextual regions rather than distinct object instances. By enforcing spatial alignment through IoU-based selection, the proposed method improves the reliability of object mask extraction in visually ambiguous and stylistically diverse cultural heritage images. Additionally, this matching strategy supports efficient batch processing and supports scalable mask generation across domains.

C. Domain-Aware Geometric Filtering

While IoU-based selection improves spatial alignment between bounding boxes and candidate masks, it does not guarantee that the selected masks correspond to well-localized object instances. This limitation is especially pronounced in

Fig. 4. Mask selection results under different filtering thresholds

the domain of traditional paintings, where visual abstraction, symbolic representation, and degraded boundaries often lead to segmentation outputs that extend beyond the intended object region or encompass excessive background context.

To address these challenges, we introduce a domain-aware geometric filtering stage that refines the selected masks using two complementary criteria. First, we compute the overflow ratio by measuring the proportion of each mask that lies outside its corresponding bounding box, and masks exceeding a predefined threshold are discarded. Second, we calculate the area ratio between the mask and its bounding box to remove candidates that are disproportionately large. As illustrated in Fig. 3, this filtering scheme effectively eliminates masks that span unrelated regions or include multiple objects. The geometric constraints are critical to producing compact and semantically meaningful masks, particularly in culturally significant artworks where object boundaries may be stylized, incomplete, or visually ambiguous.

IV. Experimental Results

We evaluated our method on a dataset consisting of 1,473 traditional paintings, each annotated with bounding boxes in XML format. The dataset spans 23 culturally significant object classes, including person, peony, lotus flower, rock, and others. For each image, we use MobileSAM [7] to generate a fixed number of candidate segmentation masks. Then, the best-matching mask for each bounding box was then selected based on the highest IoU score. To investigate the effect of geometric constraints, we varied two filtering parameters: the maximum overflow ratio and the maximum area ratio.

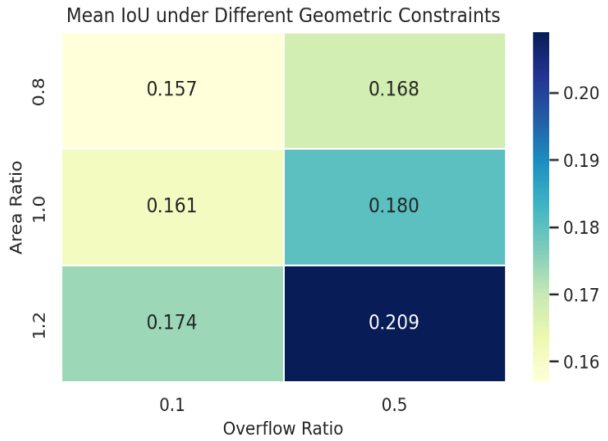

Fig. 5. Mean IoU under different overflow and area ratio settings

Fig. 6. Examples of generated object masks

Fig. 5 summarizes the mean IoU values obtained across all samples under various combinations of overflow and area ratio constraints. More relaxed settings, which correspond to higher threshold values, generally resulted in higher IoU scores. For example, a configuration with an area ratio of 1.2 and an overflow ratio of 0.5 achieved the highest mean IoU of 0.209, whereas stricter settings such as 1.0 (area) and 0.1 (overflow) yielded a lower mean IoU of 0.161. However, a higher IoU does not necessarily indicate better segmentation quality. In fact, overly permissive constraints frequently resulted in masks that extended beyond the intended object region, which is particularly problematic in traditional artworks where compositions are complex and boundaries are often ambiguous. Therefore, we complement this quantitative analysis with qualitative evaluation.

As illustrated in Fig. 4, relaxed geometric constraints often lead to over-segmentation. Selected masks encompass broader contextual regions rather than distinct object instances. Although such masks may yield higher IoU scores, they lack semantic precision. In contrast, stricter settings produce more accurate, visually coherent masks that align better with the human perception of object boundaries. Fig. 6 presents successful examples of mask generation under tighter filtering conditions, demonstrating the effectiveness of the proposed method in handling stylistic and structural ambiguity in cultural heritage images. These results suggest that, while IoU is a useful quantitative metric, domain-aware geometric filtering is necessary to produce semantically valid, practical object masks of traditional paintings.

Despite the overall effectiveness of the proposed pipeline, some failure cases remain. As illustrated in Fig. 7, the selected masks occasionally cover only partial object regions, such as garments or accessories, rather than the full object instance. This issue typically arises when the segmentation model focuses on locally distinctive visual cues instead of the

Fig. 7. Examples of failure cases where selected masks capture partial regions such as clothing

object as a whole. Addressing such cases may require integrating additional semantic priors or adopting hierarchical object modeling in future work.

V. Conclusion

This paper presents a semi-automated approach for generating object masks from bounding box annotations by integrating prompt-based segmentation and domain-aware geometric filtering. This method addresses the unique challenges of traditional paintings, where visual ambiguity, abstraction, and stylistic variation often violate the assumptions of conventional segmentation models. Through IoU-based mask selection and geometric constraints, the proposed framework enables efficient and scalable object mask generation even in visually complex and culturally significant domains. Experiments showed that although relaxed filtering conditions yielded higher IoU scores, they often resulted in over-segmented or semantically imprecise masks. In contrast, stricter filtering produced more accurate and perceptually meaningful masks, underlining the importance of domain-aware filtering in weakly supervised pipelines.

The proposed method shows promise for scalable annotation tasks in fields such as cultural heritage preservation, medical imaging, and defense. Future work will explore improved techniques for generating, selecting, and merging masks to better handle incomplete or fragmented segmentations. This will enhance the method's robustness and applicability in complex real-world scenarios.

Acknowledgment

This work was supported by Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. RS-2023-00223530, Development of the artificial intelligence technology to enhance individual soldier surveillance capabilities)

References

[1] Z. Tian, C. Shen, X. Wang, and H. Chen, "BoxInst: High-performance instance segmentation with box annotations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, 2021, pp. 5443–5452.

[2] W. Li, H. Pan, H. Zhang, and Y. Qiao, "Box-supervised instance segmentation with level set evolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 1–18.

[3] S. Lan, Z. Liu, Y. Zhang, S. Su, Y. Wang, Y. Wang, and H. Lu, "DiscoBox: Weakly supervised instance segmentation and semantic correspondence from box supervision," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, 2021, pp. 3406–3416.

[4] W. Li, H. Pan, H. Zhang, L. Zhu, and Y. Li, "Box2Mask: Box-supervised instance segmentation via level-set evolution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 7, pp. 5157–5173, Jul. 2024.

[5] T. Cheng, X. Wang, S. Chen, Q. Zhang, and W. Liu, "BoxTeacher: Exploring high-quality pseudo labels for weakly supervised instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, 2023, pp. 3145–3154.

[6] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, and R. Girshick, "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 4015–4026.

[7] C. Zhang, D. Han, Y. Qiao, and C. Xu, "Faster segment anything: Towards lightweight SAM for mobile applications," *arXiv preprint*, arXiv:2306.14289, 2023.