Active Label Correction for Semantic Segmentation with Foundation Models

Hoyoung Kim [1] **Sehyun Hwang** [2] **Suha Kwak** [1 2] **Jungseul Ok** [1 2]

Abstract

Training and validating models for semantic segmentation require datasets with pixel-wise annotations, which are notoriously labor-intensive. Although useful priors such as foundation models or crowdsourced datasets are available, they are error-prone. We hence propose an effective framework of *active label correction* (ALC) based on a design of correction query to rectify pseudo labels of pixels, which in turn is more annotator-friendly than the standard one inquiring to classify a pixel directly according to our theoretical analysis and user study. Specifically, leveraging foundation models providing useful zero-shot predictions on pseudo labels and superpixels, our method comprises two key techniques: (i) an annotator-friendly design of correction query with the pseudo labels, and (ii) an acquisition function looking ahead label expansions based on the superpixels. Experimental results on PASCAL, Cityscapes, and Kvasir-SEG datasets demonstrate the effectiveness of our ALC framework, outperforming prior methods for active semantic segmentation and label correction. Notably, utilizing our method, we obtained a revised dataset of PASCAL by rectifying errors in 2.6 million pixels in PASCAL dataset[1].

1. Introduction

Semantic segmentation has seen remarkable advancements powered by deep neural networks capable of learning from huge datasets with dense annotations for all pixels. However, such pixel-wise annotations are labor-intensive and error-prone. To address or bypass these challenges, various approaches have been studied, including crowdsourcing systems to collect large-scale human annotations (Crowston, 2012), weakly supervised learning methods to train models with image-wise annotations (Ru et al., 2023), and foundation models capable of useful zero-shot prediction on superpixels (Kirillov et al., 2023) or even semantic segmentation (Liu et al., 2023). However, those are unreliable to train and more importantly validate models for exquisite or domain-specific prediction. For instance, despite recent advances, the zero-shot prediction with foundation models (Kirillov et al., 2023; Liu et al., 2023) is considerably erroneous as demonstrated in Table 7. This can be more problematic when the semantic segmentation requiring expertise such as medical knowledge (Ma et al., 2024).

Hence, we consider the problem of active label correction (ALC) to construct a reliable pixel-wise dataset from an unreliable or unlabeled dataset with a minimum cost of user intervention. To this end, we propose an ALC framework which leverages foundation models and correction queries. Our correction query is designed to rectify the pseudo labels of pixels, only if these pseudo labels are incorrect. Unlike the standard classification query that directly requests a specific class (Cai et al., 2021; Kim et al., 2023a), our correction query allows annotators to skip labeling if the pseudo labels are correct, making it more annotator-friendly. Borrowing the information-theoretic annotation cost (Hu et al., 2020), we prove that our correction query is less costly than the classification query. Moreover, our user study in Section 4.2 reveals that the correction query is faster to complete than the classification query in practice.

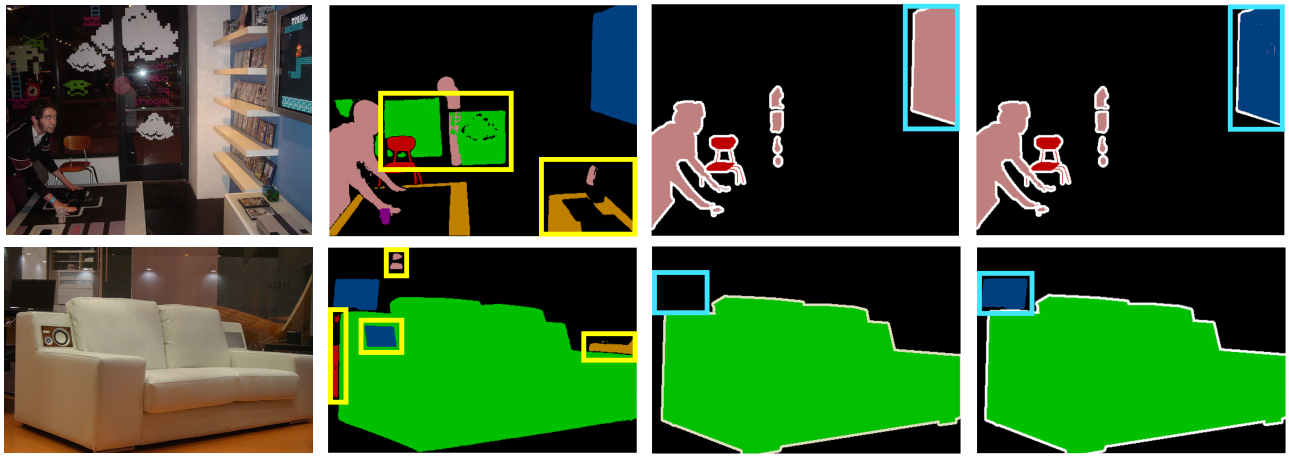
Specifically, we leverage useful zero-shot predictions on pseudo labels and superpixels from foundation models. These pseudo labels are employed in our correction query to designate pixel labels. They also allow us to warm-start, avoiding the typical cold-start problem that comes from the absence of a reliable way to evaluate data at the beginning of active learning (Mahmood et al., 2021; Chen et al., 2023). Furthermore, we fully enjoy the decent superpixels to solve the challenges of pixel-wise queries. Although pixel-wise queries can generate a flawless dataset, they require substantial time and memory to examine each pixel and lead to redundancy in the pixels chosen (Shin et al., 2021).

To address the problems, we devise superpixel-aware strategies across our entire framework. Initially, we build a diversified pixel pool consisting of partial key pixels representing

[1]Graduate School of AI, POSTECH, Pohang, Republic of Korea [2]Department of CSE, POSTECH, Pohang, Republic of Korea. Correspondence to: Jungseul Ok <jungseul@postech.ac.kr>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

[1]https://github.com/ml-postech/active-label-correction



(a) Unlabeled image (b) Grounded-SAM (c) PASCAL (d) PASCAL+ (ours)

Figure 1: *Examples of noisy and corrected labels in PASCAL.* (a, b) Initial pseudo labels are generated by applying Grounded-SAM (G-SAM) to unlabeled images. As depicted by the yellow boxes, noisy pseudo labels result in a decline in performance, as shown in Table 7. (c) PASCAL also contains noisy labels in cyan boxes. (d) By employing the superpixels from G-SAM, we construct a corrected version of PASCAL, called PASCAL+. For instance, in the first row, we correct the object labeled as person to tvmonitor, and in the second row, the object labeled as background to tvmonitor. Here, the colors black, blue, red, green, and pink represent the background, tvmonitor, chair, sofa, and person classes, respectively.

each image. As superpixels cluster pixels with similar features (Van den Bergh et al., 2012), we choose one representative pixel per superpixel and add it to our pixel pool. To solve the inefficiency of correcting each pixel individually per query, we extend the corrections from individual pixels to the entire superpixels they belong to. Accordingly, we propose a look-ahead acquisition function, which anticipates the benefits of label expansion beforehand.

The proposed framework is notably cost-efficient in constructing clean segmentation datasets. We evaluate it by constructing new segmentation datasets from the initial pseudo labels given by foundation models in different fields, including the medical domain. Our ALC framework outperforms prior methods for active semantic segmentation and label correction over a range of budgets. In particular, we highlight its practical application by enhancing the popular PASCAL dataset (Everingham et al., 2012). We call our corrected dataset PASCSAL+, which can be widely used in the literature of semantic segmentation.

Our main contributions are summarized as follows:

- We provide theoretical and empirical justifications on the efficacy of the correction query, compared to the classification query (Section 3.2 and 4.2).
- We propose an active label correction framework, leveraging the correction query and foundation models, where the look-ahead acquisition function enables selecting informative and diverse pixels to be corrected (Section 3.3 and 3.4).
- To achieve comparable performance with SOTA active semantic segmentation methods, we only use 33% to 50% of budgets on various datasets (Section 4.2).
- Using the proposed framework, we correct 2.6 million pixel labels in PASCAL and provide a revised version, called PASCAL+ (Section 5.2).

2. Related Work

Active Learning for Segmentation. Active Learning (AL) (Kim et al., 2023b; Saran et al., 2023; Yang et al., 2023) aims at increasing labeling efficiency by selectively annotating informative subsets of data. In semantic segmentation, previous work focuses on two aspects: the design of labeling units and acquisition functions. In terms of labeling unit design, classical approaches explore image-based (Yang et al., 2017; Sinha et al., 2019) and patch-based (Mackowiak et al., 2018; Casanova et al., 2019) selection. Recently, superpixel-based approaches (Siddiqui et al., 2020; Cai et al., 2021; Hwang et al., 2023; Kim et al., 2023a), are gaining attention as they only require one click for labeling each region. In terms of acquisition functions, they generally focus on selecting uncertain regions, measured with entropy (Mackowiak et al., 2018; Kasarla et al., 2019), the gap between the top-1 and the top-2 predictions (Joshi et al., 2009; Wang et al., 2016; Cai et al., 2021; Hwang et al., 2023; Kim et al., 2023a). While conventional AL methods collect labels from scratch, the proposed method starts from the initial pseudo labels from foundation models, correcting erroneous labels.

Noisy Label Detection. The studies in noisy label detection (NLD) aim to identify incorrect labels efficiently by selecting error-like samples. In computer vision, methods for robust training toward label noise often include NLD components (Natarajan et al., 2013; Xiao et al., 2015; Patrini et al., 2017; Han et al., 2018; Ren et al., 2018; Song et al., 2022), and recently, there is an increase in studies focusing solely on NLD (Müller & Markert, 2019; Northcutt et al., 2021b). NLD methods for semantic segmentation aggregate pixel-wise error scores into labeling units, like an image or superpixel. Lad & Mueller (2023) aggregate per-pixel error scores from Confident Learning (Northcutt et al., 2021a) into per-image scores, while Rottmann & Reese (2023) average error scores from locally connected components sharing the same pseudo label. Recently, the Active Label Correction (ALC) (Bernhardt et al., 2022; Kim, 2022) methods identify noisy labels and correct them in a classification task. Our work is the first ALC method for semantic segmentation, correcting pixel labels and expanding them to their corresponding superpixels.

Efficient Query Design. Designing a practical and cost-effective annotation query is crucial, as it directly impacts annotation budgets. In semantic segmentation, various approaches have been explored, including classification queries asking for a specific class (Cai et al., 2021; Kim et al., 2023a), one-bit queries requesting yes or no responses (Hu et al., 2020), and multi-class queries obtaining all classes in a superpixel (Hwang et al., 2023). Recently, there have been studies on efficiently constructing datasets using foundation models. For instance, Wang et al. (2023) leverages these models for automated labeling in remote sensing imagery, and Qu et al. (2023) focuses on building large medical datasets with them. However, its query form is stagnant in previous query types. By employing the initial pseudo labels from foundation models, we suggest correction queries that only request the correct label when the given pseudo label is incorrect.

3. Active Label Correction Framework

Given an initial noisy dataset $\mathcal{D}_0$, we consider an active label correction (ALC) scenario operating with pixel-wise labeling. Each query to an oracle annotator requests the accurate label $y \in \mathcal{C} := \{1, 2, ..., C\}$ for an associated pixel x. In contrast to active learning (AL), which commences with an unlabeled image set, ALC focuses on progressively refining a labeled dataset $\mathcal{D}_0$ which may include noisy labels. For each round t, we issue a batch $\mathcal{B}_t$ of B queries from a pixel pool $\mathcal{X}_t$ and train a model θ_t with the corrected annotations obtained so far.

In the following, we first prepare an initial dataset for correction (Section 3.1). After that, we present a correction query that requests for rectifying pseudo labels of pixels (Section 3.2). To fully enjoy the corrections, we introduce a look-ahead acquisition function, which selects from a diversified pixel pool (Section 3.3), considering the effect of label expansion (Section 3.4). The overall procedure is summarized in Algorithm 1.

Algorithm 1 Proposed Framework

Require: Batch size B, and final round T.
1: Prepare initial dataset $\mathcal{D}_0$ requiring label correction
2: Obtain model θ_0 training with $\mathcal{D}_0$ via (1)
3: **for** $t = 1, 2, \ldots, T$ **do**
4: Construct diversified pixel pool $\mathcal{X}_t^d$ via (4)
5: Correct labels of selected B pixels $\mathcal{B}_t \subset \mathcal{X}_t^d$ via (9)
6: Expand corrected labels to corresponding superpixels
7: Obtain model θ_t training with corrected $\mathcal{D}_t$ via (11)
8: **end for**
9: **return** $\mathcal{D}_T$ and θ_T

3.1. Initial Dataset Preparation

For ALC, an initial segmentation dataset is essential, and we can start with well-known datasets like Cityscapes (Cordts et al., 2016) or PASCAL VOC (PASCAL) (Everingham et al., 2012). However, the presence of labeled datasets may be impractical in many domains. Employing AL is one method for preparing labeled datasets. However, AL typically builds datasets through random pixel (Shin et al., 2021) or superpixel labeling (Cai et al., 2021) leading to lots of budgets and rounds, as it starts from unlabeled images, commonly known as the cold-start problem (Mahmood et al., 2021). Away from conventional AL methods, we utilize recent foundation models to construct segmentation datasets.

Recently, foundation models for zero-shot segmentation have been emerged. For example, Grounded-SAM, a fusion of Grounding DINO (Liu et al., 2023) and Segment Anything Model (Kirillov et al., 2023) is capable of detecting and segmenting objects based on text prompts. Each class is identified with its own text prompt, and we can obtain the initial pseudo labels by using a series of $|\mathcal{C}|$ text prompts, one for each class. We solve the problem of multi-classes in object detection by giving each object the most likely class as a pseudo-label. Figures 1a and 1b display examples of the unlabeled images in PASCAL and corresponding initial pseudo labels generated by Grounded-SAM.

Warm-start. In contrast to the cold-start problem in AL, our ALC benefits from warm-start thanks to the initial labels provided by foundation models. In Appendix A, detailed descriptions of text prompts for warm-start are provided. To obtain θ_0, we initialize θ to a model pre-trained on ImageNet (Deng et al., 2009). We then train it to reduce the following cross-entropy (CE) loss:

$$\hat{\mathbb{E}}_{(x,y)\sim\mathcal{D}_0}[\mathrm{CE}(y, f_\theta(x))] \,, \tag{1}$$

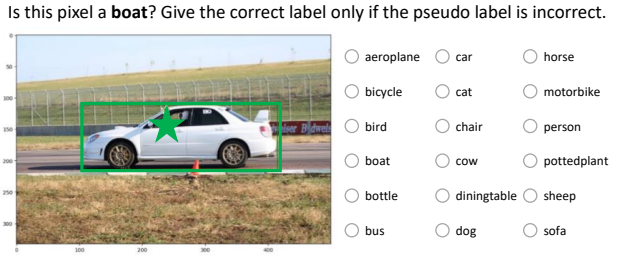


Figure 2: *An example of correction query*. Correction query presents an instruction requesting a label for a representative pixel (green star), an image displaying an object within a bounding box (green rectangle), and possible class options.

where $f_\theta(x) \in \mathbb{R}^{|\mathcal{C}|}$ represents the estimated class probability for pixel x by the model θ. Here, the difference lies in $\mathcal{D}_0$: AL uses only partial y, while ALC can access all y for each pixel x. However, compared to ground-truth in Figure 1c, the initial pseudo-labels in Figure 1b contain noisy labels. This results in negative impacts on the model's performance, as shown in Table 7. Therefore, active label correction is essential for rectifying these noisy labels.

3.2. Correction Query

Once we prepare the initial dataset for correction, we use our correction query to rectify the pseudo labels of pixels. As the number of classes increases, the classification query asking for the precise label of a pixel can become more time-consuming (Zhang et al., 2022). In contrast, our correction query lowers the overall cost by reducing the number of classification queries needed, allowing annotators to bypass labeling when the pseudo label is already correct. Specifically, we use the instruction with a pseudo label on a pixel, written as follows:

Give the correct label only if the pseudo label is incorrect.

Figure 2 and Appendix B provide detailed descriptions of our correction query. In the following, we information-theoretically compare the expected costs of classification and correction queries, denoted by C_{cls} and C_{cor}, respectively.

Theorem 3.1. *Assume the information-theoretic annotation cost (Hu et al., 2020) of selecting one out of L possible options to be $\log_2 L$. Let $L \geq 2$ be the number of classes, and p be the probability that the pseudo label is correct. Then, $C_{\text{cls}}(L) = \log_2 L$ and $C_{\text{cor}}(L, p) = p + (1-p)\log_2 L$. Thus, for any $p \in [0, 1]$ and $L \geq 2$,*

$$1 - \frac{C_{\text{cor}}(L,p)}{C_{\text{cls}}(L)} = \left(1 - \frac{1}{\log_2 L}\right) p \geq 0 \,. \tag{2}$$

Proof. The correction query can be interpreted as a binary question if the pseudo label is correct, and a L-ary one otherwise. Recalling the definition of p and $C_{\text{cls}}(L) = \log_2 L$, we have $C_{\text{cor}}(L, p) = p \log_2 2 + (1-p)\log_2 L$. □

The costs of both correction and classification queries are the same if $L = 2$. Indeed, those are logically identical when $L = 2$. In (2), the cost-saving rate using the correction query instead on the classification one is computed as $\left(1 - \frac{1}{\log_2 L}\right) p$, which is increasing in p and L. Hence, using the correction query is particularly beneficial when the number of classes is large or the pseudo labels can be obtained accurately. In addition, a user study on correction queries experimentally confirms their practical effectiveness in Section 4.2.

3.3. Diversified Pixel Pool

Employing pixel-wise queries is instrumental in constructing error-free segmentation datasets. However, examining each pixel with an acquisition function requires substantial time and memory. Furthermore, as adjacent pixels often share similar acquisition values, there exists a risk of lacking diversity in the selected pixels, i.e., pixels in a certain area of the image with high acquisition values may be picked simultaneously. To tackle these challenges at once, we propose a diversified pixel pool $\mathcal{X}^d$, which is a subset of the total pixel set $\mathcal{X}$, as follows:

$$\mathcal{X}^d := \{x_1, x_2, \ldots, x_{|\mathcal{S}|}\} \,, \tag{3}$$

where each x_i represents a key pixel from the superpixel s_i within the set of superpixels $\mathcal{S}$.

Specifically, starting with a model θ_{t-1} trained on the dataset $\mathcal{D}_{t-1}$ from the previous round, we construct a diversified pixel pool $\mathcal{X}^d_t := \{x_{t1}, x_{t2}, \ldots, x_{t|\mathcal{S}|}\}$ for the current round t. For ease of explanation, we refer to θ_{t-1} simply as θ, x_{ti} as x_i and $\mathcal{X}^d_t$ as $\mathcal{X}^d$. We select a representative pixel x_i from each superpixel s_i based on the highest cosine similarity as:

$$x_i := \underset{x \in s_i}{\arg\max} \frac{f_\theta(x) \cdot f_\theta(s'_i)}{\|f_\theta(x)\| \|f_\theta(s'_i)\|} \,, \tag{4}$$

where $f_\theta(s) := \frac{\sum_{x \in s} f_\theta(x)}{|\{x : x \in s|\}}$ represents the averaged class prediction for superpixel s. To address the flaws in superpixels and ensure more uniformity of pixel labels within them, we employ a subset s' rather than the complete set s. We start by defining the pseudo dominant label $\mathrm{D}_\theta(s)$, which serves as the representative label for superpixel s according to model θ, as follows:

$$\mathrm{D}_\theta(s) := \underset{c \in \mathcal{C}}{\arg\max} \, |\{x \in s : y_\theta(x) = c\}| \,, \tag{5}$$

where $y_\theta(x) := \arg\max_{c \in \mathcal{C}} f_\theta(c; x)$ is the estimated label for pixel x using model θ. Subsequently, we form the subset

s', consisting of pixels that align with the pseudo dominant label $\mathrm{D}_\theta(s)$, as follows:

$$s' := \{x \in s : y_\theta(x) = \mathrm{D}_\theta(s)\} \ . \tag{6}$$

After that, we select the pixel that best represents s' for each superpixel based on (4), contributing to the formation of a diverse pixel pool in (3). We highlight that the proposed diversified pixel pool reduces time and memory usage and lessens the redundancy issue in the chosen pixels.

Remarks. While various superpixel generation algorithms (Achanta et al., 2012; Van den Bergh et al., 2012) can be used for $\mathcal{S}$ in (3), these standard algorithms typically group neighboring pixels based on similar inherent properties like color and maintain nearly uniform sizes. Recent research indicates that semantically considered superpixels from a model are effective for AL in segmentation (Kim et al., 2023a). Therefore, we opt to organize superpixels based on the objects identified by Grounded-SAM.

3.4. Look-Ahead Acquisition Function

Once the set of pixels $\mathcal{X}_t^d$ for examination through an acquisition function is established, we select a pixel batch $\mathcal{B}_t \subset \mathcal{X}_t^d$ of size B to be corrected. In each round t, we iteratively select the most informative pixel, guided by the acquisition $a(x; \theta_{t-1})$:

$$x^* := \underset{x \in \mathcal{X}_t^d}{\arg\max}\, a(x; \theta_{t-1}) \ . \tag{7}$$

For simplicity, we refer to θ_{t-1} as θ. Recently, Lad & Mueller (2023) propose a confidence in label (CIL), which evaluates the confidence of a given label y for a pixel x, using the predictions of the model θ as follows:

$$a_{\mathrm{CIL}}(x; \theta) := 1 - f_\theta(y; x) \ . \tag{8}$$

The underlying assumption is that a pixel is likely mislabeled if the model demonstrates insufficient learning about that pixel's label. However, correcting only a single pixel with each query is not only inefficient but also has minimal impact on the learning process. To enhance the efficiency of pixel-wise query, we introduce a label expansion technique, which involves extending the corrected label of a pixel x into pixels in the same superpixel s.

Accordingly, we suggest a look-ahead acquisition function that not only assesses the unreliability of a pixel x as described in (8), but also takes into account the effect of label expansion into the superpixel s. Here, we rename x to x_r as it serves as a representative pixel for s. For a representative pixel x_r of s, our acquisition function is defined as follows:

$$a_{\mathrm{SIM}}(x_r; s, \theta) := \sum_{x \in s} \frac{f_\theta(x_r) \cdot f_\theta(x)}{\|f_\theta(x_r)\| \|f_\theta(x)\|} a_{\mathrm{CIL}}(x; \theta) \ , \tag{9}$$

where the cosine similarity between two feature vectors is related to the likelihood of correctly expanding the correct label of pixel x_r to another pixel x.

We note that previous acquisitions including CIL in (8) can be transformed easily to its look-ahead counterparts. For instance, the look-ahead CIL (LCIL) acquisition can be defined by adjusting the weight of each pixel from the cosine similarity to the inverse of the superpixel size as:

$$a_{\mathrm{LCIL}}(x_r; s, \theta) := \sum_{x \in s} \frac{1}{|s|} a_{\mathrm{CIL}}(x; \theta) \ . \tag{10}$$

Finally, in round t, we select the B most informative pixels from the diversified pixel pool $\mathcal{X}_t^d$ in order of SIM acquisition to form query batch $\mathcal{B}_t$.

After obtaining the clean labels of selected B pixels, we expand them to the associated superpixels. We finally construct the dataset $\mathcal{D}_t$ for round t by combining the previous dataset $\mathcal{D}_{t-1}$ with the updated annotations. Analogously to the warm-start, we initialize θ_t to a model pre-trained on ImageNet, minimizing the following CE loss:

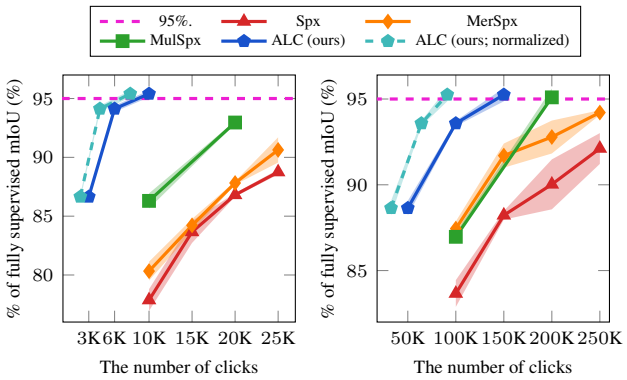
$$\hat{\mathbb{E}}_{(x,y)\sim\mathcal{D}_t}[\mathrm{CE}(y, f_\theta(x))] \ . \tag{11}$$

4. Experiments

4.1. Experimental Setup

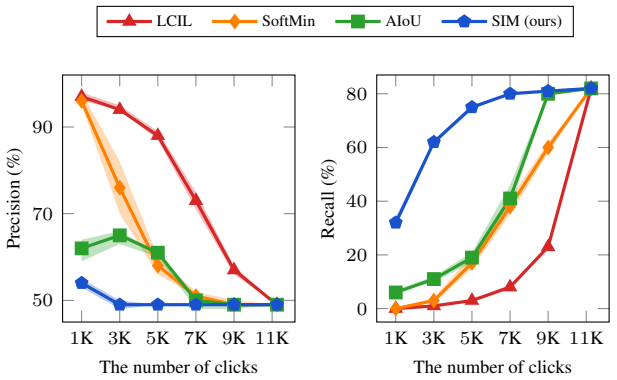
Datasets. We use three semantic segmentation datasets: Cityscapes (Cordts et al., 2016), PASCAL VOC 2012 (PASCAL) (Everingham et al., 2012), and Kvasir-SEG (Jha et al., 2020). Cityscapes comprises 2,975 training and 500 validation images with 19 classes, while PASCAL consists of 1,464 training and 1,449 validation images with 20 classes. Kvasir-SEG is a medical dataset for polyp segmentation consists of 880 training and 120 validation images with 2 classes.

Implementation Details. We adopt DeepLab-v3+ architecture (Chen et al., 2018) with Resnet101 pre-trained on ImageNet (Deng et al., 2009) as our segmentation model. During training, we use the SGD optimizer with a momentum of 0.9 and set a base learning rate of 0.1. We decay the learning rate by polynomial decay with a power of 0.9. For Cityscapes, we resize training images to 768×768 and train a model for 30K iterations with a mini-batch size 16. For PASCAL, we resize training images to 513×513 and train a model for 30K iterations with a mini-batch size 16. For Kvasir-SEG, we resize training images to 352×352 and train a model for 6.3K iterations with a mini-batch size 32. For the initial dataset generated with Grounded-SAM, we use the box threshold of 0.2 for Cityscapes and PASCAL, and 0.05 for Kvasir-SEG.



(a) PASCAL (b) Cityscapes

Figure 3: *Effect of active label correction.* ALC shows comparable results on both datasets with much fewer clicks. ALC (normalized) reflects the reduced budget of correction queries with normalization by Theorem 3.1.



(a) Precision (b) Recall

Figure 4: *Precision and recall comparisons.* Our SIM acquisition shows a high recall, indicating it corrects many noisy pixels with limited budgets.

Table 1: *User study for different queries.* Our correction query C_{cor} proves to be more cost-effective compared to classification query C_{cls}.

Query	Total time (s)	Time per query (s)	Accuracy (%)
C_{cls}	$126.1_{\pm 19.8}$	$6.31_{\pm 0.99}$	$95.0_{\pm 3.3}$
C_{cor}	$\mathbf{95.1}_{\pm 9.0}$	$\mathbf{4.76}_{\pm 0.45}$	$95.0_{\pm 4.0}$

Table 2: *Quality of corrected datasets.* The labels of 5K pixels from the initial datasets are corrected using different acquisition functions in the ALC framework.

Acquisition function	Data mIoU (%)	Model mIoU (%)
LCIL	$56.59_{\pm 0.07}$	$56.82_{\pm 0.05}$
SoftMin	$59.28_{\pm 0.59}$	$58.66_{\pm 0.89}$
AIoU	$59.95_{\pm 0.57}$	$59.04_{\pm 0.27}$
SIM (ours)	$\mathbf{83.04}_{\pm 0.62}$	$\mathbf{68.72}_{\pm 0.10}$

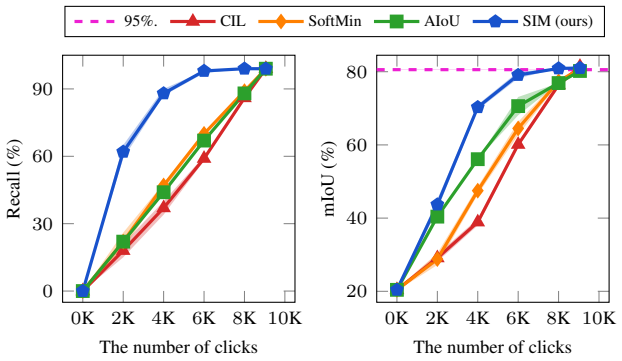
4.2. Main Experiments

Baselines. Our Active Label Correction (ALC) method is compared with the state-of-the-art (SOTA) superpixel-based active learning (AL) methods: Spx (Cai et al., 2021), MerSpx (Kim et al., 2023a), and MulSpx (Hwang et al., 2023). They are chosen for two reasons: (1) Their measure of labeling cost is the same as ours, i.e., the number of label clicks. (2) They are SOTA methods in AL for segmentation. Following conventional AL methods (Cai et al., 2021), we highlight the amount of annotation used to achieve 95% performance of the fully supervised baseline, where *95%.* denotes performance.

Evaluation Protocol. Given a limited budget, we identify and fix noisy pixel labels, and expand them to the related superpixels to construct the corrected dataset. Then, we develop a model using the dataset and evaluate its effectiveness with mean Intersection over Union (mIoU). In all experiments, we report the average results from three trials, with graph shading indicating the standard deviation. We access the model not only on the test dataset but also on the training dataset to calculate the quality of the dataset itself.

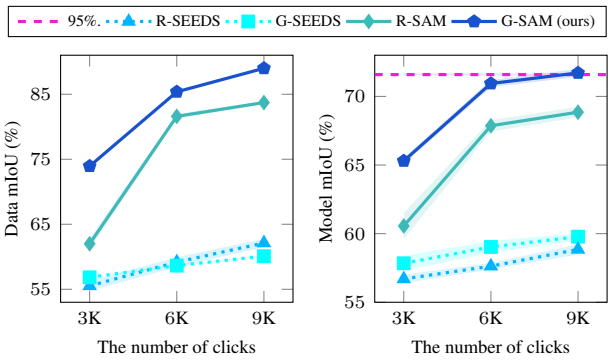
Active Label Correction vs. Active Learning. In Figure 3, we show the effectiveness of our framework, named ALC, compared with current AL methods over various budget levels, represented by the number of clicks, for both PASCAL and Cityscapes datasets. Due to variations in models and hyperparameters used in previous methods, we ensure a fair comparison by evaluating the percentage of fully supervised mIoU, where additional comparisons with absolute mIoU is reported in Appendix C. The results illustrate that our ALC substantially reduces the necessary budgets to achieve 95% target performance. Specifically, ALC achieves 95% of the fully supervised baseline performance with just 6K clicks for PASCAL and 150K clicks for Cityscapes. This is only 30% and 75% of the budget required by the previous SOTA methods, respectively. Even when considering the efficient labeling cost of correction queries in Theorem 3.1, the cost of our proposed method reduces to 68% of its original version, where p in (2) is 0.27 and 0.5 in PASCAL and Cityscapes, respectively. This result is denoted as ALC (normalized) in Figure 3.

Verification of Labeling Costs with User Study. In Theorem 3.1, we prove that the labeling cost of the correction query C_{cor} is lower than the classification query C_{cls}. In Table 1, we empirically show its effectiveness with a user study conducted by 20 annotators, where they are given 20 queries with $p = 0.5$ scenarios. Theoretically, as $L = 20$ in PASCAL, the cost ratio between the two queries is about 0.62. In Table 1, we observe that C_{cor} requires 0.75 times



(a) Recall (b) Kvasir-SEG

Figure 5: *Kvasir-SEG experiments.* The proposed SIM acquisition operate robustly on medical dataset across different budgets.



(a) Data mIoU (b) Model mIoU

Figure 6: *Advantages of foundation models.* Our ALC is called G-SAM, as it depends on Grounded-SAM. The effect of superpixels is larger than that of initial pseudo-labels.

the cost of C_{cls}, in practice. More details about user study are in the Appendix B.

4.3. Effectiveness of Proposed Acquisition Function

Baselines. In our ALC framework, we compare our SIM acquisition with previous ones for detecting noisy labels in segmentation datasets, such as LCIL, SoftMin (Lad & Mueller, 2023), and AIoU (Rottmann & Reese, 2023). For a fair comparison, we keep all other methodologies constant, including a diversified pixel pool, lookahead strategy, and label expansion, varying only the acquisition function.

Evaluation Protocol. Given a limited budget, we select unreliable pixels, correct their labels, and expand them to the corresponding superpixels. We first evaluate the efficiency of the acquisition functions in terms of precision and recall at the pixel level. Specifically, precision refers to the proportion of pixels correctly identified as mislabeled out of the selected pixels, while recall represents the fraction of pixels chosen correctly from the total number of mislabeled pixels. Then, we access models trained with each corrected dataset from different acquisitions with mIoU. The ablation experiments on acquisition are conducted in PASCAL.

Precision and Recall of Acquisition Functions. In Figure 4, we compare SIM to baseline acquisitions by calculating the precision and recall for detecting incorrect pixels. SIM outperforms the baseline acquisition functions in terms of recall while showing a comparably low precision rate. This is attributed to SIM considering the effect of label expansion as in (9), which favors large superpixels. We consider this design choice to be effective for ALC for two reasons. First, as demonstrated in Theorem 3.1, the labeling cost of reconfirming false positives, i.e., correct pseudo labels, is significantly low. Second, correcting the labels of as many pixels as possible, which is related to high recall, leads to greater improvements in data and model performance as shown in Table 2.

Quality of Corrected Datasets. In Table 2, we compare SIM to baseline acquisition functions in terms of the quality of the corrected dataset when using 5K clicks. The quality of the corrected dataset is evaluated by the accuracy of corrected labels (Data mIoU), and the performance of a model trained with them (Model mIoU). For both metrics, the dataset corrected by SIM shows the best quality. This shows that the performance of the model is more correlated to the recall in Figure 4b, as high recall indicates fewer incorrect labels in the dataset.

4.4. Further Analyses

Applicability to Medical Dataset. In Figure 5, we apply the ALC framework to the Kvasir-SEG dataset to verify the generalization ability of our framework to challenging medical domain. Here, the initial dataset shows 20% mIoU, as shown in 0K of Figure 5b. Even under such challenging initial conditions, the ALC combined with SIM reaches 93% performance of the fully supervised model only using 6K clicks. This performance can be attributed to SIM acquisition function, which consistently achieves the highest recall among baselines over various numbers of clicks, as shown in Figure 5a. We note that our approach introduces superpixel-wise sampling to Kvasir-SEG for the first time, diverging from the traditional image-wise sampling methods (Smailagic et al., 2018; Wu et al., 2021).

Decomposing the Advantages of Foundation Models. In Figure 6, we analyze the effect of the foundation model for ALC in two aspects: initial pseudo-labels and superpixels, in PASCAL. We denote the proposed method using both aspects as G-SAM. For the baseline, we initially train a model with a 3K budget through random sampling and then employ this model as the pseudo-label generator in subsequent rounds, which is denoted as R-SAM. We note that the distinction between G-SAM and R-SAM lies in the method of obtaining initial pseudo-labels, rather than in

Table 3: *Synergy of proposed components.* We conduct an ablation study, when correcting the initial dataset using 5K budgets in PASCAL.

Acquisition Diversity	Acquisition Look-ahead	Expansion	Data mIoU	Model mIoU
✗	✗	✗	$55.03_{\pm 0.25}$	$56.30_{\pm 0.56}$
✗	✓	✓	$55.38_{\pm 0.08}$	$56.01_{\pm 0.58}$
✓	✗	✓	$56.59_{\pm 0.07}$	$56.82_{\pm 0.05}$
✓	✓	✗	$55.61_{\pm 0.00}$	$56.69_{\pm 0.35}$
✓	✓	✓	$\mathbf{83.04}_{\pm 0.62}$	$\mathbf{68.72}_{\pm 0.10}$

Table 4: *Fair comparison between Spx and ALC.* For a fair comparison, we integrate two advantages of foundation models into Spx. We refine the initial dataset using 3K budgets in PASCAL.

Methods	Initial stage	Superpixels	Model mIoU (%)
Spx	Cold-start	SEEDS	$52.34_{\pm 0.85}$
Spx	Warm-start	SEEDS	$57.77_{\pm 0.70}$
Spx	Warm-start	SAM	$57.79_{\pm 0.66}$
ALC	Warm-start	SAM	$\mathbf{65.30}_{\pm 0.21}$

the acquisition itself. In subsequent rounds, namely for budgets of 6K and 9K, both R-SAM and G-SAM adhere to the same experimental settings, including the same SIM acquisition function. Another baseline is to use superpixels from SEEDS (Van den Bergh et al., 2012) instead of the ones from SAM, which is denoted as G-SEEDS. We denote R-SEEDS as a baseline combining both random sampling in the initial round and superpixels from SEEDS. As shown in Figure 6, both aspects improve both Data mIoU and Model mIoU. In particular, utilizing the superpixels from SAM shows significant performance improvement.

Synergy of Proposed Components. Table 3 quantifies the contribution of each component in our method: (1) the diversified pixel pool (Diversity) in Section 3.3, (2) the look-ahead acquisition (Look-ahead), and (3) the label expansion technique (Expansion) in Section. 3.4. The ablation study is conducted by correcting the initial dataset using 5K budgets in PASCAL, and evaluated with both the accuracy of corrected labels (Data mIoU) and the performance of a model trained with them (Model mIoU). The results show that all components improve both Data mIoU and Model mIoU. In particular, the synergy of proposed components is pronounced. Since correcting numerous pixels across various regions simultaneously is significant, omitting even one component results in significant performance degradation.

Fair Comparison with Baselines. We provide additional experiments and discussions to clarify the advantages of our method called ALC, compared to adopting Grounded-SAM (G-SAM) to Spx baseline. In turn, only our method fully leverages G-SAM mainly thanks to our acquisition function, SIM. Table 4 presents an ablation study on the advantages of G-SAM, which are two-fold: warm-start with initial pseudo-labels and SAM superpixels. The gap between the first and second rows quantifies the advantage of warm-start with G-SAM when using Spx. This is not substantial since the pseudo labels from G-SAM contain considerable noises, as shown in Figure 1b, i.e., Data mIoU 55.32% in PASCAL. In addition, comparing the second and third rows, the advantage of using SAM superpixels for Spx is negligible. The gain of our method in the fourth row is clear. This is mainly thanks to the proposed acquisition function, SIM, with the look-ahead ability. We note that MerSpx (Kim et al., 2023a) based on ClassBal of Spx has no such look-ahead. MulSpx (Hwang et al., 2023) proposes a multi-class query, which requests labeling all classes within a superpixel, making it difficult to conduct a fair comparison.

5. PASCAL+ corrected from PASCAL

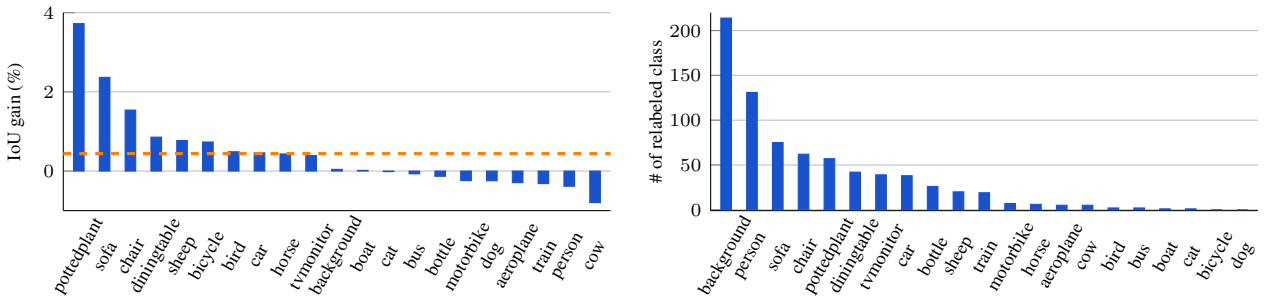
To demonstrate the practicality of the proposed framework, we apply corrections to the widely-used PASCAL dataset (Everingham et al., 2012), resulting in an enhanced version named PASCAL+ dataset (Section 5.1). Figures 1c and 1d illustrate the change in labels between PASCAL and PASCAL+ datasets, respectively. We demonstrate the enhanced model performance when using PASCAL+ compared to PASCAL and verify the cost-effectiveness of our SIM acquisition function (Section 5.2).

5.1. Construction Process

We apply our active label correction to construct the refined version of the PASCAL dataset. We first generate 81K superpixels using Grounded-SAM, where we use 0.1 as the box threshold. Considering that PASCAL has 1,464 images for training and 1,449 for validation, the average number of superpixels per image is around 29. Then we correct the pseudo label of each superpixel by annotating the true label to the corresponding representative pixel and expanding the label to the superpixel. The relabeling tasks are conducted by two annotators, each spending around 60 hours over two weeks. When labels from two annotators are different, the final annotation is determined by discussion. The qualitative result of PASCAL+ compared to PASCAL is illustrated in Figures 1 and 9. Additionally, in Figure 10, we report few failure cases for label correction due to the imperfection of superpixels, budget constraints, and human error.

5.2. Analysis of PASCAL+

Effect of PASCAL+. In PASCAL+, we make 743 superpixel label corrections in total, with 375 in the training set and 368 in the validation set. Approximately 0.5% of the



(a) Class-wise IoU gain

(b) Distribution of corrected class

Figure 7: *PASCAL+ statistics.* (a) The IoU gain is calculated by averaging the improvements in the train and valid datasets in Table 5. The orange line (---) denotes the average gain. (b) Certain classes are corrected a lot. Notably, the pottedplant, sofa, chair, and diningtable classes get many corrections, leading to a noticeable increase in IoU gain.

Table 5: *Effect of PASCAL+.* P denotes PASCAL, while P+ denotes PASCAL+. The refined train set increases model performance on both the original and refined validation sets.

Train	Valid	Data mIoU (%)	Model mIoU (%)
P	P	99.1	$75.36_{\pm 0.07}$
P+	P	100.0	$75.78_{\pm 0.12}$
P	P+	99.1	$76.18_{\pm 0.08}$
P+	P+	100.0	$\mathbf{76.42}_{\pm 0.03}$

Table 6: *Performance of corrected dataset.* With 10K budgets, we correct PASCAL to PASCAL+ with different acquisitions.

Acquisition function	Data mIoU (%)	Model mIoU (%)
LCIL	$99.16_{\pm 0.00}$	$75.68_{\pm 0.25}$
SoftMin	$99.38_{\pm 0.01}$	$75.76_{\pm 0.23}$
AIoU	$99.28_{\pm 0.02}$	$75.61_{\pm 0.22}$
SIM (ours)	$\mathbf{99.78}_{\pm 0.11}$	$\mathbf{75.87}_{\pm 0.22}$

pixel labels, equivalent to 2.6 million pixels, are altered, resulting in a 0.9% improvement in the mean Intersection over Union (mIoU) for the training set, as shown in Table 5. Regardless of whether the valid set is PASCAL or PASCAL+, corrections to the training data enhance the mIoU by around 0.3%. In particular, Figure 7a represents that IoU scores for the pottedplant and sofa classes are increased by more than 2%. This trend is related to the distribution of the corrected classes in Figure 7b. Excepting the background and person classes, which already achieve high IoU scores with PASCAL in Figure 12, the IoU scores tend to improve in line with the number of corrections applied to classes that initially have more errors. PASCAL+ not only enhances the reliability of segmentation model evaluations but also has the potential to reduce both false negatives and false positives in the literature of detecting noisy labels for segmentation tasks, thereby contributing to more reliable and precise outcomes in this field.

Various Acquisitions for PASCAL+. Since it is possible to access both the noisy PASCAL and clean PASCAL+ datasets at the same time, we analyze which acquisition function is effective in real-world. Table 6 indicates that our SIM acquisition achieves nearly 100% Data mIoU, i.e., almost similar to PASCAL+, with selecting 10K pixels for correction. As the training dataset's quality improves, there is a corresponding slight increase in model performance.

6. Conclusion

In this work, we propose a framework for active label correction in semantic segmentation operating with foundation models. Our framework includes cost-efficient correction queries, which are verified theoretically and empirically, that ask for a pixel label to be corrected if needed. We fully enjoy the benefits of foundation models, namely initial pseudo-labels and decent superpixels, resulting in significant budget reduction across various datasets in different domains. In addition, we demonstrate the practicality of our framework by constructing PASCAL+, a corrected version of the PASCAL dataset.

Limitations. Our framework depends on foundation models, particularly Grounded-SAM (Liu et al., 2023), and shares the same inherent limitations as these models, like generating incomplete superpixels for minor domains. However, we demonstrate the effectiveness of our framework in the medical field, and we expect these issues to be resolved as foundation models continue to improve over time.

Acknowledgements. This work was partly supported by the IITP grants and the NRF grants funded by Ministry of Science and ICT, Korea (No.RS-2019-II191906, Artificial Intelligence Graduate School Program (POSTECH); No.RS-2021-II212068, Artificial Intelligence Innovation Hub; No.RS-2023-00217286; No.RS-2022-II220926).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., and Süsstrunk, S. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012.

Bernhardt, M., Castro, D. C., Tanno, R., Schwaighofer, A., Tezcan, K. C., Monteiro, M., Bannur, S., Lungren, M. P., Nori, A., Glocker, B., et al. Active label cleaning for improved dataset quality under resource constraints. *Nature communications*, 13(1):1161, 2022.

Cai, L., Xu, X., Liew, J. H., and Foo, C. S. Revisiting superpixels for active learning in semantic segmentation with realistic annotation costs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10988–10997, 2021.

Casanova, A., Pinheiro, P. O., Rostamzadeh, N., and Pal, C. J. Reinforced active learning for image segmentation. In *International Conference on Learning Representations*, 2019.

Chen, L., Bai, Y., Huang, S., Lu, Y., Wen, B., Yuille, A., and Zhou, Z. Making your first choice: to address cold start problem in medical active learning. *Proceedings of Machine Learning Research–nnn*, 1:30, 2023.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 801–818, 2018.

Chollet, F. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251–1258, 2017.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Crowston, K. Amazon mechanical turk: A research tool for organizations and information systems scholars. In *Shaping the Future of ICT Research. Methods and Approaches: IFIP WG 8.2, Working Conference, Tampa, FL, USA, December 13-14, 2012. Proceedings*, pp. 210–221. Springer, 2012.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results, 2012.

Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., and Sugiyama, M. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31, 2018.

He, T., Zhang, Z., Zhang, H., Zhang, Z., Xie, J., and Li, M. Bag of tricks for image classification with convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 558–567, 2019.

Hu, H., Xie, L., Du, Z., Hong, R., and Tian, Q. One-bit supervision for image classification. *Advances in Neural Information Processing Systems*, 33:501–511, 2020.

Hwang, S., Lee, S., Kim, H., Oh, M., Ok, J., and Kwak, S. Active learning for semantic segmentation with multi-class label query. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Jha, D., Smedsrud, P. H., Riegler, M. A., Halvorsen, P., de Lange, T., Johansen, D., and Johansen, H. D. Kvasir-seg: A segmented polyp dataset. In *MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26*, pp. 451–462. Springer, 2020.

Joshi, A. J., Porikli, F., and Papanikolopoulos, N. Multi-class active learning for image classification. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

Kasarla, T., Nagendar, G., Hegde, G. M., Balasubramanian, V., and Jawahar, C. Region-based active learning for efficient labeling in semantic segmentation. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1109–1117. IEEE, 2019.

Kim, H., Oh, M., Hwang, S., Kwak, S., and Ok, J. Adaptive superpixel for active learning in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 943–953, October 2023a.

Kim, K. I. Active label correction using robust parameter update and entropy propagation. In *European Conference on Computer Vision*, pp. 1–16. Springer, 2022.

Kim, Y.-Y., Cho, Y., Jang, J., Na, B., Kim, Y., Song, K., Kang, W., and Moon, I.-C. Saal: sharpness-aware active learning. In *Proc. International Conference on Machine Learning (ICML)*, pp. 16424–16440. PMLR, 2023b.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollar, P., and Girshick, R. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, October 2023.

Lad, V. and Mueller, J. Estimating label quality and errors in semantic segmentation data via any model. In *ICML Workshop on Data-centric Machine Learning Research*, 2023.

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

Ma, J., He, Y., Li, F., Han, L., You, C., and Wang, B. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.

Mackowiak, R., Lenz, P., Ghori, O., Diego, F., Lange, O., and Rother, C. Cereals-cost-effective region-based active learning for semantic segmentation. In *BMVC*, 2018.

Mahmood, R., Fidler, S., and Law, M. T. Low-budget active learning via wasserstein distance: An integer programming approach. In *International Conference on Learning Representations*, 2021.

Müller, N. M. and Markert, K. Identifying mislabeled instances in classification datasets. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2019.

Natarajan, N., Dhillon, I. S., Ravikumar, P. K., and Tewari, A. Learning with noisy labels. *Advances in neural information processing systems*, 26, 2013.

Northcutt, C., Jiang, L., and Chuang, I. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, 2021a.

Northcutt, C. G., Athalye, A., and Mueller, J. Pervasive label errors in test sets destabilize machine learning benchmarks. In *Proceedings of the 35th Conference on Neural Information Processing Systems Track on Datasets and Benchmarks*, December 2021b.

Patrini, G., Rozza, A., Krishna Menon, A., Nock, R., and Qu, L. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1944–1952, 2017.

Qu, C., Zhang, T., Qiao, H., Liu, J., Tang, Y., Yuille, A., and Zhou, Z. Abdomenatlas-8k: Annotating 8,000 ct volumes for multi-organ segmentation in three weeks. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

Ren, M., Zeng, W., Yang, B., and Urtasun, R. Learning to reweight examples for robust deep learning. In *International conference on machine learning*, pp. 4334–4343. PMLR, 2018.

Rottmann, M. and Reese, M. Automated detection of label errors in semantic segmentation datasets via deep learning and uncertainty quantification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3214–3223, 2023.

Ru, L., Zheng, H., Zhan, Y., and Du, B. Token contrast for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3093–3102, June 2023.

Saran, A., Yousefi, S., Krishnamurthy, A., Langford, J., and Ash, J. T. Streaming active learning with deep neural networks. In *Proc. International Conference on Machine Learning (ICML)*, pp. 30005–30021. PMLR, 2023.

Shin, G., Xie, W., and Albanie, S. All you need are a few pixels: semantic segmentation with pixelpick. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Siddiqui, Y., Valentin, J., and Nießner, M. Viewal: Active learning with viewpoint entropy for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9433–9443, 2020.

Sinha, S., Ebrahimi, S., and Darrell, T. Variational adversarial active learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5972–5981, 2019.

Smailagic, A., Costa, P., Noh, H. Y., Walawalkar, D., Khandelwal, K., Galdran, A., Mirshekari, M., Fagert, J., Xu, S., Zhang, P., et al. Medal: Accurate and robust deep active learning for medical image analysis. In *2018 17th IEEE international conference on machine learning and applications (ICMLA)*, pp. 481–488. IEEE, 2018.

Song, H., Kim, M., Park, D., Shin, Y., and Lee, J.-G. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

Van den Bergh, M., Boix, X., Roig, G., de Capitani, B., and Van Gool, L. Seeds: Superpixels extracted via energy-driven sampling. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VII 12*, pp. 13–26. Springer, 2012.

Wang, D., Zhang, J., Du, B., Xu, M., Liu, L., Tao, D., and Zhang, L. Samrs: Scaling-up remote sensing segmentation dataset with segment anything model. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

Wang, K., Zhang, D., Li, Y., Zhang, R., and Lin, L. Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 2016.

Wu, X., Chen, C., Zhong, M., and Wang, J. Hal: Hybrid active learning for efficient labeling in medical domain. *Neurocomputing*, 456:563–572, 2021.

Xiao, T., Xia, T., Yang, Y., Huang, C., and Wang, X. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2691–2699, 2015.

Yang, J., Wang, H., Wu, S., Chen, G., and Zhao, J. Towards controlled data augmentations for active learning. In *Proc. International Conference on Machine Learning (ICML)*, pp. 39524–39542. PMLR, 2023.

Yang, L., Zhang, Y., Chen, J., Zhang, S., and Chen, D. Z. Suggestive annotation: A deep active learning framework for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pp. 399–407. Springer, 2017.

Zhang, Y., Zhang, X., Xie, L., Li, J., Qiu, R. C., Hu, H., and Tian, Q. One-bit active query with contrastive pairs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9697–9705, 2022.

A. Text Prompts for Warm-start

In Section 3.1, for the warm-start process, we generate initial pseudo labels through a sequence of text prompts described. For example, we employ "Road. Sidewalk. Building. ... Bicycle." prompts for Cityscapes and "Aeroplane. Bicycle. Bird. ... Tvmonitor." for PASCAL, where each word aligns with the respective target class. However, each prompt, such as "Diningtable", can be segmented into multiple tokens, such as "Dining" and "table". Therefore, we assign each token to its corresponding class to derive the initial labels for the warm-start process.

B. User Study with Different Queries

To verify the efficiency of the proposed correction query C_{cor} in Active Label Correction (ALC) compared to classification query C_{cls} in conventional Active Learning (AL), we conduct a user study focusing on actual labeling costs, specifically annotation time. The example of the correction query questionnaire is illustrated in Figure 2, and the results are summarized in Table 1. Each question presents the user with instructions, an image with an object highlighted, and options for classifying the object. For the correction query scenario, the instructions include the pseudo label of the foundation model, and users only need to correct if the pseudo label is incorrect. The detailed instruction for correction query is given as follows:

*Is this pixel a **TV**?*
Give the correct label only if the pseudo label is incorrect.

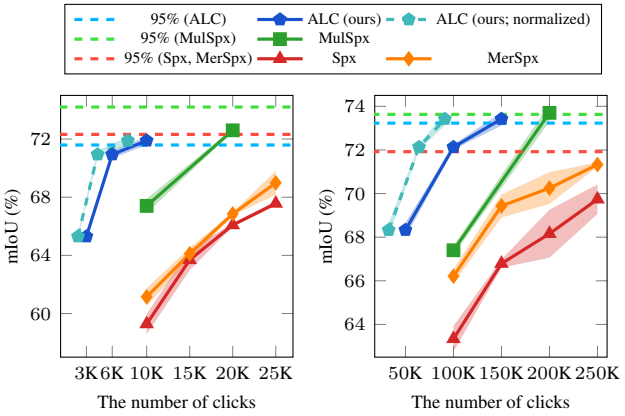
On the other hand, the example instruction for classification query is given as follows:

Give the correct label of the pixel.

Based on the ground-truth, we collect 20 images consisting of 10 images with correct pseudo labels, and 10 images with incorrect pseudo labels counterparts, i.e., $p = 0.5$. We ask for labels of these images in both correction queries and classification queries. A total of 20 volunteers participates in the survey. To prevent the user from memorizing images, we only ask one type of query per user, which means we ask the correction queries to 10 users and the classification queries to the others. The responses from annotators are evaluated by calculating the accuracy of the classification prediction. As shown in Table 1, the correction query only requires 75% labeling time of that of the classification query. In terms of accuracy, both queries show the same 95%.

C. Absolute Performance of ALC vs. AL

While conventional AL for semantic segmentation methods use the same DeepLab-v3+ (Chen et al., 2018) segmentation decoder combined with backbone pre-trained with the ImageNet (Deng et al., 2009) dataset, the architecture of their backbones are slightly different. ALC (ours) utilize plain ResNet101, MulSpx (Hwang et al., 2023) use ResNet101 combined with deepstem tricks (He et al., 2019), and MerSpx (Kim et al., 2023a) and Spx (Cai et al., 2021) employ Xception-65 (Chollet, 2017). Figure 3 presents the performance in terms of recovery rate relative to a fully supervised model, calculated as the ratio of our model's performance to that of the fully supervised model.



(a) PASCAL (b) Cityscapes

Figure 8: *Effect of active label correction.* Our ALC shows comparable results on both datasets with much fewer clicks. Our ALC (normalized) reflects the reduced budget of correction queries in Theorem 3.1.

Here, we additionally report the comparison with absolute mIoU in Figure 8 over various budget levels, represented by the number of clicks, for both PASCAL and Cityscapes datasets. The 95% performance of each baseline's fully supervised model is illustrated with a dashed line labeled as 95% (·). Our proposed ALC method consistently demonstrate the most efficient performance.

D. Ablation Studies

D.1. Initial Pseudo Labels

Table 7: *Performance of initial pseudo labels from Grounded-SAM.* Noisy pseudo labels cause the data and model mIoU to worsen.

Box-thresohld	# of objects	Data mIoU (%)	Model mIoU (%)
0.2	11,257	55.32	59.04
0.3	5,995	65.14	66.15
0.4	3,890	66.71	65.30
0.5	2,798	60.87	59.50

To evaluate the quality of pseudo labels generated by

(a) Unlabeled image (b) Grounded-SAM (c) PASCAL (d) PASCAL+ (ours)

Figure 9: *Additional examples of noisy and corrected labels in PASCAL.* We correct PASCAL into PASCAL+ utilizing the superpixels of Grounded-SAM.

Grounded-SAM (Liu et al., 2023) on the PASCAL dataset, we measure the data and model mIoU while adjusting a hyperparameter. Grounded-SAM operates with two hyperparameters: box-threshold and text-threshold. The text-threshold aims to identify all potential classes with a potential value exceeding the threshold. As we only focus on a specific class per a object, we employ the argmax function on the potential classes. The box-threshold determines the confidence level in the bounding box of the identified object. With a lower box-threshold, the foundation model can detect more objects, as demonstrated in Table 7. However, this often leads to numerous incorrectly labeled objects, resulting in decreased mIoU for both data and model. Yet, the benefit of detecting lots of objects lies in the potential for enhanced performance when correcting the pseudo labels of all detected objects, resulting in model mIoU of 72.59%, 70.90%, and 66.97% for box-thresholds of 0.2, 0.3, and 0.4, respectively.

D.2. Similarity Threshold for Label Expansion

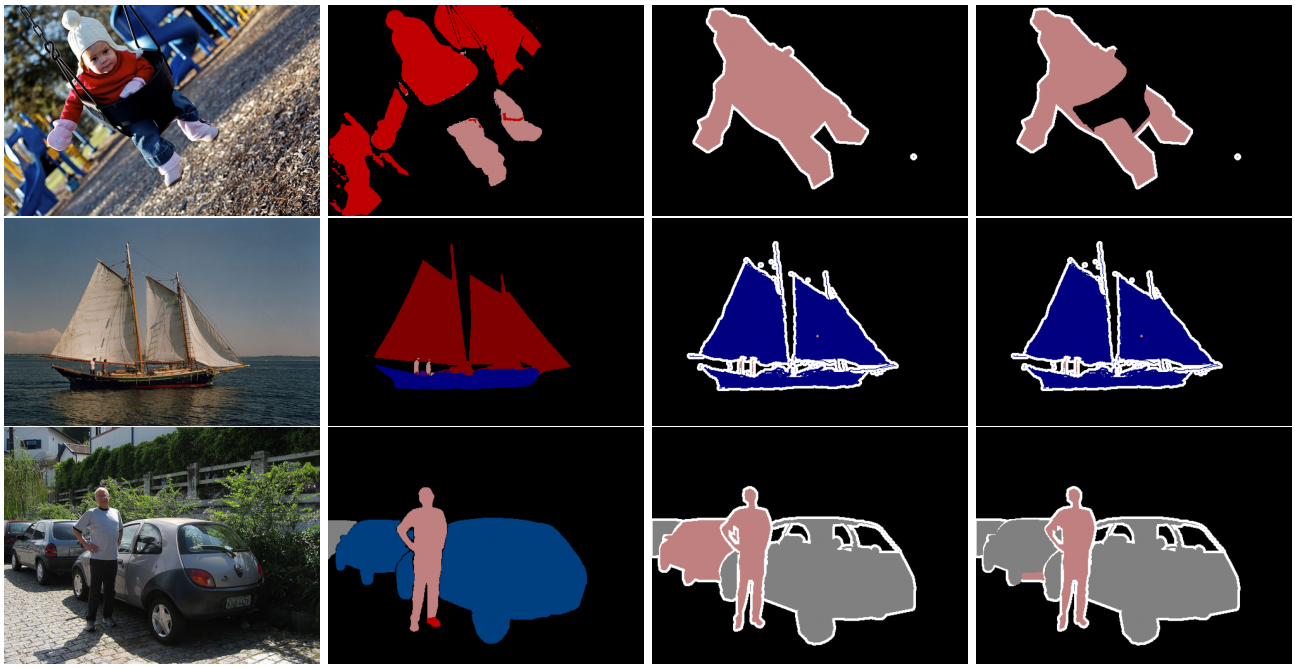
During the label expansion phase detailed in Section 3.4, a challenge can emerge when superpixels contain pixels belonging to various classes, potentially diminishing the dataset's overall quality. To this end, we propose expanding the clean label of a pixel x_i only to similar pixels within its corresponding superpixel s_i as follows:

$$s_i(x_i, \epsilon) := \{x \in s_i : \cos\big(f_\theta(x_i), f_\theta(x)\big) \geq \epsilon)\}\,, \quad (12)$$

where the degree of expansion is determined by hyperparameter ϵ. The more incomplete the superpixel, the larger ϵ is required. For our main experiments in Section 4, we set ϵ as 0, indicating complete expansion, where $s_i(x_i, \epsilon) = s_i$. Here, we investigate how the value of ϵ in (12) affects results. Since foundation models accurately generate superpixel boundaries in PASCAL, we observe that setting ϵ to 0, thereby allowing the corrected pixel label to cover the entire superpixel, yields the best performance, as demonstrated in Table 8.

Table 8: *Similarity threshold.* For correction, we select 5K pixels from the initial labels and adjust the extent of label expansion.

ϵ	Data mIoU (%)	Model mIoU (%)
0.0	83.34	68.71
0.2	82.85	68.48
0.4	82.11	68.58
0.6	81.17	68.48
0.8	80.18	68.32
1.0	55.61	56.05



(a) Unlabeled image (b) Grounded-SAM (c) PASCAL (d) PASCAL+ (ours)

Figure 10: *Uncorrectable examples of noisy and corrected labels in PASCAL.* We correct PASCAL into PASCAL+ utilizing the superpixels of Grounded-SAM, however, due to the inherent limitations of superpixels, some failure cases can be observed.

D.3. Comparison with Other Diversified Pixel Pool

Table 9: *Experiments for diversified pixel pools.* With 5K budgets, we select pixels from different pixel pools and correct the initial labels to PASCAL.

Methods	Data mIoU (%)	Model mIoU (%)
PixelPick	66.88	62.59
ALC	83.60	68.71

To solve the issue of picking similar pixels, as described in Section 3.3, PixelPick employs an acquisition function to rank all pixels, subsequently uniformly selecting them from the top 5% ranked pixels in each image (Shin et al., 2021). Thus, we contrast our diversified pixel pool based on superpixels with the PixelPick method. For a fair comparison, we incorporate all other techniques, including SIM acqusition equipped with the concept of look-ahead and label expansion. As shown in Table 9, our ALC performs better than PixelPick in terms of both data and model mIoU.

D.4. Comparison with Other Acquisitions

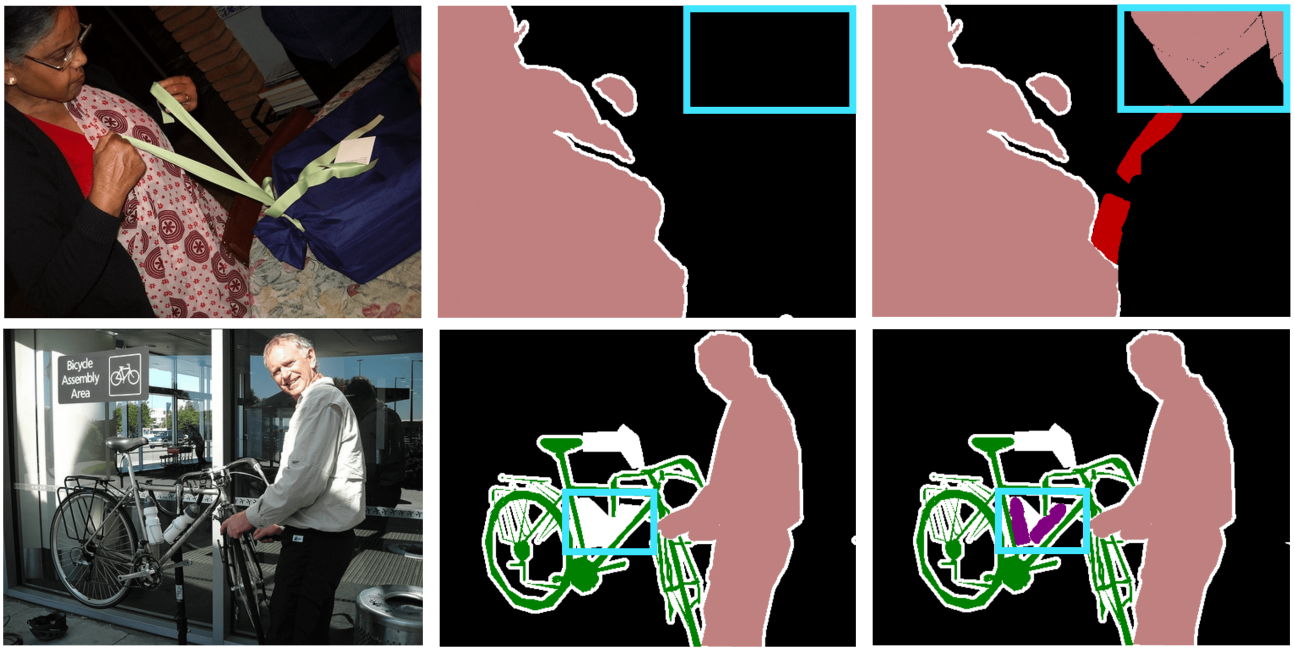
In Table 10, our SIM acquisition outperforms other various acquisitions including Entropy, Best-versus-Second-Best (BvSB), and Class-Balanced (ClassBal), employed in active learning, due to the incorporation of the look-ahead concept. We concentrate on adjusting the acquisition function, while simultaneously applying other techniques such as diversified pixel pool and expansion techniques. We correct the labels of 3K pixels selected using various acquisition functions, and expand the labels to their corresponding superpixels.

Table 10: *Experiments with other acquisitions.* With 5K budgets, we select pixels from different pixel pools and correct the initial labels to PASCAL.

Methods	Model mIoU (%)
Entropy	57.09 ± 0.40
BvSB	57.58 ± 0.41
ClassBal	57.51 ± 0.67
SIM	65.30 ± 0.21

D.5. Class IoU on PASCSAL

We provide the rationale of IoU gain in Figure 7a. For a detail, thanks to the corrected PASCAL+, we observe that the IoU values of pottedplant, sofa, chair, and diningtable classes increase. This is related to the number of corrections in Figure 7b, as those class are corrected lots than other classes. However, in case of background and person classes,



(a) Image (b) PASCAL (c) PASCAL+

Figure 11: *Correction that appears to cause negative IoU gains.* Here, the colors black, red, purple, green, and pink represent the background, chair, bottle, bicycle, and person classes, respectively.

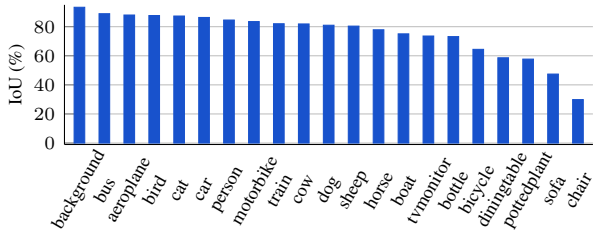


Figure 12: *Class IoU on PASCAL.* The IoU values of diningtable, pottedplant, sofa, and, chair classes are relatively low when trained with PASCAL.

we cannot obtain IoU gain as those classes already attain high IoU with PASCAL as depicted in Figure 12.

E. Additional Results of PASCAL+

E.1. Qualitative Results

Additional qualitative results of corrected labels using our proposed method are depicted in Figure 9. These results demonstrate that our proposed correction method effectively identifies objects overlooked in the original labels.

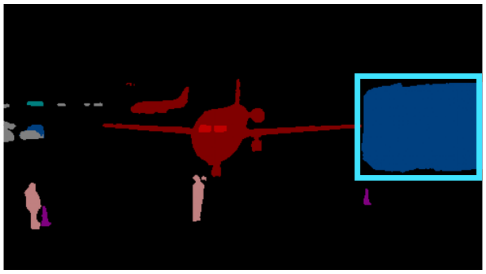
E.2. Uncorrectable Cases

Figure 10 presents examples where corrections made by our proposed method are not entirely successful. Specifically, the examples in the first and second rows of Figure 10 illustrate situations where annotators mistakenly assign pixel clicks to the wrong classes. Such errors can occur under limited budgets. In the last row of Figure 10, an area mislabeled as person class is effectively corrected to car class. However, due to the insufficient granularity of the superpixels, small areas remain uncorrected. This limitation can be mitigated by employing more refined superpixels or utilizing improved foundational models.

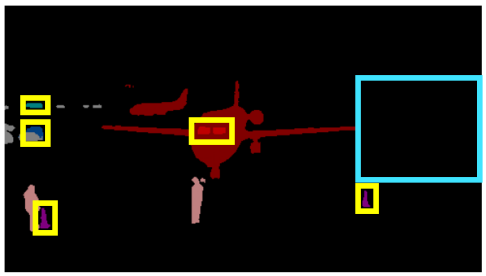
E.3. Negative IoU Gains of PASCAL+

Figure 6a represents negative IoU gains for certain classes such as person, bottle, and cow. Here, we provide the rationale for these negative gains. The final IoU gain is determined by the positive and negative impacts of corrections. Although corrections generally aim to reduce noisy labels, yielding positive effects, they can also have negative effects, especially on challenging objects, as shown in Figure 11.

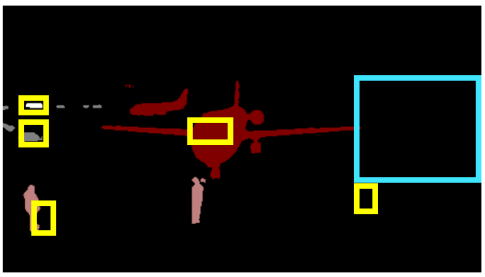
(a) Unlabeled image



(b) Round 0 (Grounded-SAM)



(c) Round 1



(d) Round 2

Figure 13: *Segmentation changes through active label correction.* (b) The initial pseudo labels obtained from Grounded-SAM contain numerous noisy labels, exemplified by instances like tvmonior inside the cyan box. (c) In the first round, the object labeled as tvmonitor is corrected to background. Nonetheless, many noisy labels exist within the yellow boxes. (c) In the second round, we rectify all remaining noisy labels. With the help of the proposed look-ahead acquisition function, we prioritize correcting large objects before addressing small ones. Here, the colors black, blue, red, dark red, purple, and pink represent the background, tvmonitor, chair, airplane, bottle and person classes, respectively.