A^{2}LC: Active and Automated Label Correction for Semantic Segmentation

Youjin Jeon*, Kyusik Cho*, Suhan Woo, Euntai Kim†

Yonsei University
{ngbrjyj, ks.cho, wsh112, etkim}@yonsei.ac.kr

Abstract

Active Label Correction (ALC) has emerged as a promising solution to the high cost and error-prone nature of manual pixel-wise annotation in semantic segmentation, by actively identifying and correcting mislabeled data. Although recent work has improved correction efficiency by generating pseudo-labels using foundation models, substantial inefficiencies still remain. In this paper, we introduce A^{2}LC, an Active and Automated Label Correction framework for semantic segmentation, where manual and automatic correction stages operate in a cascaded manner. Specifically, the automatic correction stage leverages human feedback to extend label corrections beyond the queried samples, thereby maximizing cost efficiency. In addition, we introduce an adaptively balanced acquisition function that emphasizes underrepresented tail classes, working in strong synergy with the automatic correction stage. Extensive experiments on Cityscapes and PASCAL VOC 2012 demonstrate that A^{2}LC significantly outperforms previous state-of-the-art methods. Notably, A^{2}LC exhibits high efficiency by outperforming previous methods with only 20% of their budget, and shows strong effectiveness by achieving a 27.23% performance gain under the same budget on Cityscapes.

Code — https://github.com/ngbrjyj/A2LC

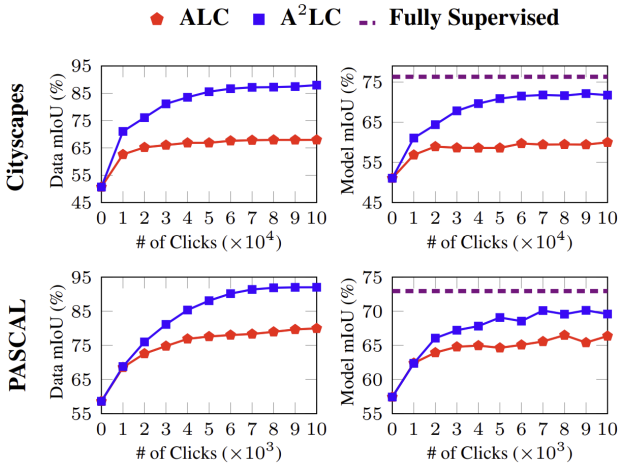


Figure 1: Effectiveness of A^{2}LC framework. Our model demonstrates clear effectiveness, saturating by round 6 on Cityscapes and round 5 on PASCAL. At these points, it outperforms the baseline by 11.81 and 4.44 mIoU and reaches 94% and 95% of full supervision.

Introduction

The high cost and labor-intensive nature of pixel-wise annotation pose a significant challenge in semantic segmentation, where large scale labeled datasets are essential for deep learning models (Edlund et al. 2021; Li et al. 2023a). Furthermore, such datasets frequently contain noisy labels due to manual annotation errors, which hinder the learning process of deep neural networks (Li et al. 2023b; Plank 2022). In response to this challenge, active label correction methods (Kim et al. 2024; Kim 2022; Kremer, Sha, and Igel 2018) offer a promising solution, as they iteratively identify and refine noisy labels through human review.

While active label correction has been actively explored in other vision tasks, its application to semantic segmentation remains limited due to the inherent difficulty of correcting pixel-wise labels. Recent advancements in foundation models (Lüddecke and Ecker 2022; Ren et al. 2024; Zou et al. 2023) have helped mitigate these challenges by enabling zero-shot prediction, generating high quality pseudo-labels that can serve as a better starting point for label refinement. Building on these advancements, Kim et al. (2024) developed a framework that generates pseudo-labels using a foundation model for semantic segmentation and actively selects uncertain pixels for correction by annotators.

This framework effectively incorporates a pretrained deep neural network into the conventional active learning paradigm, establishing an efficient pipeline that significantly reduces the need for human intervention. Nevertheless, we found that their integration of deep learning techniques remained incomplete, particularly in addressing redundancy during correction. Specifically, prior methods rely exclusively on annotators for label correction, and the corrected labels are not generalizable beyond the selected samples. Consequently, many similar pixels are repeatedly sampled

*These authors contributed equally.
†Corresponding author.

and corrected, leading to inefficient querying and increased annotation costs. This redundancy ultimately slows down performance improvement, as illustrated in Figure 1.

To address this limitation, we propose an additional automatic correction stage that exploits annotator feedback to perform label correction beyond the queried samples. Unlike conventional pipelines that rely solely on manual correction, our method incorporates a secondary correction source, the Label Correction Module (LCM), forming a dual-source correction framework. The primary advantage of the LCM lies in its ability to maximize the utility of human-provided corrections without incurring additional annotation overhead, thereby improving cost-efficiency. Furthermore, we introduce an adaptively balanced acquisition function, designed based on the pseudo-label statistics. This function prioritizes tail classes during the querying process, thereby enabling efficient mitigation of the class imbalance problem. Notably, our proposed acquisition function synergizes well with the LCM, as it directs annotator effort toward a subset of samples targeted for tail classes. This, in turn, guides the LCM's automatic label correction toward improved class balance, enhancing overall annotation efficiency. Extensive experiments on Cityscapes (Cordts et al. 2016) and PASCAL VOC 2012 (Everingham et al. 2015) demonstrate that A^{2}LC significantly outperforms previous state-of-the-art methods, as shown in Figure 1.

Our contributions can be summarized as follows:

- We develop a label correction module that enables automatic correction and maximizes the utility of human effort, significantly improving cost efficiency.
- By introducing an adaptive class weight and integrating it into the acquisition function, we effectively address the class imbalance challenge.
- A^{2}LC achieves high efficiency by outperforming previous methods while using only 20% and 60% of their budget on Cityscapes and PASCAL, respectively.
- A^{2}LC demonstrates strong effectiveness by yielding 27.23% and 14.30% performance improvements under equivalent budget constraints on Cityscapes and PASCAL, respectively.

Related Works

Active Label Correction

Active label correction aims to construct a clean labeled dataset with minimal cost by selectively querying and correcting samples that are likely to be mislabeled. As cost is directly tied to the query unit, extensive research has explored its optimal label query unit. Early studies used image-level querying (Chen et al. 2024; Yang et al. 2023), but this led to excessive labeling overhead as entire images required annotation regardless of the informativeness of specific regions. To improve efficiency, pixel-level querying (Didari et al. 2024; Ma, Karakus, and Rosin 2025; Rückin et al. 2024a,b; Schachtsiek, Rossi, and Hannagan 2023) was introduced, though it still incurred substantial costs. A two-step querying approach (Ribeiro Marnet et al. 2024; van Marrewijk et al. 2024) refined this by first selecting uncertain images and then querying only the most informative pixels. More recently, superpixel-level querying (Cai et al. 2021; Ge et al. 2024; Hwang et al. 2023; Kim et al. 2023; Wu et al. 2025) has been explored to balance granularity and efficiency. Due to the practical challenge of obtaining perfectly clean labels, active label correction has been increasingly applied across various tasks, e.g., text classification (Hou et al. 2025; Taneja and Goel 2024; Wang et al. 2024), image classification (Ekambaram et al. 2016; Beck et al. 2024; Bernhardt et al. 2022; Khanal et al. 2024; Kim 2022; Kremer, Sha, and Igel 2018; Li et al. 2022; Rebbapragada et al. 2012; Wang et al. 2024), and semantic segmentation (Kim et al. 2024). Despite the integration of ALC into semantic segmentation, prior studies have not optimized annotation cost efficiency, as they assign individual annotation costs to each similar pixel. We effectively address this issue by proposing a module that performs automatic correction for similar pixels. To the best of our knowledge, this is the first study to introduce automatic label correction in semantic segmentation.

Acquisition Function

Since accurately identifying mislabeled data is essential to avoid unnecessary expenditure, the acquisition function is crucial for the success of active label correction. It can be broadly categorized into three types. Uncertainty-based functions prioritize sampling data with high uncertainty, e.g., BALD (Gal, Islam, and Ghahramani 2017), Confidence (Wang and Shang 2014), Entropy (Safaei et al. 2024), Margin (Roth and Small 2006), MeanSTD (Kampffmeyer, Salberg, and Jenssen 2016). Diversity-based functions aim to sample data that effectively represent the overall dataset, e.g., BADGE (Ash et al. 2020), Cluster-Margin (Citovsky et al. 2021), CoreSet (Sener and Savarese 2018). Hybrid functions consider both uncertainty and diversity, e.g., CLAUS (Rana and Rawat 2023), which performs frame selection based on model uncertainty and diverse video sampling through deep clustering. In this work, we effectively sample informative data by incorporating our newly defined adaptive class weight into the acquisition function, which dynamically modulates diversity consideration based on dataset imbalance.

A^{2}LC Framework

Overview

An overview of our proposed framework is shown in Figure 2. We construct the initial dataset $\mathcal{D}_0 = \{(X_i, \hat{Y}_i)\}_{i=1}^{N}$ using zero-shot predictions from the foundation model (Ren et al. 2024) as pseudo-labels, and initialize the model θ_0. The set of masks generated by SAM, denoted as $\mathcal{M}_0 = \{(m_i, \hat{y}_i)\}_{i=1}^{M}$, serves as the correction unit. At round r, all masks are scored based on an acquisition function computed by the model θ_{r-1}. The top-B most informative masks are sampled and their true labels are obtained by querying an annotator. Whereas prior research concluded the correction process at this manual stage, we cascade an additional automatic correction stage for further refinement. Specifically, the model ψ_r in Label Correction Module (LCM) is trained using the set of queried masks $\mathcal{M}_r^Q$, which contains only

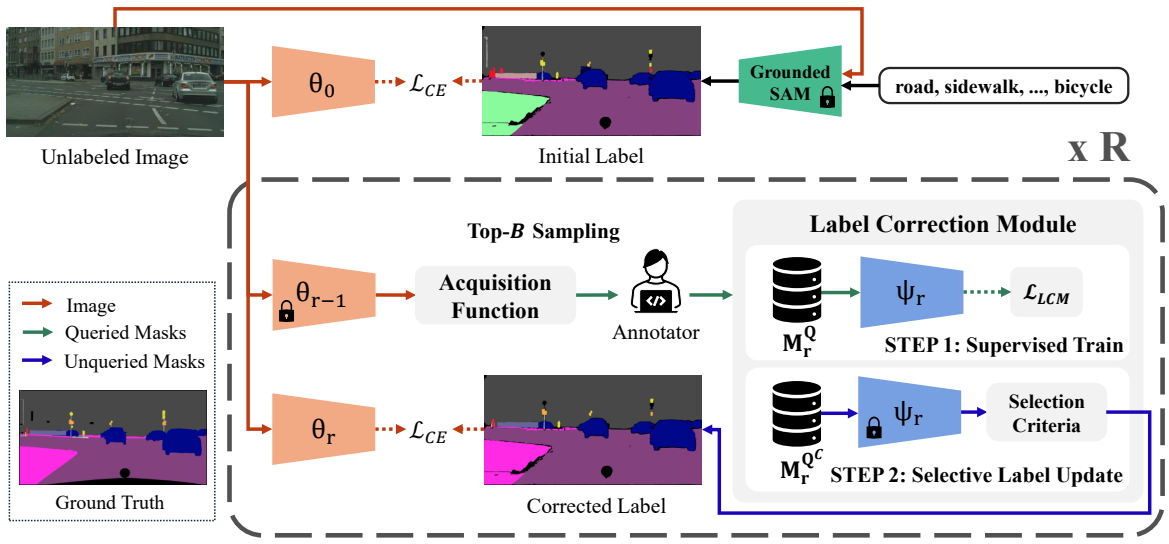


Figure 2: Overview of A^{2}LC framework. We execute Grounded SAM on unlabeled images to generate initial pseudo-labels. For each of the R rounds, the model θ_{r-1} trained with pseudo-labels selects B masks via the acquisition function to query for manual correction. Then, the model ψ_r in Label Correction Module (LCM) is trained using the queried masks and corrects the labels of unqueried masks. After both manual and automatic corrections, the pseudo-labels and model are updated, completing a single round of the correction cycle.

clean labels obtained from the manual correction stage. Subsequently, the trained ψ_r performs automatic label correction by inferring the unqueried masks $\mathcal{M}_r^{Q^c} = \mathcal{M}_{r-1} \backslash \mathcal{M}_r^Q$. This secondary correction is applied only to the masks that satisfy all selection criteria. Once both the manual and automatic correction stages are completed, the model θ_r is updated using the corrected dataset $\mathcal{D}_r$. The overall procedure is then iteratively repeated until the allocated budget is exhausted.

The following sections introduce two key techniques that constitute our Active and Automated Label Correction (A^{2}LC) framework. First, we describe the automatic label correction stage, which plays a crucial role in enhancing cost efficiency. Next, we introduce a novel acquisition function that strengthens both correction stages through data-driven balanced sampling.

Label Correction Module

An analysis of prior studies (Xu et al. 2023, 2021) demonstrates that mislabeling predominantly arises between semantically similar categories, such as *traffic sign* and *traffic light*, rather than between disparate classes. This occurs because the foundation model, trained on broad and diverse domains, often fails to capture the subtle distinctions required in fine-grained domains. As a result, it inevitably generates biased pseudo-labels, and pixels with similar features naturally cluster within comparable acquisition score ranges. This leads to redundant queries of similar masks across rounds, resulting in inefficient use of annotation resources. To address this issue, we propose a module that utilizes annotator-provided true labels to automatically correct noisy labels exhibiting similar features, enabling the correction of a more diverse set of noisy labels in the following rounds. This module is executed through two sequential steps in each round: the learning phase and the correction phase. The overall pipeline of our proposed LCM is illustrated in Figure 3.

Components of LCM. The LCM is built upon two core components: a learnable model and selection criteria. The model ψ_r at round r takes the mask feature $f_{\theta_{r-1}}(m)$ as input, defined as

$$f_{\theta_{r-1}}(m) = \frac{1}{|m|} \sum_{x \in m} f_{\theta_{r-1}}(x), \tag{1}$$

where the feature $f_{\theta_{r-1}}(x)$ of pixel x is extracted by the model θ trained at round $r-1$, and $f_{\theta_{r-1}}(m)$ denotes the average feature over all pixels $x \in m$. For simplicity, we refer to mask feature $f_{\theta_{r-1}}(m)$ as m. A four-layer fully connected network with ReLU activations reduces input features to 256, 128, and 64 dimensions, followed by a final classification layer with softmax activation to produce class probabilities. The selection criteria are designed to identify predictions with high reliability, so that automatic label updates are applied exclusively to the subset of confidently inferred labels.

STEP 1: Supervised Train with Queried Masks. In the learning phase, the masks queried in round r, denoted as $m \in \mathcal{M}_r^Q$ are used to train ψ_r under supervision, with a weighted cross-entropy loss:

$$\mathcal{L}_{LCM} := \lambda_{y(m)} \cdot \mathcal{L}_{CE}(y(m), \psi_r(m)), \tag{2}$$

where $y(m)$ denotes the annotator-corrected label of mask m, and $\lambda_{y(m)}$ represents the corresponding normalized class

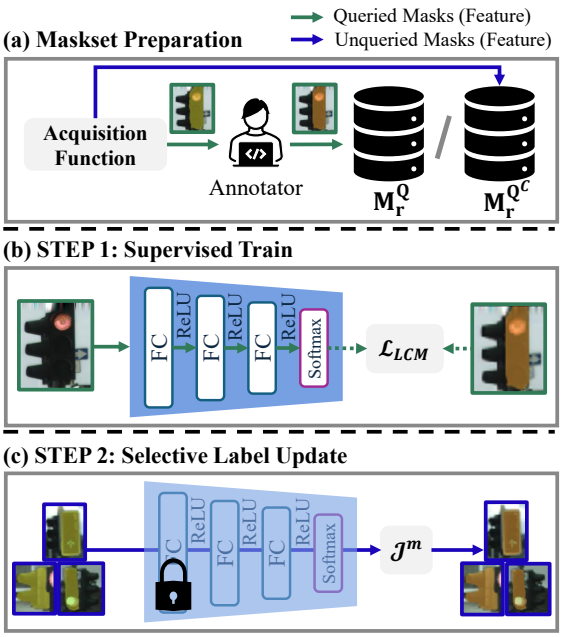


Figure 3: LCM pipeline. LCM operates in two sequential steps: the model is first trained with accurately labeled masks queried in the current round, followed by correcting potentially mislabeled masks based on the model's predictions. Here, a *traffic light (orange)* mask, mislabeled as a *traffic sign (yellow)*, was corrected by the annotator. This manually corrected mask is then used to train the model, enabling automatic refinement of similar cases. As a result, three additional similar masks were automatically corrected.

weight. The weight λ_k for class k is defined as $\lambda_k := \frac{N/N_k}{\sum_{c\in\mathcal{C}} N/N_c}$, where N_k and N denote the number of training samples in class k and the total number of training samples, respectively.

STEP 2: Selective Label Update for Unqueried Masks. In the correction phase, the model trained during the learning phase attempts to correct all unqueried masks. Label prediction is performed solely on the masks $m \in \mathcal{M}_r^{Q^c}$ unqueried in round r, using model ψ_r trained in the previous step. However, direct label updates to predicted classes pose a risk of erroneous correction, as class imbalance in $\mathcal{M}_r^Q$ can lead to biased learning. To mitigate this, we introduce a conservative label update strategy based on three selection criteria. A mask is eligible for automatic correction only if all three of the following conditions are satisfied: (1) the model's prediction confidence exceeds a confidence threshold τ, (2) the predicted class is not among the tail classes, and (3) the original pseudo-label is also not from a tail class. These three constraints jointly serve to minimize the risk of incorrect label updates. We define the overall selection criteria $\mathcal{J}^m$ for mask m as:

$$\mathcal{J}^m = \mathcal{J}_1^m \wedge \mathcal{J}_2^m \wedge \mathcal{J}_3^m, \tag{3}$$

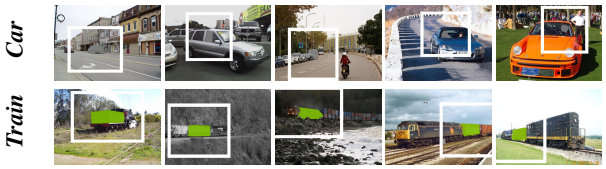


Figure 4: Visualization of automatically corrected masks for the *car* and *train* classes.

$$\mathcal{J}_1^m = \mathbb{I}\Big(\max_{c\in\mathcal{C}} \psi_r(c;m) \geq \tau\Big), \tag{4}$$

$$\mathcal{J}_2^m = \mathbb{I}\Big(\hat{y}_{\psi_r}(m) \notin \{c \mid \text{rank}(c) \geq (1-\alpha)\cdot|\mathcal{C}|,\ c\in\mathcal{C}\}\Big), \tag{5}$$

$$\mathcal{J}_3^m = \mathbb{I}\Big(\hat{y}(m) \neq \arg\max_{c\in\mathcal{C}} \text{rank}(c)\Big), \tag{6}$$

where $\mathbb{I}(\cdot)$ is the indicator function, $\text{rank}(c)$ represents the index of class c after sorting all classes by sample count in descending order, and α is a hyperparameter specifying the threshold ratio for tail classes (e.g., $\alpha = 0.5$ corresponds to the bottom 50% of classes). We set $\alpha = 0.5$, and τ is initialized at 0.99 and incrementally increased for careful correction over time. Figure 4 visualizes the masks corrected by LCM.

Adaptively Balanced Acquisition Function

Based on the assumption that the label of a pixel predicted with low confidence by the model is more likely to be incorrect, numerous acquisition functions (Kim et al. 2024; Roth and Small 2006; Safaei et al. 2024; Wang and Shang 2014) have been proposed. However, directly applying the conventional acquisition functions results in marginal performance improvements due to a sampling bias toward head classes, as illustrated in Figure 5. Moreover, since the model within the LCM is trained on the data sampled by the acquisition function, class imbalance in the sampled data can also degrade the overall effectiveness of the LCM. To address this, we introduce an adaptive class weight $w(x)$, applied multiplicatively to the base acquisition function. This factor dynamically adjusts the degree of balanced sampling according to the pseudo-label statistics at each round. The weight consists of two components: the class rarity score and dataset imbalance score. For simplicity, we refer to $\mathcal{M}_{r-1}$ as $\mathcal{M}$ and θ_{r-1} as θ.

Class Rarity Score. The class rarity score $\hat{w}(x)$ is designed to prioritize tail classes and is defined as follows:

$$\hat{w}(x) := \frac{\min_{c\in\mathcal{C}} |\{x' \in \mathcal{M} : \hat{y}(x') = c\}|}{|\{x' \in \mathcal{M} : \hat{y}(x') = \hat{y}(x)\}|}, \tag{7}$$

where $\hat{y}(x)$ is the pseudo-label assigned to pixel x, and $\mathcal{C}$ denotes total classes. $\hat{w}(x)$ is a class-wise value assigned to each pixel within the range (0, 1], where pixels labeled with less frequent classes in the dataset receive values closer to 1, and those with more frequent classes receive values closer to 0. In particular, a pixel labeled with the rarest class is assigned a class rarity score of 1.

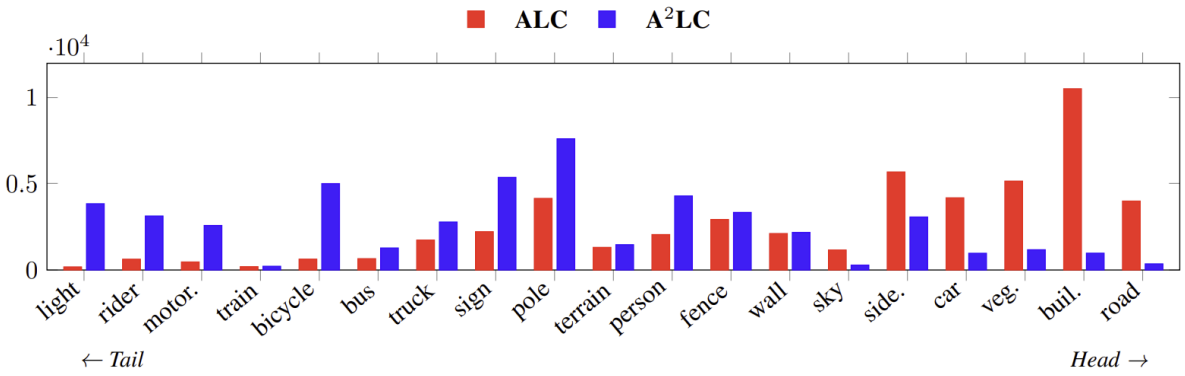


Figure 5: Class distribution of sampled data. The x-axis shows classes sorted by pseudo-label frequency, and the y-axis shows the number of sampled masks. Unlike the baseline, which largely concentrates on head classes, our proposed ABC acquisition function substantially increases the sampling of tail classes.

Dataset Imbalance Score. While the previously defined $\hat{w}(x)$ effectively emphasizes tail classes by reflecting their relative frequency within the dataset, it does not consider the overall distributional characteristics of the dataset. As a result, relying solely on $\hat{w}(x)$ may lead to suboptimal weighting, particularly in datasets with low levels of class imbalance. To address this limitation, we introduce a dataset imbalance score that modulates the strength of $\hat{w}(x)$ based on the dataset class distribution. We measure dataset imbalance using the KL divergence between the dataset's pixel distribution $\mathbb{P}_{\text{dist}}$ and a uniform distribution $\mathbb{U}_{\text{dist}}$:

$$\text{KL}(\mathbb{P}_{\text{dist}} \parallel \mathbb{U}_{\text{dist}}) = \sum_{c \in C} \mathbb{P}(c) \log \frac{\mathbb{P}(c)}{\mathbb{U}(c)}, \tag{8}$$

where $\mathbb{P}(c) = \frac{|\{x \in \mathcal{M} : \hat{y}(x) = c\}|}{|\{x \in \mathcal{M}\}|}$ quantifies the proportion of pixels $x \in \mathcal{M}$ whose pseudo-label $\hat{y}(x)$ belongs to class c, $\mathbb{U}(c) = \frac{1}{|C|}$ represents the uniform class distribution, and C denotes the total classes.

Adaptive Class Weight. Finally, we define an adaptive class weight $w(x)$ that emphasizes tail class pixels based on current pseudo-label statistics updated at every round:

$$w(x) := \hat{w}(x)^{\text{KL}^3(\mathbb{P}_{\text{dist}} \| \mathbb{U}_{\text{dist}})}, \tag{9}$$

ensuring $w(x) \in (0, 1]$. The magnitude of $\text{KL}(\mathbb{P}_{\text{dist}} \parallel \mathbb{U}_{\text{dist}})$ controls the strength of $\hat{w}(x)$, where its larger value in the highly imbalanced datasets amplifies $\hat{w}(x)$, guiding the correction process toward tail classes, whereas the smaller value in mildly imbalanced datasets suppresses $\hat{w}(x)$, reducing its effect.

Adaptively Balanced CIL. Confidence In Label (CIL) formulates the label quality of each pixel as the model's predictive confidence for a given label (Lad and Mueller 2023). To encode the CIL of a mask into a single representative pixel, Kim et al. (2024) introduce a SIM acquisition function by computing a weighted sum of the CILs of pixels within each mask based on their cosine similarity. Building upon this, we propose Adaptively Balanced CIL (ABC), which addresses the class imbalance issue by incorporating the adaptive class weight $w(x)$ formulated in Eq. (9).

$$a_{\text{CIL}}(x; \theta) := 1 - f_\theta(\hat{y}(x); x), \tag{10}$$

$$a_{\text{ABC}}(x; \theta) := w(x) \cdot a_{\text{CIL}}(x; \theta). \tag{11}$$

The acquisition function for each mask is defined as the weighted sum of pixel-level acquisition scores $a_{\text{ABC}}(x; \theta)$.

$$a_{\text{ABC}}(m; \theta) := \sum_{x \in m} \frac{f_\theta(x) \cdot f_\theta(m')}{\|f_\theta(x)\| \|f_\theta(m')\|} \cdot a_{\text{ABC}}(x; \theta), \tag{12}$$

where the $m' := \{x \in m : y_\theta(x) = \mathcal{D}_\theta(m)\}$ denotes the subset of pixels x in mask m whose predicted label $y_\theta(x)$ matches the pseudo dominant label $\mathcal{D}_\theta(m)$. The pseudo dominant label $\mathcal{D}_\theta(m) = \arg\max_{c \in \mathcal{C}} |\{x \in m : y_\theta(x) = c\}|$ is defined as the most frequently predicted class by the model θ within mask m (Kim et al. 2023, 2024). The feature $f_\theta(x)$ is the model representation at pixel x, and $f_\theta(m')$ is the average feature over m'. Finally, the top-B most informative masks, $\mathcal{M}_r^Q$, are sampled through the ABC acquisition function as follows:

$$\mathcal{M}_r^Q := \arg\max_{m \in \mathcal{M}}^B a_{\text{ABC}}(m; \theta), \tag{13}$$

where the superscript B denotes the budget assigned for the current round r.

Experiments

Experimental Setup

Baselines. We compare our framework primarily with ALC (Kim et al. 2024), the first state-of-the-art approach to introduce active label correction for semantic segmentation, as well as with state-of-the-art segmentation methods incorporating active learning, namely Spx (Cai et al. 2021), MerSpx (Kim et al. 2023), and MulSpx (Hwang et al. 2023). Our proposed acquisition function is evaluated with a broad range of acquisition functions, including Random, Entropy (Safaei et al. 2024), Margin (Roth and Small 2006), CIL (Lad and Mueller 2023), LCIL (Lad and Mueller

Dataset	mIoU (%)	Methods	Init.	1R	2R	3R	4R	5R
Cityscapes	Data	ALC	$50.68_{\pm 0.00}$	$62.28_{\pm 1.25}$	$65.36_{\pm 0.56}$	$66.19_{\pm 0.60}$	$66.87_{\pm 0.29}$	$67.01_{\pm 0.30}$
		A^{2}LC (Ours)	$50.68_{\pm 0.00}$	$\mathbf{70.01}_{\pm 1.05}$	$\mathbf{75.60}_{\pm 0.58}$	$\mathbf{80.84}_{\pm 0.53}$	$\mathbf{82.69}_{\pm 0.92}$	$\mathbf{85.26}_{\pm 0.55}$
	Model	ALC	$51.55_{\pm 0.71}$	$56.61_{\pm 0.42}$	$58.27_{\pm 0.63}$	$58.58_{\pm 0.08}$	$58.52_{\pm 0.26}$	$58.59_{\pm 0.05}$
		A^{2}LC (Ours)	$51.55_{\pm 0.71}$	$\mathbf{60.83}_{\pm 0.66}$	$\mathbf{63.89}_{\pm 0.44}$	$\mathbf{67.50}_{\pm 0.31}$	$\mathbf{68.87}_{\pm 0.78}$	$\mathbf{70.51}_{\pm 0.32}$
PASCAL	Data	ALC	$58.63_{\pm 0.00}$	$\mathbf{68.19}_{\pm 0.34}$	$72.72_{\pm 0.16}$	$74.84_{\pm 0.29}$	$76.41_{\pm 0.70}$	$77.06_{\pm 0.73}$
		A^{2}LC (Ours)	$58.63_{\pm 0.00}$	$67.49_{\pm 2.03}$	$\mathbf{74.88}_{\pm 1.45}$	$\mathbf{80.88}_{\pm 0.83}$	$\mathbf{84.81}_{\pm 0.48}$	$\mathbf{88.08}_{\pm 0.44}$
	Model	ALC	$56.94_{\pm 0.44}$	$\mathbf{62.11}_{\pm 0.61}$	$\mathbf{64.12}_{\pm 0.31}$	$64.15_{\pm 0.68}$	$65.00_{\pm 0.33}$	$65.48_{\pm 0.83}$
		A^{2}LC (Ours)	$56.94_{\pm 0.44}$	$60.87_{\pm 2.93}$	$64.08_{\pm 2.43}$	$\mathbf{66.45}_{\pm 0.95}$	$\mathbf{67.76}_{\pm 0.06}$	$\mathbf{68.42}_{\pm 0.87}$

Table 1: Quantitative results across multiple rounds. 'Init.' represents the performance of the initial pseudo-labels generated by Grounded SAM, before any label correction is performed.

Methods	Model mIoU (%)
Fully Supervised	76.31
Spx (Cai et al. 2021)	63.77
MerSpx (Kim et al. 2023)	66.53
MulSpx (Hwang et al. 2023)	66.60
ALC (Kim et al. 2024)	70.71
A^{2}LC (Ours)	**72.37**

Table 2: Quantitative comparison with state-of-the-art superpixel-based active learning methods.

2023), AIoU (Rottmann and Reese 2023), BvSB, and ClassBal from (Cai et al. 2021), MerSpx (Kim et al. 2023), and SIM (Kim et al. 2024).

Evaluation Metrics. Following previous work (Kim et al. 2024), we adopt mean Intersection over Union (mIoU) as the primary evaluation metric. This metric is used in two distinct contexts: Data mIoU and Model mIoU. Data mIoU denotes the mIoU between pseudo-labels and the ground truth, whereas Model mIoU refers to the mIoU between model predictions and the ground truth on the validation set. In addition, we employ Overall Accuracy (OA) to evaluate the accuracy of LCM by measuring the correctness of its automatically corrected masks.

Correction Details. A^{2}LC achieves further efficiency through two key strategies: (1) mask-level correction and (2) non-redundant correction. Previous studies (Kim et al. 2024) perform pixel-level correction by sampling highly unconfident pixels, querying them for ground truth labels, and propagating the labels to similar pixels, referred to as label expansion. However, this label expansion technique can lead to erroneous corrections, as many pixels depend on a single queried pixel. We address this limitation in a simple but effective manner by performing mask-level correction, where the mask is the prediction unit of the foundation model (Ren et al. 2024) employed in generating initial pseudo-labels. We further enhance efficiency through non-redundant correction, by excluding previously queried masks from the set of candidates for label correction. Non-redundant correction is highly synergistic with mask-level correction in enhancing cost efficiency. As mask-level correction aggregates multiple pixels into a single query, the number of unqueried pixels becomes relatively small after several rounds of label correction, which maximizes the effect of non-redundant correction.

Dataset	# of Corrected Masks	OA (%) Before	OA (%) After
Cityscapes	22,105	19.13	**60.99**
PASCAL	1,629	4.85	**82.26**

Table 3: Accuracy analysis of label correction module.

Acquisition Function	Cityscapes	PASCAL
Init.	50.68	58.63
Random	63.78	68.84
Entropy (Safaei et al. 2024)	58.12	59.95
Margin (Roth and Small 2006)	62.68	58.72
CIL (Lad and Mueller 2023)	61.51	59.29
LCIL (Kim et al. 2024)	62.63	59.80
AIoU (Rottmann and Reese 2023)	60.59	59.15
BvSB (Cai et al. 2021)	64.35	59.60
ClassBal (Cai et al. 2021)	65.65	60.56
MerSpx (Kim et al. 2023)	65.82	60.13
SIM (Kim et al. 2024)	78.98	**87.32**
ABC (Ours)	**84.92**	87.24

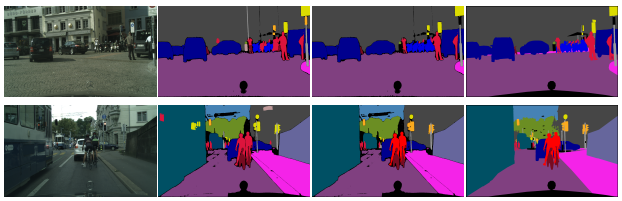
Table 4: Comparative analysis of adaptively balanced acquisition function (Data mIoU, %).

Semantic Segmentation Performance

In Table 1, we benchmark our proposed method under extremely restricted budget constraints across multiple rounds. Label correction is performed over five rounds with a budget of 10k and 1k per round on Cityscapes and PASCAL, respectively. To ensure statistical reliability, each reported result is averaged over three independent runs and presented with its corresponding standard deviation. Our method out-

Methods			Data mIoU (%)					Model mIoU (%)				
LCM	ABC	Mask	1R	2R	3R	4R	5R	1R	2R	3R	4R	5R
			62.60	65.20	66.03	66.85	66.86	56.86	58.93	58.63	58.59	58.59
✓			65.66	71.53	76.00	78.38	80.59	59.02	61.39	64.19	65.58	66.94
	✓		67.31	70.06	74.49	78.32	81.58	59.09	60.86	62.05	65.27	67.75
✓	✓		70.62	**77.29**	80.39	82.80	84.15	60.32	**65.09**	67.36	68.71	70.11
✓	✓	✓	**71.04**	76.08	**81.13**	**83.53**	**85.57**	**61.06**	64.39	**67.83**	**69.63**	**70.88**

Table 5: Ablation analysis of each component. Columns LCM, ABC, and Mask denote the Label Correction Module, Adaptively Balanced CIL, and Mask-level correction, respectively.



(a) (b) (c) (d)

Figure 6: Qualitative results: (a) Image, (b) ALC based pseudo-label (Kim et al. 2024), (c) A^{2}LC based pseudo-label (Ours), (d) Ground Truth. A^{2}LC provides more fine-grained corrections particularly for tail classes (e.g., *traffic light*, *rider*, *traffic sign*), which were largely ignored in the baseline.

performs the baseline by a considerable margin in every round on both datasets, except for the early rounds on PASCAL. Notably, A^{2}LC achieves high efficiency by outperforming the baseline while using only 20% and 60% of its budget on Cityscapes and PASCAL, respectively. It also demonstrates strong effectiveness, yielding 27.23% and 14.30% performance improvements under the same budget constraints. Figure 6 demonstrates that our method produces pseudo-labels of substantially higher quality compared to the baseline. In Table 2, we compare our proposed method not only with the baseline (Kim et al. 2024) but also with state-of-the-art superpixel-based active learning methods (Cai et al. 2021; Hwang et al. 2023; Kim et al. 2023). The evaluation is conducted on a single round with a fixed annotation budget of 100k on Cityscapes. A^{2}LC achieves the highest mIoU, outperforming the baseline by 1.66 and reaching 95% of the fully supervised performance.

Effectiveness of Label Correction Module

In Table 3, we evaluate the correction accuracy of LCM on Cityscapes and PASCAL. On Cityscapes, LCM improves mask accuracy by 41.86 by correcting 22,105 masks, increasing clean masks from 4,229 to 13,481. On PASCAL, it improves accuracy by 77.41 by correcting 1,629 masks, where clean masks increased from 79 to 1,340. While such manual corrections would require multiple rounds, our A^{2}LC framework, empowered by LCM, achieves them instantly without human intervention.

Effectiveness of Adaptively Balanced Acquisition Function

In Table 4, we compare the pseudo-labels obtained using our ABC function with those produced by other acquisition functions, including active learning baselines. Our method achieves strong performance on both datasets. This improvement stems from the pixel-wise adaptive class weight, which assigns higher acquisition scores to false positive pixels belonging to tail classes. By integrating this weight into the acquisition function, our method prioritizes the correction of masks misclassified as tail class labels. Notably, the performance gain is more pronounced on Cityscapes, likely due to its more severe class imbalance.

Ablation Studies

In Table 5, we perform an ablation study to demonstrate the impact of each component in our method. Label correction is conducted over five rounds with a 10k budget per round on Cityscapes, and all experiments follow the non-redundant correction setting. Each component individually improves performance, while the best results are obtained when all are combined. Notably, the combination of LCM and ABC yields greater gains than either alone, highlighting the complementary and synergistic nature of these two components.

Conclusion

In this study, we introduce A^{2}LC, a semi-automated label correction framework for semantic segmentation built upon cascading stages of manual and automatic correction. The A^{2}LC framework incorporates two key components: an LCM that performs automatic correction by propagating human-provided corrections beyond the queried samples, and an ABC acquisition function that strengthens both correction stages through pixel-wise adaptive class weighting. Extensive experiments on Cityscapes and PASCAL VOC 2012 show that our method achieves state-of-the-art performance with superior cost efficiency, highlighting its potential for real-world deployment.

Acknowledgments

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2022-0-01025, Development of core technology for mobile manipulator for 5G edge-based transportation and manipulation)

References

Ash, J. T.; Zhang, C.; Krishnamurthy, A.; Langford, J.; and Agarwal, A. 2020. Deep Batch Active Learning by Diverse, Uncertain Gradient Lower Bounds. In *International Conference on Learning Representations*.

Beck, N.; Killamsetty, K.; Kothawade, S.; and Iyer, R. 2024. Beyond active learning: Leveraging the full potential of human interaction via auto-labeling, human correction, and human verification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2881–2889.

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. 2022. Active label cleaning for improved dataset quality under resource constraints. *Nature communications*, 13(1): 1161.

Cai, L.; Xu, X.; Liew, J. H.; and Foo, C. S. 2021. 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*, 10988–10997.

Chen, J.; Ma, B.; Cui, H.; and Xia, Y. 2024. Think twice before selection: Federated evidential active learning for medical image analysis with domain shifts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11439–11449.

Citovsky, G.; DeSalvo, G.; Gentile, C.; Karydas, L.; Rajagopalan, A.; Rostamizadeh, A.; and Kumar, S. 2021. Batch Active Learning at Scale. In *Neural Information Processing Systems*.

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3213–3223.

Didari, S.; Hu, W.; Woo, J. O.; Hao, H.; Moon, H.; and Min, S. 2024. Bayesian Active Learning for Semantic Segmentation. *arXiv preprint arXiv:2408.01694*.

Edlund, C.; Jackson, T. R.; Khalid, N.; Bevan, N.; Dale, T.; Dengel, A.; Ahmed, S.; Trygg, J.; and Sjögren, R. 2021. LIVECell—A large-scale dataset for label-free live cell segmentation. *Nature methods*, 18(9): 1038–1045.

Ekambaram, R.; Fefilatyev, S.; Shreve, M.; Kramer, K.; Hall, L. O.; Goldgof, D. B.; and Kasturi, R. 2016. Active cleaning of label noise. *Pattern Recognition*, 51: 463–480.

Everingham, M.; Eslami, S. A.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2015. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111: 98–136.

Gal, Y.; Islam, R.; and Ghahramani, Z. 2017. Deep bayesian active learning with image data. In *International conference on machine learning*, 1183–1192. PMLR.

Ge, J.; Zhang, Z.; Phan, M. H.; Zhang, B.; Liu, A.; and Zhao, Y. 2024. Esa: Annotation-efficient active learning for semantic segmentation. *arXiv preprint arXiv:2408.13491*.

Hou, C.; Jiang, K.; Li, T.; Zhou, M.; and Jiang, J. 2025. Coactive: an efficient selective relabeling model for resource constrained edge AI. *Wireless Networks*, 1–14.

Hwang, S.; Lee, S.; Kim, H.; Oh, M.; Ok, J.; and Kwak, S. 2023. Active learning for semantic segmentation with multiclass label query. *Advances in Neural Information Processing Systems*, 36: 27020–27039.

Kampffmeyer, M.; Salberg, A.-B.; and Jenssen, R. 2016. Semantic Segmentation of Small Objects and Modeling of Uncertainty in Urban Remote Sensing Images Using Deep Convolutional Neural Networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

Khanal, B.; Dai, T.; Bhattarai, B.; and Linte, C. 2024. Active Label Refinement for Robust Training of Imbalanced Medical Image Classification Tasks in the Presence of High Label Noise. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 37–47. Springer.

Kim, H.; Hwang, S.; Kwak, S.; and Ok, J. 2024. Active Label Correction for Semantic Segmentation with Foundation Models. In *Forty-first International Conference on Machine Learning*.

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

Kim, K. I. 2022. Active Label Correction Using Robust Parameter Update and Entropy Propagation. In *European Conference on Computer Vision*.

Kremer, J.; Sha, F.; and Igel, C. 2018. Robust active label correction. In *International conference on artificial intelligence and statistics*, 308–316. PMLR.

Lad, V.; and Mueller, J. 2023. Estimating label quality and errors in semantic segmentation data via any model. *arXiv preprint arXiv:2307.05080*.

Li, J.; Zhu, G.; Hua, C.; Feng, M.; Bennamoun, B.; Li, P.; Lu, X.; Song, J.; Shen, P.; Xu, X.; et al. 2023a. A systematic collection of medical image datasets for deep learning. *ACM Computing Surveys*, 56(5): 1–51.

Li, P.; Purkait, P.; Ajanthan, T.; Abdolshah, M.; Garg, R.; Husain, H.; Xu, C.; Gould, S.; Ouyang, W.; and Van Den Hengel, A. 2023b. Semi-supervised semantic segmentation under label noise via diverse learning groups. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1229–1238.

Li, S.-Y.; Shi, Y.; Huang, S.-J.; and Chen, S. 2022. Improving deep label noise learning with dual active label correction. *Machine Learning*, 1–22.

Lüddecke, T.; and Ecker, A. 2022. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7086–7096.

Ma, W.; Karakus, O.; and Rosin, P. L. 2025. Integrating Semi-Supervised and Active Learning for Semantic Segmentation. *arXiv preprint arXiv:2501.19227*.

Plank, B. 2022. The'Problem'of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. *arXiv preprint arXiv:2211.02570*.

Rana, A. J.; and Rawat, Y. S. 2023. Hybrid active learning via deep clustering for video action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18867–18877.

Rebbapragada, U.; Brodley, C. E.; Sulla-Menashe, D.; and Friedl, M. A. 2012. Active Label Correction. In *2012 IEEE 12th International Conference on Data Mining*, 1080–1085.

Ren, T.; Liu, S.; Zeng, A.; Lin, J.; Li, K.; Cao, H.; Chen, J.; Huang, X.; Chen, Y.; Yan, F.; et al. 2024. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*.

Ribeiro Marnet, L.; Brodskiy, Y.; Grasshof, S.; and Wasowski, A. 2024. Uncertainty driven active learning for image segmentation in underwater inspection. In *International Conference on Robotics, Computer Vision and Intelligent Systems*, 66–81. Springer.

Roth, D.; and Small, K. 2006. Margin-based active learning for structured output spaces. In *Machine Learning: ECML 2006: 17th European Conference on Machine Learning Berlin, Germany, September 18-22, 2006 Proceedings 17*, 413–424. Springer.

Rottmann, M.; and Reese, M. 2023. 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*, 3214–3223.

Rückin, J.; Magistri, F.; Stachniss, C.; and Popović, M. 2024a. Active Learning of Robot Vision Using Adaptive Path Planning. *arXiv preprint arXiv:2410.10684*.

Rückin, J.; Magistri, F.; Stachniss, C.; and Popović, M. 2024b. Semi-supervised active learning for semantic segmentation in unknown environments using informative path planning. *IEEE Robotics and Automation Letters*, 9(3): 2662–2669.

Safaei, B.; Vibashan, V.; de Melo, C. M.; and Patel, V. M. 2024. Entropic open-set active learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 4686–4694.

Schachtsiek, M.; Rossi, S.; and Hannagan, T. 2023. Class Balanced Dynamic Acquisition for Domain Adaptive Semantic Segmentation using Active Learning. *arXiv preprint arXiv:2311.14146*.

Sener, O.; and Savarese, S. 2018. Active Learning for Convolutional Neural Networks: A Core-Set Approach. In *International Conference on Learning Representations*.

Taneja, K.; and Goel, A. 2024. Can Active Label Correction Improve LLM-based Modular AI Systems? *arXiv preprint arXiv:2401.05467*.

van Marrewijk, B. M.; Dandjinou, C.; Rustia, D. J. A.; Gonzalez, N. F.; Diallo, B.; Dias, J.; Melki, P.; and Blok, P. M. 2024. Active learning for efficient annotation in precision agriculture: a use-case on crop-weed semantic segmentation. *arXiv preprint arXiv:2404.02580*.

Wang, D.; and Shang, Y. 2014. A new active labeling method for deep learning. In *2014 International joint conference on neural networks (IJCNN)*, 112–119. IEEE.

Wang, Y.; Zhao, J.; Hong, J.; Askin, R. G.; and Maciejewski, R. 2024. A simulation-based approach for quantifying the impact of interactive label correction for machine learning. *IEEE Transactions on Visualization and Computer Graphics*.

Wu, F.; Marquez-Neila, P.; Rafi-Tarii, H.; and Sznitman, R. 2025. Active Learning with Context Sampling and One-vs-Rest Entropy for Semantic Segmentation. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*.

Xu, N.; Liu, B.; Lv, J.; Qiao, C.; and Geng, X. 2023. Progressive purification for instance-dependent partial label learning. In *International Conference on Machine Learning*, 38551–38565. PMLR.

Xu, N.; Qiao, C.; Geng, X.; and Zhang, M.-L. 2021. Instance-dependent partial label learning. *Advances in Neural Information Processing Systems*, 34: 27119–27130.

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

Zou, X.; Yang, J.; Zhang, H.; Li, F.; Li, L.; Wang, J.; Wang, L.; Gao, J.; and Lee, Y. J. 2023. Segment everything everywhere all at once. *Advances in neural information processing systems*, 36: 19769–19782.