A Training-free Synthetic Data Selection Method for Semantic Segmentation

Hao Tang[1], **Siyue Yu**[2], **Jian Pang**[1], **Bingfeng Zhang**[1*]

[1]China University of Petroleum (East China)
[2]XJTLU
{haotang, jianpang}@s.upc.edu.cn, siyue.yu@xjtlu.edu.cn, Bingfeng.Zhang@upc.edu.cn

Abstract

Training semantic segmenter with synthetic data has been attracting great attention due to its easy accessibility and huge quantities. Most previous methods focused on producing large-scale synthetic image-annotation samples and then training the segmenter with all of them. However, such a solution remains a main challenge in that the poor-quality samples are unavoidable, and using them to train the model will damage the training process. In this paper, we propose a training-free Synthetic Data Selection (SDS) strategy with CLIP to select high-quality samples for building a reliable synthetic dataset. Specifically, given massive synthetic image-annotation pairs, we first design a Perturbation-based CLIP Similarity (PCS) to measure the reliability of synthetic image, thus removing samples with low-quality images. Then we propose a class-balance Annotation Similarity Filter (ASF) by comparing the synthetic annotation with the response of CLIP to remove the samples related to low-quality annotations. The experimental results show that using our method significantly reduces the data size by half, while the trained segmenter achieves higher performance.

Code — https://github.com/tanghao2000/SDS

Introduction

Semantic segmentation is a fundamental task in computer vision (Chen et al. 2017a; He et al. 2017; Liu et al. 2018; Zhao et al. 2017). Its goal is to assign semantic labels to each pixel in an image, which is crucial for applications such as autonomous driving (Liu et al. 2020), semantic editing (Ling et al. 2021), and medical image segmentation (Ronneberger, Fischer, and Brox 2015).

With the increasing demand for large-scale datasets in semantic segmentation tasks, the use of synthetic data has attracted widespread attention from researchers. Previous researchers utilize Generative Adversarial Networks (GANs) (Creswell et al. 2018) and their variants, like DatasetGAN (Zhang et al. 2021) and BigDatasetGAN (Li et al. 2022) to effectively generate synthetic dataset, thereby reducing manual annotation. In recent years, Denoising Diffusion Probabilistic Models (DDPMs) (Ho, Jain, and Abbeel 2020; Rombach et al. 2022) have achieved astonishing text-to-image synthesis ability. Such models are also utilized in sample generation for semantic segmentation and can be divided into two pipelines: mask-to-image (Xue et al. 2023; Yang et al. 2024) and image-to-mask (Wu et al. 2023; Nguyen et al. 2024). The mask-to-image pipeline can ensure the accuracy of annotations, but time-consuming with manual annotations. Instead, the image-to-mask method does not need human annotation at all, providing a more efficient pipeline. For example, DiffuMask (Wu et al. 2023) and Dataset-Diffusion (Nguyen et al. 2024), which effectively generate abundant pairs of synthetic images and annotations. Specifically, DiffuMask (Wu et al. 2023) leverages the cross-attention mapping text to image and trains the Affinity Net (Ahn and Kwak 2018) to extend text-driven image synthesis to semantic mask generation. Furthermore, Dataset-Diffusion (Nguyen et al. 2024) makes innovations based on DiffuMask. It rewrites prompts by a large language mode (LLM), generating realistic images and simultaneously producing corresponding segmentation masks. Note that after generating massive samples, these methods use **all the samples** to train the segmenter.

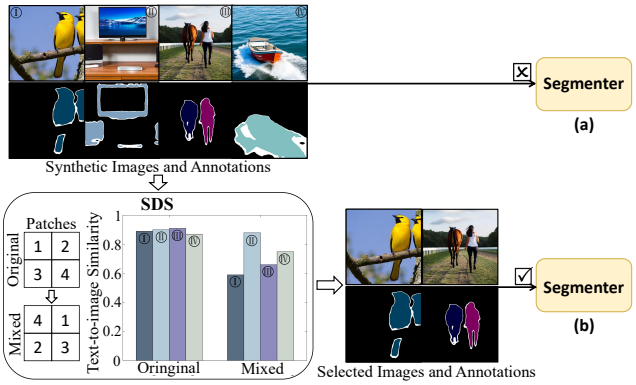


Figure 1: Comparison of synthetic data training methods. (a) Previous methods use all the synthetic data to train a segmenter. (b) Our training-free synthetic data selection method (SDS) selects higher-quality samples to train the segmented, where we select images with significant text-to-image similarity differences before and after mixed image patches.

*Corresponding author.

However, it is hard to control the generation process of the synthetic samples (synthetic image-annotation pairs), making it inevitable to generate samples whose distribution or domain is different from real samples, *i.e.*, low-quality samples. In this case, training the segmenter with them makes it easy to learn unreliable information, impeding the segmentation performance, as shown in Fig. 1(a). Therefore, if we can recognize and select high-quality samples, *i.e.*, find images fitting the distribution of the real-world image, and with accurate synthetic annotation, the effectiveness of the whole training process would be better guaranteed.

Based on the above analysis, we propose a training-free Synthetic Data Selection (SDS) strategy with the Contrastive Language Image Pretraining (CLIP) model (Radford et al. 2021) to select high-quality samples, as shown in Fig. 1 (b). Our intuition is that CLIP is trained on a large amount of real data. In theory, it fits the distribution of real data, making it possible to distinguish whether the synthetic image (Wang, Chan, and Loy 2023) belong to the same distribution of real images. To evaluate it, We randomly select several synthetic images using the Dataset-Diffusion (Nguyen et al. 2024) and calculate the text-to-image similarity. As shown on the left bar chart in Fig. 1 (b), we find that all of them generate high text-to-image similarities, which is hard to regard as a confidence metric directly. Then we mix up the order of the image patches (Lee et al. 2024) and re-calculate the similarities. The results show a significant difference among the images, as shown on the right bar chart in Fig. 1(b). Considering the CLIP model is pre-trained with the natural object order of images, using the image with mixed patches as input, the text-to-image similarity should be low since mixing operation damages object structures, the model cannot receive common object relationships or the proper object order in the image. On the contrary, if the text-to-image similarity of patch-mixed images remains high, it can be treated as that the CLIP model relies on uncommon or even incorrect object relationships, *i.e.*, unrepresentative information in the image makes the network produce high-confidence decisions. Such samples are unreliable in training a model. Therefore, we claim that a high-quality image should have low text-to-image similarity after mixing patches, while a low-quality image should have high similarity after mixing patches.

Following our observations, we design a Perturbation-based CLIP Similarity (PCS) approach to select reliable synthetic data for training semantic segmenter. Specifically, we first calculate text-to-image similarities for the original image and the patch-mixed image, respectively. For a high-quality image, its text-to-image similarity with the original patches should be high to guarantee it has the correct classes and objects. Meanwhile, its text-to-image similarity with mixed patches should be low to ensure it shares a similar distribution with real images. To accurately quantify the similarity degree, we use the similarity difference between the original patches and mixed patches to replace only considering the similarity of the mixed patches. The similarity difference is defined as the PCS score in this paper. Only samples that satisfy the above two rules will remain for further processing.

Besides, it is necessary for a high-quality sample to require both high-quality images and annotations. The above module only selects high-quality images, ignoring the annotation quality, as shown in Fig. 1, the PCS score of the boat is high, but it mistakenly labeled the wave as the boat. To solve this problem, we propose a class-balance Annotation Similarity Filter (ASF) to remove low-quality annotation samples by comparing the synthetic annotation with the response of CLIP. While selecting the synthetic images, we utilize the generation ability of CLIP to generate a set of reference annotations. Considering that the quality of annotations varies among different classes, we classify them into different groups and finally select high-quality annotation samples by computing the mIoU between reference annotations and synthetic annotations.

To evaluate the effectiveness of our selection strategy, we use our approach to select samples from synthetic datasets to train the segmenter and then evaluate the performance on two real-world datasets, PASCAL VOC 2012 (Everingham et al. 2015) and MS COCO 2017 (Lin et al. 2014). Experimental results show our method achieves higher performance while significantly reducing the dataset scales.

In summary, the contributions of our work are as follows:

- We observe that CLIP has different performances on different samples in synthetic data. Based on this, we propose a training-free Synthetic Data Selection (SDS) pipeline that can effectively select synthetic samples.
- We design a Perturbation-based CLIP Similarity (PCS) method based on CLIP to select high-quality synthetic images. In addition, we also propose a class-balance Annotation Similarity Filter (ASF) module that compares synthetic annotations with the response of CLIP to remove samples associated with low-quality annotations.
- The experiment shows that with our data selection pipeline, the number of datasets is significantly reduced, and better performance is achieved in training segmenter, *e.g.*, the synthetic training dataset can be reduced by half but generate a 3% mIoU increase to 62.5% when evaluating on the Pascal VOC 2012 dataset.

Related Work

Fully Supervised Semantic Segmentation

Semantic segmentation is a pixel-level image analysis technique aimed at identifying and differentiating objects of all classes in an image, precisely locating them within the image. Currently, mainstream semantic segmentation methods primarily fall into two categories: fully convolutional neural networks (FCN) (Long, Shelhamer, and Darrell 2015) and Transformer-based approaches. Among them, FCN includes models such as U-Net (Ronneberger, Fischer, and Brox 2015), SegNet (Badrinarayanan, Kendall, and Cipolla 2017) and the DeepLab series (Liang-Chieh et al. 2015; Chen et al. 2017a,b, 2018), and Mask2Former (Cheng et al. 2022) for Transformer-based methods. All these methods are trained with images collected from real-world with manual pixel-level annotations, while in this work, we focus on synthetic image-annotation pairs.

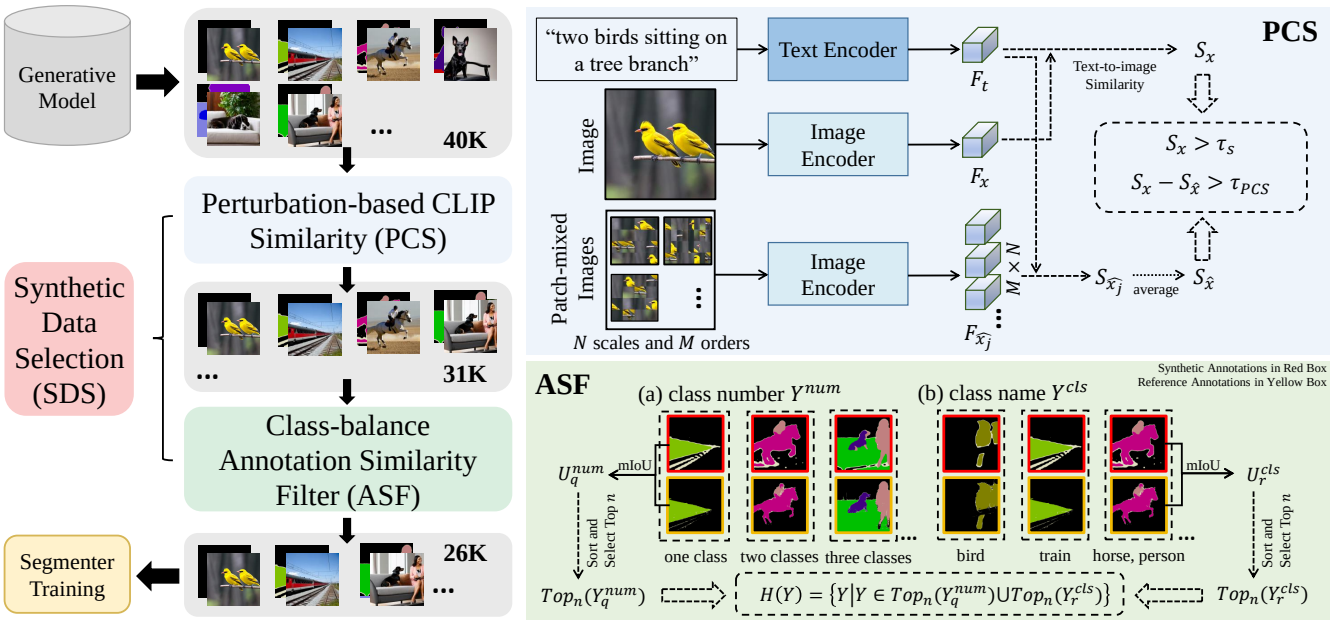


Figure 2: Overview of our Synthetic Data Selection (SDS) framework. (1) The initial synthetic dataset consists of 40k image-annotation pairs generated by a generative model. (2) Our SDS includes two modules: PCS and ASF. In the PCS module, the text caption and original image are encoded to feature represents F_t and F_x. For patch-mixed image, the multi-level patch-mixed strategy are designed to obtain $F_{\hat{x_j}}$. We calculate the text-to-image similarity and select image samples with high similarity and high PCS scores. In the ASF module, we design the rule (a) and (b) to calculate the mIoU between synthetic annotations and the response of CLIP. The mIoU scores are sorted and we select the Top n annotation samples. (3) The selected dataset remaining 26k images-annotation pairs are used to train a segmenter.

Synthetic Data for Semantic Segmentation

Semantic segmentation of synthetic data is predominantly executed through image-to-mask and mask-to-image methods. For image-to-mask, DiffuMask (Wu et al. 2023) exploits cross-attention maps between text and images and trains an Affinity Net to generate pixel-level annotations. Dataset-Diffusion (Nguyen et al. 2024) removes the Affinity Net and rewrites prompts by LLM, which only utilizes the diffusion model to generate accurate segmentation masks for synthetic images. For mask-to-image, FreestyleNet (Xue et al. 2023) introduces Rectified Cross-Attention (RCA), seamlessly integrating semantic masks into the image generation process. FreeMask (Yang et al. 2024) makes improvements based on FreestyleNet. It significantly improves the performance of semantic segmenter by filtering noise and prioritizing the sampling of hard-to-learn masks.

Method

Problem Setting

We first utilize LLM to generate captions for real images. Then, images and their captions are input to the generative model (Nguyen et al. 2024) to synthetic the dataset $D = (X_m, Y_m)_{m=1}^{M}$, where X_m is the image and Y_m is the corresponding annotation. These images and annotations capture both the semantic and location information of the target classes $C = \{c_1, c_2, ..., c_L\}$, where L represents the number of classes. Our objective is to select high-quality samples from D to perform a reliable synthetic dataset $D_f = (X_m, Y_m)_{m=1}^{M_f}$, where M_f represents the number of selected samples. Finally, we can train a segmenter with D_f.

Overview

The overall framework of our approach is shown in Fig. 2, which can be divided into the following steps:

1) We generate massive synthetic images with corresponding annotations following Dataset-Diffusion (Nguyen et al. 2024). The original synthetic images are input to our PCS module to evaluate their quality by comparing the similarity and our designed PCS scores. We select high-fidelity images that exhibit both high similarity and PCS scores.
2) Meanwhile, the reference annotation for each image is generated using softmax Class Activation Maps (Lin et al. 2023). Both the reference annotations and the synthetic annotations are input into the ASF module to assess their quality. We select reliable annotations based on our designed mIoU-based rule.
3) We combine the selected images from the PCS module with annotations from the ASF module to create a reliable dataset for training a segmenter.

Synthetic Dataset Generation

Recent studies indicate that high-quality and diverse results can be synthesized by training large-scale text-to-image dif-

fusion models. Motivated by this, we adopt the most recent work Dataset-Diffusion (Nguyen et al. 2024), to synthesize additional training image-annotation pairs based on the provided real datasets. Specifically, for each real image, we generate K captions through the LLM, such as ChatGPT (OpenAI 2023), then captions are input to the diffusion model to obtain K synthetic images. Meanwhile, the corresponding annotations are generated using the attention maps from the diffusion model. These synthetic samples (image-annotation pairs) build an initial dataset.

Perturbation-based CLIP Similarity (PCS)

A straightforward way to evaluate image quality is to use the CLIP model to calculate the cosine similarity between the text prompt features and the corresponding image features directly. Specifically, the text prompt T and the image X are input into the text encoder f_θ and image encoder g_θ, respectively, obtaining two feature representations:

$$F_t = f_\theta(T), F_x = g_\theta(X), \tag{1}$$

where $F_t \in \mathbb{R}^{1\times d}$ and $F_x \in \mathbb{R}^{1\times d}$ are the feature representations of the text prompt T and the image X, respectively, where d is the channel number.

The common text-to-image similarity $S_x \in [0, 1]$ is calculated as:

$$S_x = \frac{F_t F_x^\top}{\parallel F_t \parallel \cdot \parallel F_x \parallel}, \tag{2}$$

where $\top$ represents the matrix transpose, and $\parallel \cdot \parallel$ represents the l_2 norm. Our experiments indicate that synthetic data usually have high text-to-image similarity, so only relying on Eq. (2) cannot evaluate the quality of synthetic data.

To address this problem, we propose Perturbation-based CLIP Similarity (PCS). Perturbing in synthetic images can be achieved through image transformations such as pixel-mixed, patch-mixed, or random occlusion. Each strategy possesses distinct characteristics. With pixel-mixed, the mean color of the image is maintained, but it becomes difficult to discern the object feature. Patch-mixed disrupts the shape of the object in the image but preserves local features through the patches. Random occlusion allows for the preservation of partial information. But, when the object is not exceptionally large, it may fully occlude the object (Lee et al. 2024). In our PCS module, we need both the object features in images before and after perturbation. Therefore, we apply the patch-mixed strategy in our method, which divides an image into patches with different scales and then mixes the order of patches as new images to keep semantic integrity.

Specifically, given an original image $X \in R^{H\times W\times 3}$, we design a multi-level patch-mixed strategy, which sets N_s patch scales and produces N_o different patch orders for each scale to mitigate the influence of the random. If we only perform once patch-mixed on the image, different patch sizes and orders will lead to unstable similarity and affect the reproducibility of the experiment. Finally, we can generate $N_s \times N_o$ new mixed images. In detail, to generate the j-th mixed image, suppose the original image is divided into n patches and the produced random patch order is $\{j_1, j_2, ..., j_n\}$, for example, 4 patches with the order $\{2, 1, 4, 3\}$, then the mixed image is defined as follows:

$$X_j = \{patch_{j_1}, patch_{j_2}, \ldots, patch_{j_n}\}, \tag{3}$$

where $patch_{j_*}$ means the j_* patch in the original image. X_j is the j-th mixed image, and $j \in \{1, 2, \ldots, N_s \times N_o\}$. For each X_j, we obtain the visual feature $F_{\hat{x}_j} = g_\theta(X_j)$ through Eq. (1). Then, we compute the similarity between the mixed image and the text as follows:

$$S_{\hat{x}_j} = \frac{F_t F_{\hat{x}_j}^\top}{\parallel F_t \parallel \cdot \parallel F_{\hat{x}_j} \parallel}. \tag{4}$$

Finally, we average all similarities computed from $N_s \times N_o$ patch-mixed images as follows:

$$S_{\hat{x}} = \frac{1}{N_s \times N_o} \sum_{j=1}^{N_s \times N_o} S_{\hat{x}_j}, \tag{5}$$

$S_{\hat{x}}$ is the averaged similarity, it integrates representative information from multiple mixed images, making the similarity more reliable. With the averaged similarity $S_{\hat{x}}$ and text-to-image similarity S_x, we design Perturbation-based CLIP Similarity to select high-quality samples:

$$G(x) = \{x \mid S_x > \tau_s, PCS(x, \hat{x}) > \tau_{PCS}\}, \tag{6}$$

where

$$PCS(x, \hat{x}) = S_x - S_{\hat{x}}, \tag{7}$$

In Eq. (6), $G(x)$ is the selected image set that includes images maintaining both high text-to-image similarity and high PCS score, τ_s and τ_{PCS} are thresholds. Eq. (7) computes the PCS score for the images. The selected images have two advantages: a) rich semantics of objects, the condition $S_x > \tau_s$ ensures a strong correlation between the semantics of the object and text prompt; b) Fit the real distribution, $PCS(x, \hat{x}) > \tau_{PCS}$ ensures synthetic images fit in a similar distribution with real images and contain representative information. As shown in Fig. 3, the selected images

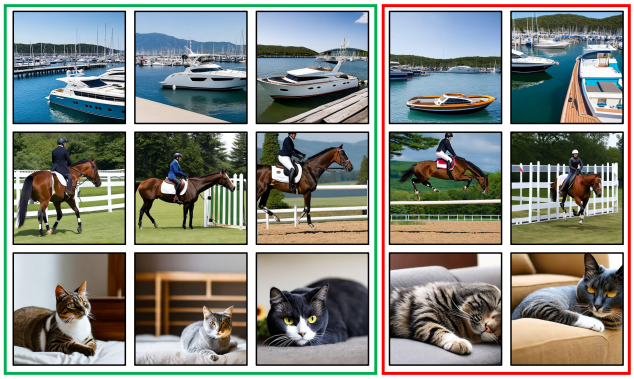


Figure 3: Visualization of selected images. High-quality images are with green box and low-quality with red box.

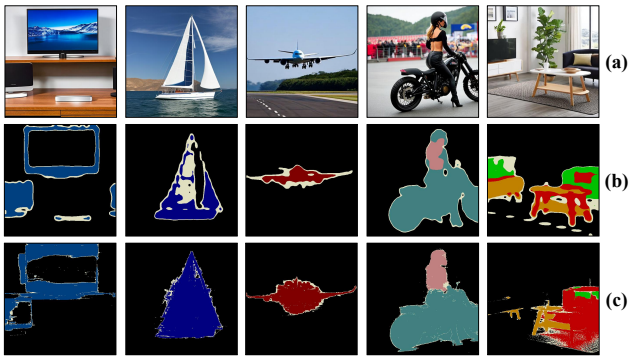


Figure 4: Visualization of low-quality annotation examples. (a) Synthetic Images. (b) Synthetic Annotations. (c) Reference Annotations from CLIP.

are highlighted with the green box. We observe that the images in the green box are visually pleasant, objects in such images contain relatively rich semantics. Yet, the images in the red box are less realistic, and our PCS can help discard these low-quality samples.

Class-balance Annotation Similarity Filter (ASF)

Due to the limitations of the generative models, the synthetic annotations unavoidably contain multiple objects that are relatively inaccurate (Nguyen et al. 2024). As shown in Fig. 4(b), there are some low-quality annotation examples. To remove these samples, we propose a class-balance Annotation Similarity Filter (ASF). The most important metric for evaluating annotations is mIoU which measures the degree of overlap between two segmentation regions, with higher values indicating better segmentation accuracy. Owing to the synthetic images not being manually annotated, we generate a set of reference annotations for the selected images in previous section from CILP, following (Lin et al. 2023).

Considering the impact of the class, our ASF module follows two rules: a) The quality of annotations containing fewer classes is generally higher. To balance the number of classes in an annotation, we should group the dataset according to the class numbers in an annotation; b) The quality of annotations among the classes is unbalanced, *e.g.*, images of the monitor have coarse annotations while images of sheep all have accurate annotations. Thus, we should group the dataset according to the object class in an annotation to prevent the classes with all low-quality annotations from being completely removed in the selection process. We define Y^{num} and Y^{cls} to represent two grouped annotations based on the above two rules, respectively:

$$Y^{num} = \left\{ Y_1^{num}, Y_2^{num}, \ldots, Y_q^{num}, \ldots, Y_Q^{num} \right\}, \quad (8)$$

$$Y^{cls} = \left\{ Y_1^{cls}, Y_2^{cls}, \ldots, Y_r^{cls}, \ldots, Y_R^{cls} \right\}. \quad (9)$$

In Eq. (8), Y_q^{num} is the q-th grouped annotation subset in Y^{num}, where $q \in \{1, 2, .., Q\}$. Q represents the maximum number of simultaneous classes among all the images. For example, $Q = 6$ means up to six classes appear among all images. Y_r^{cls} is the r-th subset in Y^{cls} and $r \in \{1, 2, .., R\}$. R represents the maximum index of classes in the dataset, *e.g.*, $R = 20$ means the 20th class in the dataset. Then for each subset, we calculate the mIoU score as follows:

$$U_q^{num} = \left\{ U_{q_1}^{num}, U_{q_2}^{num}, \ldots, U_{q_{\max}}^{num} \right\}, \quad (10)$$

$$U_r^{cls} = \left\{ U_{r_1}^{cls}, U_{r_2}^{cls}, \ldots, U_{r_{\max}}^{cls} \right\}, \quad (11)$$

where $U_{q_i}^{num}$ is the mIOU score for the i-th image in the Y_q^{num}. Similarly, $U_{r_j}^{cls}$ is the mIoU for the j-th image in the Y_r^{cls}. $q_{\max}$ and $r_{\max}$ are the maximum numbers of the corresponding grouped annotations. The $U_{q_i}^{num}$ and $U_{r_j}^{cls}$ are computed as follows:

$$U_{q_i}^{num} = \text{mIoU}(Y_{q_i}^{num}, \hat{Y}_{q_i}^{num}), q_i \in \{q_1, q_2, ..., q_{\max}\}, \quad (12)$$

$$U_{r_j}^{cls} = \text{mIoU}(Y_{r_j}^{cls}, \hat{Y}_{r_j}^{cls}), r_j \in \{r_1, r_2, ..., r_{\max}\}, \quad (13)$$

where $Y_{q_i}^{num}$ and $Y_{r_j}^{cls}$ are synthetic annotations, $\hat{Y}_{q_i}^{num}$ and $\hat{Y}_{r_j}^{cls}$ are reference annotations generated from CLIP. A higher mIoU score means the synthetic annotation is more reliable. Note that each element index in Y_q^{num} and U_q^{num} is a one-to-one correspondence. Hence, we can use the mIoU set, U_q^{num}, to select the annotations from Y_q^{num}, *i.e.*, annotations with Top n mIoU scores in each grouped annotation remain to build a new subset, and all other annotations are removed, as follows:

$$\begin{aligned} \text{Top}_n(Y_q^{num}) &= \left\{ Y_{q_{k_1}}^{num}, Y_{q_{k_2}}^{num}, \ldots, Y_{q_{k_n}}^{num} \right\}, \\ \text{where} \quad & U_{q_{k_1}}^{num} \geq U_{q_{k_2}}^{num} \geq \cdots \geq U_{q_{k_n}}^{num}. \end{aligned} \quad (14)$$

For annotation set Y_r^{cls}, we obtain $\text{Top}_n(Y_r^{cls})$ following the above process. The reliable annotation set is selected as:

$$H(Y) = \left\{ Y \mid Y \in \text{Top}_n(Y_q^{num}) \cup \text{Top}_n(Y_r^{cls}) \right\} \quad (15)$$

where $H(Y)$ represents the selected annotation set which is the union of $\text{Top}_n(Y_q^{num})$ and $\text{Top}_n(Y_r^{cls})$.

Finally, we combine the two modules to build a reliable synthetic training dataset consisting of high-fidelity images with their corresponding high-quality pixel-level semantic annotations.

Experiments

Dataset and Evaluation Metrics

We evaluate our method on PASCAL VOC 2012 (Everingham et al. 2015) and MS COCO 2017 (Lin et al. 2014).

PASCAL VOC 2012 has 20 object classes and 1 background class, which is augmented by SBD (Hariharan et al. 2011) to obtain 10,584 training, 1,449 validation, and 1,456 test images. For the synthetic dataset, we follow Dataset Diffusion (Nguyen et al. 2024) to produce 40k image-annotation pairs as the initial dataset. After applying SDS on the initial dataset, 26k high-quality samples are selected to form the SDS-VOC dataset.

Method	Segmenter	Backbone	Data Size	mIoU (%)
VOC's training (Everingham et al. 2015)	DeepLabV3 (Chen et al. 2017b)	ResNet50	11.5k	77.4
Dataset Diffusion (Nguyen et al. 2024)			40k	58.1*
SDS-VOC (Ours)			26k	**60.4**
VOC's training (Everingham et al. 2015)		Resnet101	11.5k	79.9
Dataset Diffusion (Nguyen et al. 2024)			40k	56.9*
SDS-VOC (Ours)			26k	**59.1**
VOC's training (Everingham et al. 2015)	Mask2Former (Cheng et al. 2022)	Resnet50	11.5k	77.3
DiffuMask (Wu et al. 2023)			60k	57.4
Dataset Diffusion (Nguyen et al. 2024)			40k	57.8*
SDS-VOC (Ours)			26k	**59.8**
Dataset Diffusion (Nguyen et al. 2024)	CDL (Zhang et al. 2023)	Resnet101	40k	59.6*
SDS-VOC (Ours)			26k	**62.5**

Table 1: Comparisons with other methods on the PASCAL VOC 2012 dataset. * means our reproduced results.

Method	Segmenter	Backbone	Data Size	mIoU (%)
COCO's training (Lin et al. 2014)	DeepLabV3 (Chen et al. 2017b)	ResNet50	118k	48.9
Dataset Diffusion (Nguyen et al. 2024)			80k	28.7*
SDS-COCO (Ours)			50k	**31.0**
COCO's training (Lin et al. 2014)		Resnet101	118k	54.9
Dataset Diffusion (Nguyen et al. 2024)			80k	29.2*
SDS-COCO (Ours)			50k	**31.8**
Dataset Diffusion (Nguyen et al. 2024)	CDL (Zhang et al. 2023)	Resnet101	80k	30.3*
SDS-COCO (Ours)			50k	**33.4**

Table 2: Comparisons with other methods on the MS COCO 2017 dataset. * means our reproduced results.

MS COCO 2017 contains 80 object classes and 1 background class with 118k training and 5k validation images. Similarly, we follow Dataset Diffusion to obtain 80k synthetic image-annotation pairs and then use our method to get 50k high-quality samples to form the SDS-COCO dataset.

The mean Intersection over Union (mIoU) is used as the evaluation metric.

Implementation Details

Model settings: We employ the CLIP pre-trained model ViT-B-16 with our method to select high-quality samples. We involve three segmenters for evaluation: DeepLabv3 (Chen et al. 2017a), Mask2Former (Cheng et al. 2022), and CDL (Zhang et al. 2023). All segmenters follow the default settings in the original paper.

Hyperparatemers: In Perturbation-based CLIP Similarity (PCS), $N_s \in \{8, 16, 32\}$ represents the patch scale. N_o is the number of patch orders, we set $N_o = 3$. In Eq. 6, thresholds τ_s and τ_{PCS} are set to 0.8 and 0.1, respectively. In class-balance Annotation Similarity Filter (ASF), we set n to 0.6 $\times$ the number of samples within groups in Eq. (15), *i.e.*, top 60% samples are selected. All experiments are running on NVIDIA RTX 3090 GPUs.

Performance Comparison

Table 1 presents the evaluation results of three segmenters. The datasets used in Table 1 are divided into three types, (1) real dataset; (2) initial datasets where samples are generated by diffusion models (Dataset Diffusion); (3) SDS dataset, comprising only high-quality samples selected from the initial dataset. From comparison, our method achieves 62.5% mIoU when compared to the Dataset diffusion of 59.6% mIoU. Further, ours outperforms DiffuMask by 2.4% mIoU using the same ResNet50 backbone. The results reveal the diffusion model produces many noise samples which hinder the semantic segmentation, with our method, only high-quality samples remain, and achieve better performance.

Fig. 5 shows our predicted annotation results on the validation set of VOC, which overall align with the ground truth.

MS COCO 2017: Table 2 shows the performance on two segmenter. The dataset settings are the same as above. Our method achieves a promising result of 33.4% mIoU compared to 30.3% mIoU of the Dataset Diffusion.

Ablation Studies

We select DeepLabV3 (Chen et al. 2017b) with ResNet50 as the segmenter. All experiments are conducted on the PASCAL VOC 2012 dataset unless otherwise stated.

Effectiveness of Each Module in SDS: Table 3 demonstrates the effectiveness of each module in SDS by progressively adding PCS and ASF. It can be seen both PCS and ASF bring increased performance. By combining PCS and ASF to select a dataset for training segmenter, the segmenter has the best result, reaching 60.4% mIoU. PCS module selects high-quality images and the ASF module selects high-quality annotations, which are complementary and consistent with our design targets.

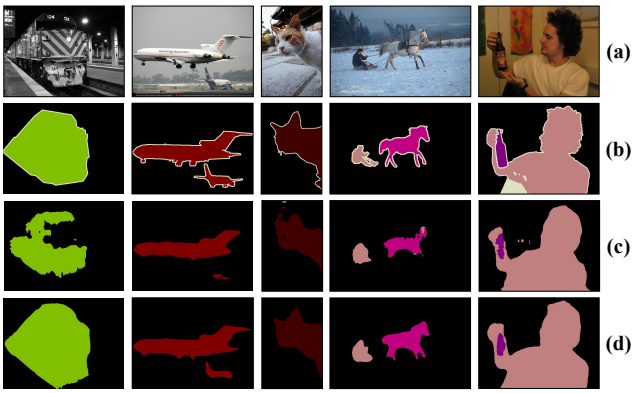


Figure 5: Segmentation results on the PASCAL VOC 2012 validation set. (a) Images. (b) Ground-truth annotation. (c) Predicted annotation from Dataset Diffusion. (d) Our predicted annotation.

base	PCS	ASF	mIoU (%)
✓			58.1
✓	✓		59.5
✓	✓	✓	**60.4**

Table 3: Effectiveness of PCS and ASF modules. "base" means the synthetic dataset of Dataset Diffusion.

N_s	N_o	mIoU (%)
{8}	3	59.9
{16}	3	60.2
{32}	3	59.5
{8, 16, 32}	3	**60.4**

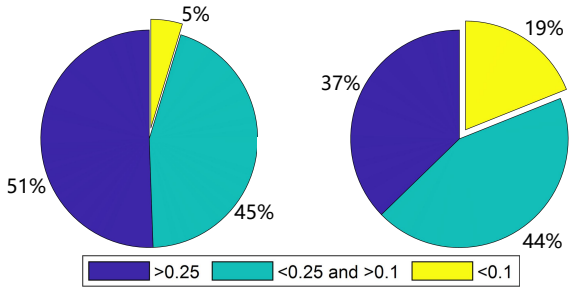
Table 4: Ablation studies of multi-level patch-mixed strategy.

Moreover, we make a comparison of our proposed PCS score between the real dataset and the initial synthetic dataset. As shown in Fig. 6, the proportion of low PCS scores in the synthetic dataset is much higher than that in the real dataset, which indicates that using PCS is an effective method to evaluate the image quality.

Multi-level Patch-mixed Strategy: Table 4 illustrates the influence of N_s and N_o in Multi-level Patch-mixed Strategy. N_s represents the scales of patches in the image and N_o represents the number of orders for each scale. It can be observed that the optimal result is achieved when we take average of N_s patch scales and set the $N_o = 3$.

Effectiveness of Rules in ASF Module: Table 5 demonstrates the effectiveness of each rule in the ASF module. We observe that both Rule (a) and Rule (b) bring increased performance, which indicates the necessity of our class-balance strategy.

Data Size and the Performance of Segmenter: Table 6 illustrates the relationship between data size and the segmenter performance. Our method can reduce the synthetic training dataset by half but generate a 2.3% mIoU increase.



PCS in Real Dataset PCS in Synthetic Dataset

Figure 6: A significant difference in PCS scores between the real dataset and the synthetic dataset.

Direct	Rule (a)	Rule (b)	mIoU (%)
✓			57.5
	✓		59.2
		✓	59.6
	✓	✓	**60.4**

Table 5: Effectiveness of rules in ASF module. "Direct" means calculating the mIoU in all synthetic annotations and taking annotations with Top n mIoU scores.

Data Size	40k	30k	26k	15k
mIoU (%)	58.1	59.7	**60.4**	58.0

Table 6: The relationship between data size and the performance of segmenter.

Conclusion

In this work, we propose a training-free Synthetic Data Selection (SDS) method with CLIP to select high-quality samples from synthetic dataset. To achieve this, we design two novel modules: the PCS module, which introduces Perturbation in images and selects high-quality images without incorrect object relationships, and the ASF module, which applies a class-balance strategy and selects high-quality annotations based on mIoU scores. With our two processing strategies, the segmenter trained on the selected dataset achieves a better performance.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No.62301613, No.62301451), the Taishan Scholar Program of Shandong (No. tsqn202306130), the Shandong Natural Science Foundation (Grant No. ZR2023QF046), Independent Innovation Research Project of China University of Petroleum (East China) (No.22CX06060A).

References

Ahn, J.; and Kwak, S. 2018. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4981–4990.

Badrinarayanan, V.; Kendall, A.; and Cipolla, R. 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12): 2481–2495.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2017a. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4): 834–848.

Chen, L.-C.; Papandreou, G.; Schroff, F.; and Adam, H. 2017b. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*.

Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision*, 801–818.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1290–1299.

Creswell, A.; White, T.; Dumoulin, V.; Arulkumaran, K.; Sengupta, B.; and Bharath, A. A. 2018. Generative adversarial networks: An overview. *IEEE Signal Processing Magazine*, 35(1): 53–65.

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.

Hariharan, B.; Arbeláez, P.; Bourdev, L.; Maji, S.; and Malik, J. 2011. Semantic contours from inverse detectors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 991–998. IEEE.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2961–2969.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.

Lee, J.; Jung, D.; Lee, S.; Park, J.; Shin, J.; Hwang, U.; and Yoon, S. 2024. Entropy is not Enough for Test-time Adaptation: From the Perspective of Disentangled Factors. In *Proceedings of the International Conference on Learning Representations*.

Li, D.; Ling, H.; Kim, S. W.; Kreis, K.; Fidler, S.; and Torralba, A. 2022. Bigdatasetgan: Synthesizing imagenet with pixel-wise annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21330–21340.

Liang-Chieh, C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. 2015. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *Proceedings of the IEEE/CVF International Conference on Learning Representations*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, 740–755. Springer.

Lin, Y.; Chen, M.; Wang, W.; Wu, B.; Li, K.; Lin, B.; Liu, H.; and He, X. 2023. Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15305–15314.

Ling, H.; Kreis, K.; Li, D.; Kim, S. W.; Torralba, A.; and Fidler, S. 2021. Editgan: High-precision semantic image editing. *Advances in Neural Information Processing Systems*, 34: 16331–16345.

Liu, S.; Qi, L.; Qin, H.; Shi, J.; and Jia, J. 2018. Path aggregation network for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8759–8768.

Liu, X.; Han, Y.; Bai, S.; Ge, Y.; Wang, T.; Han, X.; Li, S.; You, J.; and Lu, J. 2020. Importance-aware semantic segmentation in self-driving with discrete wasserstein training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 11629–11636.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3431–3440.

Nguyen, Q.; Vu, T.; Tran, A.; and Nguyen, K. 2024. Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation. *Advances in Neural Information Processing Systems*, 36.

OpenAI, R. 2023. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2(5).

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, 8748–8763. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 234–241. Springer.

Wang, J.; Chan, K. C.; and Loy, C. C. 2023. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2, 2555–2563.

Wu, W.; Zhao, Y.; Shou, M. Z.; Zhou, H.; and Shen, C. 2023. Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1206–1217.

Xue, H.; Huang, Z.; Sun, Q.; Song, L.; and Zhang, W. 2023. Freestyle layout-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14256–14266.

Yang, L.; Xu, X.; Kang, B.; Shi, Y.; and Zhao, H. 2024. Freemask: Synthetic images with dense annotations make stronger segmentation models. *Advances in Neural Information Processing Systems*, 36.

Zhang, B.; Xiao, J.; Wei, Y.; and Zhao, Y. 2023. Credible dual-expert learning for weakly supervised semantic segmentation. *International Journal of Computer Vision*, 131(8): 1892–1908.

Zhang, Y.; Ling, H.; Gao, J.; Yin, K.; Lafleche, J.-F.; Barriuso, A.; Torralba, A.; and Fidler, S. 2021. Datasetgan: Efficient labeled data factory with minimal human effort. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10145–10155.

Zhao, H.; Shi, J.; Qi, X.; Wang, X.; and Jia, J. 2017. Pyramid scene parsing network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2881–2890.