Iterative Self-Training with Class-Aware Text-to-Image Synthesis for Visual Task Learning

Xiang Zhang, Wanqing Zhao*, Pengyang Li, Ying Liu*, Hangzai Luo, Sheng Zhong, Jinye Peng, Jianping Fan

Northwest University, Xi'an, China

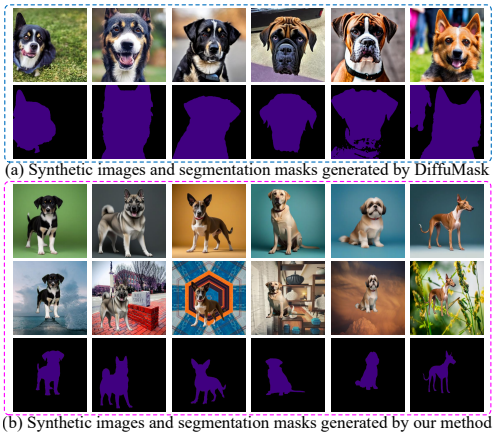
{xiangz@, zhaowq@, lipengyang@stumail., liuying6@stumail., hzluo@, szhong@, pjy@, jfan@}nwu.edu.cn

Abstract

Generative models are widely used to produce synthetic images with annotations, alleviating the burden of image collection and annotation for training deep visual models. However, challenges such as limited image diversity, noisy pseudo labels, and domain gaps between synthetic and real images often undermine their effectiveness in downstream visual tasks. This paper introduces the Iterative Self-Training with Class-Aware Text-to-Image Synthesis (IST-CATS) framework, which addresses these challenges by integrating a class-aware text-to-image synthesis (CATS) component with an iterative self-training (IST) strategy. CATS innovatively introduces a class-aware chain approach to generate detailed descriptions. These descriptions act as prompts for a diffusion model, enabling the creation of a diverse of images accompanied by distinguishable objects against the background. The generated images can be easily pseudo-labeled by an unsupervised instance segmentation method, and then noisy pseudo labels can be effectively purified by a novel feature similarity-based filtering mechanism. The generated images underpin our IST, which progressively enhances vision models and refines pseudo labels through self-training and our proposed label filtering strategy (LabFilt). LabFilt meticulously improves the quality of pseudo labels by employing class-adaptive techniques at both the pixel and object levels, ensuring refined pseudo-label accuracy. IST-CATS demonstrates superior performance in object detection and semantic segmentation compared to traditional synthetic and semi/weakly-supervised methods, effectively addressing data collection and annotation challenges.

Introduction

Existing data-hungry deep vision models typically demand extensive images with precise annotations to achieve notable advancements. However, manually annotating large-scale training images at the pixel or bounding box level is extremely labor-intensive and poses significant challenges for humans to provide consistent labeling quality. In recent years, advancements have been made in alleviating the manual burden of image annotation through the development of semi-supervised (Shehzadi et al. 2024; Sun et al. 2024) and weakly-supervised (Feng et al. 2024; Yoon et al. 2024) learning techniques for visual tasks. However, semi-supervised techniques rely on annotated images and require similar domain distributions for labeled and unlabeled datasets. Meanwhile, weakly-supervised approaches may overlook detailed spatial information inherent in annotations, potentially compromising model precision. Some researchers (Wu et al. 2019; Zheng and Yang 2021) have also explored using synthetic images generated by 3D models or game engines to train vision models. Nevertheless, the creation of virtual scenes and 3D models still demands significant involvement from skilled professionals.



(a) Synthetic images and segmentation masks generated by DiffuMask

(b) Synthetic images and segmentation masks generated by our method

Figure 1: Comparing synthetic images generated by our method with those from DiffuMask (Wu et al. 2023b).

Artificial Intelligence Generated Content (AIGC) presents an appealing solution to diminish the dependence on manually annotated data for visual model training. Certain studies (Bosquet et al. 2023; Yang et al. 2024b) have leveraged Generative Adversarial Networks (GANs) to produce training data for object detection or semantic segmentation, thus minimizing the need for manual annotation. However, GAN-based approaches often struggle to generate high-quality realistic images, which can limit the performances of the trained models. Recently, the advent of text-to-image diffusion models (Rombach et al. 2022; Yang et al. 2024a) offers a promising alternative, capable of generating highly realistic images from detailed textual descriptions. Such models, exemplified by initiatives

*Wanqing Zhao and Ying Liu are the corresponding authors.

like DiffuMask (Wu et al. 2023b), DiffusionSeg (Ma et al. 2023), and Dataset diffusion (Nguyen et al. 2024) have demonstrated their potential in generating synthetic datasets complete with precise annotations (pixel-level or bounding-box annotations) by utilizing concise text prompts. Despite reducing image generation and annotation costs, these approaches face several key challenges: 1) Image Diversity: Simplistic text prompts limit the variety of generated images, leading to visually similar outputs within the same category, as shown in Figure 1(a); 2) Label Quality: Coarse labels (e.g., incomplete mask and excessive mask in Figure 1(a)) produced by generative models can adversely affect model performance and generalization; 3) Domain Gap: Data generated by generative models typically rely on features from the training set, while real images exhibit more variability. This distribution shift causes discrepancies between generated and real images, affecting the performance of models.

In addressing these challenges that beset visual task learning, particularly in scenarios where labeled images are scarce or costly to obtain, this paper introduces a novel approach termed Iterative Self-Training with Class-Aware Text-to-Image Synthesis (IST-CATS). IST-CATS employs class-aware text-to-image synthesis (CATS) to generate diverse synthetic images with annotations and incorporates iterative self-training (IST) to improve the precision of the generated annotations and the performance of the visual model iteratively. At the core of CATS is a class-aware chain method that produces detailed image descriptions, ensuring that the generated images are not only diverse but also conducive to the segmentation of objects from their backgrounds. To attain pseudo labels, an unsupervised instance segmentation model is employed alongside a noise filtering mechanism predicated on feature similarity. A critical impediment to the efficacy of visual models trained on synthetic images is the domain gap that often exists between synthetic and real images, which can severely limit the models' generalization capabilities. To this end, we propose an iterative self-training (IST) strategy for training visual models on synthetic images. This strategy progressively refines both the model and pseudo labels through self-training and LabFilt, improving the performance of downstream visual tasks. LabFilt improves pseudo-label quality at both pixel and object levels through class-adaptive filtering mechanisms. The contribution of our framework is three-fold:

- We propose a class-aware text-to-image synthesis (CATS) component that automatically generates diverse synthetic images with accurate annotations, effectively addressing the laborious and time-intensive challenges of data collection and annotation in visual tasks.
- An iterative self-training (IST) strategy is developed to progressively optimize the model and pseudo labels through self-training and LabFilt. LabFilt enhances pseudo-label quality at pixel and object levels by employing class-adaptive filtering mechanisms.
- Extensive experiments on PASCAL VOC (Everingham et al. 2010) and MSCOCO (Lin et al. 2014) show that IST-CATS outperforms most existing synthetic, semi-supervised, and weakly-supervised methods in object detection and semantic segmentation.

Related Works

Text-to-Image Diffusion Models. Text-to-image diffusion models represent a new revolution in the field of image generation, bringing significant innovations and enabling the creation of high-quality images. Successful examples include LDMs (Rombach et al. 2022), Muse (Chang et al. 2023), and CONPREDIFF (Yang et al. 2024a). Although these methods can generate high-quality images that closely resemble the real world, they cannot directly generate images with bounding-box or pixel-level annotations. In this paper, we propose a class-aware text-to-image synthesis framework for automatically generating high-quality images with bounding-box and pixel-level annotations, aiming to alleviate the burden of data collection and annotation.

Synthetic Image Generation for Vision Tasks. To mitigate the extensive annotation requirements in vision tasks, image synthesis has emerged as a key strategy for generating labeled datasets. Early approaches focused on using synthetic images from 3D models or game engines for training vision models (Wu et al. 2019; Zheng and Yang 2021). Later, GANs were employed to create synthetic datasets, reducing reliance on manual annotation (Bosquet et al. 2023; Yang et al. 2024b). However, these methods often require expert input for 3D scene creation and tend to generate images focused on isolated objects, failing to capture the complexity of natural environments. More recently, diffusion models have demonstrated strong capabilities in generating diverse datasets for vision tasks, with notable works such as DiffuMask, DiffusionSeg, and Dataset diffusion. Despite this progress, previous methods have not fully explored the potential of sophisticated text-prompting strategies, limiting the diversity of generated images. Our approach leverages Large Language Models (LLMs) to generate more nuanced prompts, enabling the creation of a broader and more varied image set, thereby enhancing the performance of vision models through a richer training dataset.

Learning with Noisy Labels. Most existing methods for training models based on noisy labels typically involve correcting the loss function (Ma et al. 2020; Jiang et al. 2021), or implementing robust regularization techniques (Shorten and Khoshgoftaar 2019; Liu et al. 2022) to mitigate the impact of noisy labels on model performance. However, the task of designing a dependable metric to identify noise-affected samples is fraught with difficulties, often leading to the accumulation of errors due to incorrect sample classification. Recent studies have explored sample selection (Yang et al. 2022; Zhang et al. 2022), which seek to isolate accurately labeled examples from noisy data, thereby improving model performance. Despite their efficacy, they face challenges in excluding low-quality pseudo labels, particularly those associated with complex images. Our method, LabFilt, diverges from existing approaches by introducing a dual-strategy filtering mechanism that operates at both the pixel and object levels. LabFilt employs class-adaptive filtering techniques to meticulously refine pseudo labels, enhancing the precision of the training data.

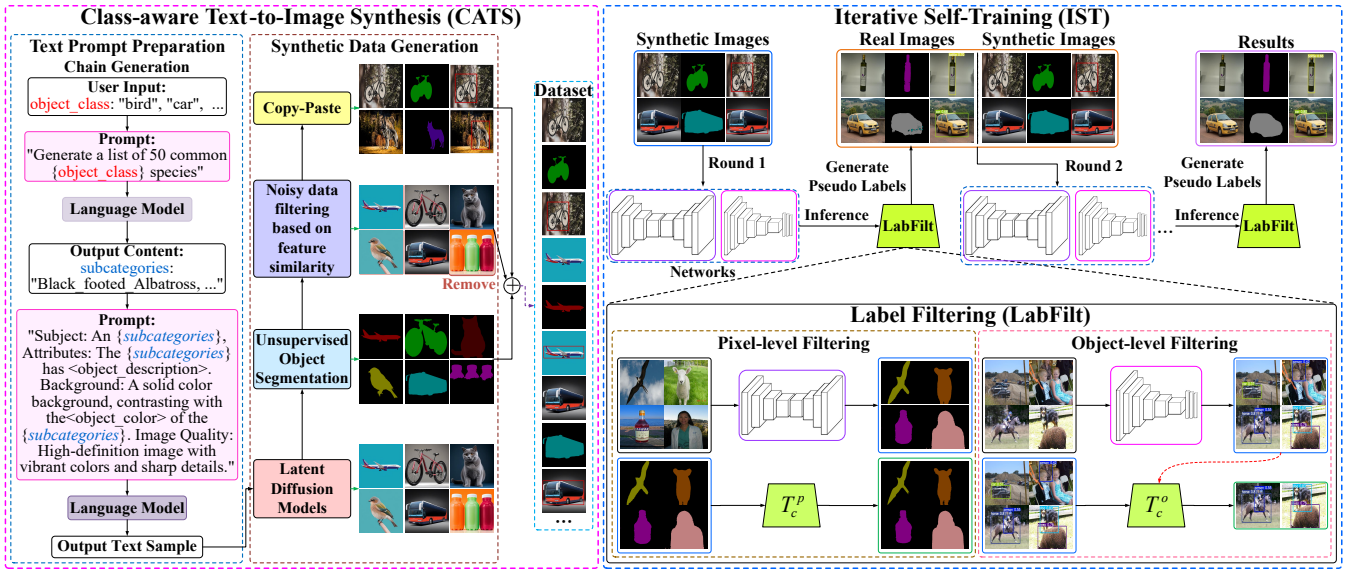


Figure 2: The framework of the proposed IST-CATS.

Methodology

As shown in Figure 2, IST-CATS consists of a class-aware text-to-image synthesis (CATS) component and an iterative self-training (IST) strategy with LabFilt. The CATS aims to generate diverse images with accurate annotations to reduce data collection and labeling efforts. IST leverages these generated images to progressively optimize downstream visual models and pseudo labels, mitigating domain gaps between generated and real data to enhance model robustness. LabFilt dynamically purifies pseudo-label quality through class-adaptive filtering mechanisms at pixel and object levels.

Class-aware Text-to-image Synthesis Strategy

To effectively support downstream vision tasks, text-to-image synthesis must address two key challenges: image diversity and label quality. To this end, we propose a novel class-aware text-to-image synthesis (CATS) strategy consisting of two main stages: preparing text prompts for latent diffusion models and generating synthetic images with annotations. Initially, for a given thematic target, we design an innovative class-aware chain approach to create detailed and diverse descriptions via LLMs. These are then input into LDMs, which produce high-quality images with varied appearances and clear distinctions from their backgrounds. Next, we use an unsupervised instance segmentation model and our feature similarity-based filtering mechanism to generate high-quality pseudo labels, and apply a copy-paste strategy (Ghiasi et al. 2021) to synthesize images with complex backgrounds, enhancing dataset realism.

Preparing Text Prompts for Latent Diffusion Models. Most existing methods (Wu et al. 2023a; Yoshihashi et al. 2023) rely on straightforward text prompts to generate images. They typically employ a "Direct Generation" strategy, where generic category labels like "bird" are slotted into predefined templates to craft image descriptions. This approach often fails to capture the diverse range of objects (e.g., "Black-footed Albatross") in the real world, resulting in visually similar images within the same category and reducing both the diversity of generated data and the generalization capability of downstream visual tasks. To this end, we introduce a class-aware chain approach, depicted in Figure 2 as *Chain Generation*. This approach initially utilizes templates filled with a class label to generate fine-grained subcategories through an LLM. The next round will refine these subcategories into detailed image descriptions, thus enhancing the diversity and specificity of the generated images. To make the generated image easy to segment, we add a cue for a clean background in the prompt. This approach mitigates the limitations associated with broad class prompts and facilitates a richer diversity of image outputs.

Synthetic Image Generation. After generating detailed textual descriptions, we input them into LDMs to create high-quality images with clear object-background separation. We then use the unsupervised instance segmentation model CutLER (Wang et al. 2023) to extract object masks. Despite efforts to ensure clean backgrounds, CutLER may still produce low-quality masks.

To further enhance segmentation quality, we introduce a noisy data filtering strategy based on feature similarity. Given an object feature cropped from its mask, if its features are very inconsistent with the full-image features, it can be suggested that the object mask may be missing essential parts of the semantic object. Conversely, if it is too similar to the image features, it might indicate that the mask has included too much of the background, making it not truly representative of the object alone. To implement this, we use the pre-trained VGG16 (Simonyan and Zisserman 2014) for feature extraction and compute the cosine similarity between the full-image and cropped-object features.

$$s_i^c = cos\_sim(VGG16(I_i), VGG16(O_i^c)), \quad (1)$$

where s_i^c is the similarity between the image I_i and the object O_i^c cropped from its mask in class c, $VGG16(\cdot)$ will output 4096-dimension feature vectors and $cos\_sim(\cdot)$ signifies cosine similarity. To identify and exclude outliers, we establish thresholds using the mean and standard deviation of these similarity scores across the dataset:

$$\textit{LabStatus}_i^c = \begin{cases} Positive, & if\ |\mu_c - \sigma_c| < s_i^c < |\mu_c + \sigma_c|, \\ Negative, & otherwise, \end{cases}$$
$$\textit{where} \quad \mu_c = \frac{1}{N^c}\sum_{i=1}^{N^c} s_i^c,\ \sigma_c = \frac{1}{N^c}\sum_{i=1}^{N^c}(s_i^c - \mu_c)^2 \tag{2}$$

Here μ_c denotes the mean of all images in class c, σ_c represents the variance, and N^c is the total number of images for class c. $\textit{LabStatus}_i^c$ indicates whether the label of the object in image I_i from class c should be retained (Positive) or removed (Negative), ensuring that only the most representative and highest-quality labels are used for training.

To create a training dataset with complex backgrounds, we employ a copy-paste strategy that merges generated object images with various background scenes. This strategy not only increases dataset complexity but also enhances the model's ability to discern and understand the relationships between objects and their environments, deepening its comprehension of common semantic themes.

(a) Pixel-Level Noise

(b) Object-Level Noise

Figure 3: Pseudo labels may introduce noise at both pixel and object levels. (a) Pixel-level noise occurs when pseudo labels mix objects with labels from other categories. (b) Object-level noise arises when objects in pseudo labels are incorrectly classified as other categories.

Iterative Self-Training

Domain gaps between synthetic and real images can hinder visual model generalization. We propose iterative self-training (IST) to bridge this gap, thereby improving the performance of downstream visual tasks. The framework of the IST is outlined in Figure 2. In this process, a segmenter and detector are utilized to respectively generate pixel-level and bounding-box pseudo labels for the unlabeled real images. We have developed an innovative label filtering strategy, termed LabFilt, to enhance the fidelity of these pseudo labels. The models for segmentation and detection are trained iteratively, employing both synthetic and refined pseudo-labeled real images. More specifically, our IST takes both synthetic data and unlabeled images as inputs. Initially, we train segmentation and detection models on synthetic data $S^l = \{(x_i^l, y_i^l, z_i^l)\}_{i=1}^N$, where N indicates the number of synthetic images, y_i^l and z_i^l represents the pixel-level and bounding-box annotations of the i^{th} image x_i^l, respectively. Such models are then utilized to generate pseudo labels (i.e., pixel-level pseudo labels and bounding-box pseudo labels) for unlabeled images $T^u = \{x_i^u\}_{i=1}^M$, where M is the number of unlabeled images. Subsequently, LabFilt is employed to filter noisy pseudo labels. Afterward, the segmentation and detection networks are trained on a combination of synthetic data and unlabeled images with pseudo labels. This iterative cycle is designed to continuously refine the pseudo labels and optimize the models' performance. To avoid overfitting, the model weights are trained from scratch, with the initialization parameters set to pre-trained weights on ImageNet (Deng et al. 2009) in each round.

Label Filtering

During the training of models for generating pseudo labels from synthetic data, notable discrepancies between synthetic and real images can lead to the production of noisy pseudo labels for real images. We categorize this noise into two levels, pixel-level and object-level, as depicted in Figure 3. To counteract the detrimental impact of these noisy labels on model performance, we develop a label filtering strategy, LabFilt, which includes both class-adaptive pixel-level and object-level pseudo-label filtering mechanisms.

Class-adaptive Pixel-level Pseudo-label Filtering. Traditional methods (Zoph et al. 2020; Feng et al. 2022) often involve manually setting confidence thresholds to filter out noise in pseudo labels, a process that can be imprecise and inconsistent due to the varying characteristics of different pixel categories. To tackle this issue, our approach involves a class-adaptive mechanism where we dynamically calculate the mean and variance of the confidence scores for all pixels within each class during each iteration of the IST. We then establish a filtering threshold based on the difference between the mean and the variance to selectively remove noisy pixels, thereby enhancing pseudo-label accuracy. The formula for this threshold is given by:

$$T_c^p = \mu_c^p - \sigma_c^p,$$
$$\textit{where} \quad \mu_c^p = \frac{1}{M^c}\sum_{i=1}^{M^c} p_i^c,\ \sigma_c^p = \frac{1}{M^c}\sum_{i=1}^{M^c}(p_i^c - \mu_c^p)^2, \tag{3}$$

where μ_c^p represents the average confidence score of all pixels in class c across all unlabeled images, while σ_c^p denotes the variance of these scores. The total number of pixels in class c is given by M^c, and p_i^c is the confidence score for the i^{th} pixel. T_c^p is the adaptive threshold of class c. The confidence score of the pseudo label of the c^{th} class pixel is lower than T_c^p to reclassify it into the background class.

This filtering mechanism is effective because segmentation models typically provide accurate predictions for the majority of data; hence, the mean serves as a robust filter for removing noise within pseudo labels. Nevertheless, the iterative nature of the training process can sometimes introduce

variability, potentially leading to inflated average confidence values μ_c^p. To accommodate this, we adjust the threshold by using the mean minus the variance, ensuring a more rigorous exclusion of noisy pixels. This class-adaptive pixel-level threshold significantly boosts the filtering capabilities of our approach, ensuring that only the most reliable pseudo labels are used for model training.

Class-adaptive Object-level Pseudo-label Filtering. While much research has focused on pixel-level pseudo-label filtering, the filtering of noisy objects at the object level has often been overlooked. For example, Figure 3(b) illustrates a case where the model misidentifies a "person" as a "motorbike". To address this issue, we introduce a class-adaptive object-level pseudo-label filtering strategy. In each iteration of the IST, we calculate the mean and variance of confidence scores for all objects within each class in the pseudo boxes generated by the detection model. We then use the disparity between the mean and variance as the threshold to eliminate noisy pseudo labels at the object level. The threshold formula is as follows:

$$\begin{aligned} & T_c^o = \mu_c^o - 2 * \sigma_c^o, \\ \textit{where} \quad & \mu_c^o = \frac{1}{K^c} \sum_{i=1}^{K^c} o_i^c, \ \sigma_c^o = \frac{1}{K^c} \sum_{i=1}^{K^c} (o_i^c - \mu_c^o)^2 \end{aligned} \tag{4}$$

where μ_c^o is the average confidence score for all objects of class c, σ_c^o denotes the variance, and K^c is the total number of objects for class c in all pseudo labels. T_c^o is the adaptive threshold of class c. To ensure sufficient training data, if the confidence score of at least one object in an image exceeds the threshold, the image and all its pseudo labels are retained. This approach emphasizes the use of LabFilt to filter pseudo labels consistently across both the detection and segmentation task training phases, employing the same training data for both tasks.

Experimental

Experimental Setup

Datasets and Evaluation Metrics. Due to the dual capabilities of our IST-CATS, which encompasses both semantic segmentation and object detection tasks, we evaluate the performance of our method separately using datasets tailored to each task in our experiments. To evaluate the performance of our IST-CATS in semantic segmentation tasks, we utilize category information from the PASCAL VOC 2012 and MSCOCO datasets as inputs for our class-aware text-to-image synthesis framework, resulting in the creation of the Syn-VOC and Syn-COCO datasets. The Syn-VOC dataset consists of $51,924$ images with 20 object classes, which are further divided into $46,731$ training images and $5,193$ validation images. The Syn-COCO includes $154,092$ images with 80 object classes, and it is split into $138,682$ training images and $15,410$ validation images. In this paper, we employ the mean Intersection-over-Union (mIoU) as a metric to evaluate the segmentation results on the PASCAL VOC 2012 and MSCOCO datasets. Regarding object detection tasks, our evaluation focused on the PASCAL VOC 2007 and 2012 *test* sets, as well as the MSCOCO *val* set. Drawing inspiration from prior research (Yin et al. 2023; Feng et al. 2024), we employ Average Precision (AP) and mean Average Precision (mAP) as metrics to evaluate the detection results. A prediction is considered a true positive only when the Jaccard overlap between the predicted bounding box and the corresponding ground-truth box exceeds 0.5.

Segmenter	Text Prompt	mIoU
DeepLabv3plus+SDA	*Direct Generation*	39.4
DeepLabv3plus+SDA+CRF	*Direct Generation*	39.9
DeepLabv3plus+SDA	*Chain Generation*	48.2
DeepLabv3plus+SDA+CRF	*Chain Generation*	48.7

Table 1: Performance of different text prompt selections. Both *Direct Generation* and *Chain Generation* use 46,731 training images and 5,193 validation images.

Implementation Details. In our experiment, the models are trained on an NVIDIA RTX 2080 Ti GPU using PyTorch. We employ the pretrained ResNet101 on ImageNet as the backbone for the segmentation network (i.e., DeepLabv3+ (Chen et al. 2018)). The network is trained with mini-batch stochastic gradient descent (SGD) using a batch size of 8, weight decay of 0.0002, and momentum of 0.9 over 60 epochs. We apply data augmentation techniques such as random horizontal flipping and random cropping, which resized the images to 513×513. Moreover, we integrate strong data augmentations (SDA) (DeVries and Taylor 2017) into the training images to introduce a more challenging optimization objective, thus enhancing the model's generalization capabilities. The initial learning rate for DeepLabv3+ is set to $4e^{-3}$ and decreases gradually using polynomial decay with a power of 0.9. During inference, we apply multi-scale testing and use conditional random field (CRF) with the hyperparameters recommended in (Chen et al. 2014) for post-processing. For object detection, we utilize YOLOv5x (Jocher et al. 2021) as our detector. During training, we opt for a batch size of 16 and initialize the learning rate to 0.00334, alongside a weight decay of 0.00025 and momentum of 0.74832. Input images are resized to 512×512 pixels, and training lasts for 50 epochs.

Ablation Studies

To evaluate the effectiveness of IST-CATS, we conduct ablation experiments on the PASCAL VOC dataset.

Comparison of Various Text Prompt Templates. In Table 1, we compare different text prompt selection methods. We find that the descriptive sentences generated for the target by *Chain Generation* can significantly improve the performance of the semantic segmentation compared to using *Direct Generation*. This is mainly attributed to the diverse and rich data generated with the assistance of the text descriptions produced by *Chain Generation*, which enhances the robustness of the model.

Effectiveness of Synthetic Images. We verify the effectiveness of synthetic data from two perspectives: 1) As shown in Table 2, the model trained on a combination of both simple and synthetic data exhibits significantly enhanced per-

Datasets	Training	Val	mIoU	07-mAP_{50}	12-mAP_{50}
Sim	20,769	5,193	35.4	53.3	49.6
Syn	20,769	5,193	39.5	54.4	52.0
Sim&Syn	46,731	5,193	48.7	63.9	57.3

Table 2: Performance of various synthetic data on PASCAL VOC with DeepLabv3plus+SDA+CRF for segmentation and YOLOV5x for detection.

Number of Images	mIoU	07-mAP_{50}	12-mAP_{50}
6492 (5842+650)	46.8	57.9	55.6
12982 (11683+1299)	47.4	58.5	56.2
25963 (23366+2597)	47.6	59.2	56.5
51924 (46731+5193)	48.7	63.9	57.3

Table 3: Comparing the performance of our method under different numbers of synthetic images.

formance compared to those trained solely on either simple or synthetic data. One potential reason for this performance enhancement may lie in the enriched diversity and increased data volume resulting from the fusion of simple and synthetic data in the training set. In Table 2, the abbreviations "Sim", "Syn", and "Sim&Syn" refer to "Simple", "Synthetic", and "Simple & Synthetic" respectively; 2) We investigate the influence of the quantity of synthetic data on performance. We train visual models using 12.5%, 25%, 50%, and 100% of images from the Syn-VOC dataset, respectively, and test them on the PASCAL VOC dataset. According to the results presented in Table 3, we can observe that the performance of the models gradually improves as the number of training images increases.

Effectiveness of Data Selective. To validate the potential effectiveness of our LabFilt, we compare the impact on network performance of removing noisy pseudo labels using LabFilt, removing noisy pseudo labels using class-adaptive pixel-level pseudo labels filtering (Filt-PL), and not removing noisy pseudo labels. Experimental results, as presented in Table 4, demonstrate that LabFilt significantly enhances performance. This is attributed to LabFilt's effective use of class-adaptive techniques at both pixel and object levels, which enhances the quality of pseudo labels and consequently improves model performance.

Analyzing the Impact of Different Models. We conducted experiments using PSPNet (Zhao et al. 2017) and Faster R-CNN (Ren et al. 2015) trained on the Syn-VOC dataset.

Iteration	Filter	Training	Val	mIoU	07-mAP_{50}	12-mAP_{50}
iter0	-	46,731	5,193	48.7	63.9	57.3
iter1	all	57,313	6,642	55.8	64.7	57.9
iter1	Filt-PL	57,313	6,642	56.5	65.1	58.2
iter1	LabFilt	49,383	5,581	56.7	66.0	58.8

Table 4: Comparative evaluation of filtering strategies.

Iteration	VOC dataset				
	Training	Val	mIoU	07-mAP_{50}	12-mAP_{50}
iter0	46,731	5,193	48.7	63.9	57.3
iter1	49,383	5,581	56.7	66.0	58.8
iter2	53,282	6,127	57.4	67.2	60.1
iter3	**53,902**	**6,207**	**59.3**	67.9	60.9
iter4	53,498	6,149	58.4	**68.6**	**63.0**
Iteration	MSCOCO dataset				
	Training	Val	mIoU	mAP_{50}	
iter0	138,682	15,410	17.89	19.3	
iter1	215,347	17,695	20.48	23.4	
iter2	**224,394**	**18,762**	**20.51**	25.8	
iter3	223,580	18,563	20.09	**30.5**	

Table 5: The iterative evaluation results of IST-CATS.

The initial segmentation performance on the PASCAL VOC 2012 *val* set was 46.8% mIoU. Detection results on the PASCAL VOC 2007 and 2012 *test* sets yielded 62.2% mAP and 56.2% mAP, respectively. These results suggest that stronger baseline models could enhance performance.

Effectiveness of Iterative Training. There is currently no mechanism to directly define the number of iterations for the model. In this study, we primarily determine when to stop the iterations by monitoring the growth of training data, specifically by stopping when the number of training images no longer increases. For example, for experiments conducted on the PASCAL VOC dataset, except for the initial round, where $5,193$ images from Syn-VOC are used as the *val* set, subsequent *val* sets are all from the *val* set of the PASCAL VOC 2012 dataset, and the annotations for these images are generated based on the inference results from the previous round. We calculate the model's performance after each iteration, which is shown in Table 5. It can be seen that the model's performance improves with more training rounds, eventually reaching a plateau after several iterations.

Semantic Segmentation

Comparison with Synthetic Data Methods. To validate the effectiveness of our IST-CATS, we perform a comparative evaluation against state-of-the-art semantic segmentation methods based on synthetic data, including GranSAM (Kundu et al. 2023), Goyal *et al.* (Goyal et al. 2018), Zhang *et al.* (Zhang et al. 2018), Attn2mask (Yoshihashi et al. 2023), DatasetDM (Wu et al. 2023a), and DiffuMask (Wu et al. 2023b). Experimental results are summarized in Table 6. On the PASCAL VOC 2012 dataset, our IST-CATS outperforms the state-of-the-art Attn2mask by 1.0% mIoU. On the MSCOCO dataset, IST-CATS achieves the highest mIoU score of 20.5%. An essential factor contributing to this is the capacity of our method to generate training data of high quality and diversity, thereby bolstering the model's generalization prowess.

Comparison with Semi-Supervised Semantic Segmentation Approaches. Our IST-CATS utilizes synthetic data and unlabeled data to gradually learn segmentation knowledge, which is akin to the fundamental principle of semi-

Methods	Training set	Val	mIoU
GranSAM	Syn(4k)	VOC12	25.2
Goyal *et al.*	Weak(10k)+Syn(2k)		55.5
Zhang *et al.*	Real(20k)+Syn(4k)		58.2
DiffuMask	DiffuMask		57.4
Attn2mask+BECO	Synth.imgs-168679		58.3
IST-CATS	Syn-VOC		**59.3**
GranSAM	Syn(16k)	MSCOCO	8.6
DatasetDM	Syn(8k)		17.6
IST-CATS	Syn-COCO		**20.5**

Table 6: Comparisons with competitive synthetic image methods on PASCAL VOC 2012 and MSCOCO datasets.

Dataset	Labeled	Unlabeled	Methods	mIoU
Syn-VOC	46,731	10,582	ST++	49.19
			UniMatch	31.98
			CorrMatch	50.97
			IST-CATS	**59.30**
Syn-COCO	138,682	118,288	ST++	14.84
			UniMatch	17.50
			CorrMatch	18.31
			IST-CATS	**20.51**

Table 7: Comparisons with competitive SSSS methods.

supervised semantic segmentation (SSSS) techniques that leverage a limited number of labeled images and a plethora of unlabeled images to enhance performance. Consequently, we employ synthetic data to substitute the labeled data required in SSSS, facilitating fair comparative experiments with semi-supervised methods (e.g., ST++ (Yang et al. 2022), UniMatch (Yang et al. 2023), and CorrMatch (Sun et al. 2024)). Specifically, we utilize the training set of the Syn-VOC/Syn-COCO dataset as our labeled data, while employing the training set of PASCAL VOC 2012/MSCOCO as unlabeled data. The *val* set of the Syn-VOC/Syn-COCO dataset serves as our *val* set, whereas the *val* set of PASCAL VOC 2012/MSCOCO is employed for testing purposes. As demonstrated in Table 7, our method exhibits notably superior performance compared to other competitors, surpassing them by a significant margin. The reason may come from the following two aspects: 1) Semi-supervised approaches necessitate high-quality labeled data that aligns with the domain distribution of the unlabeled data, posing challenges in achieving satisfactory performance when confronted with disparate domain distributions; 2) Our IST incrementally boosts model performance and enhances pseudo-label accuracy through self-training and our LabFilt strategy.

Object Detection

In Table 8, we compare our IST-CATS with existing weakly-supervised (i.e., PCL (Tang et al. 2018), MIST (Ren et al. 2020), CASD (Huang et al. 2020), CBL (Yin et al. 2023), NDI-MIL (Wang et al. 2024), and Feng *et al.* (Feng et al. 2024)) and synthetic data-based (i.e., Ge *et al.* (Ge et al. 2023), Peng *et al.* (Peng et al. 2015), Zhang *et al.* (Zhang et al. 2022)) object detection techniques on the PASCAL VOC 2007, 2012, and MSCOCO datasets. Our approach achieves the state-of-the-art performance of 68.6% mAP, 63.0% mAP, and 30.5% mAP on these three datasets, respectively. It surpasses the previous state-of-the-art weakly-supervised method by 11.2%, 9.1%, and 2.9%, outperforms the prior synthetic data-based method (i.e., Zhang *et al.*) by 9.3% and 7.9% on PASCAL VOC 2007 and 2012 datasets. Furthermore, as shown in Table 5, while our segmentation model has reached its peak performance, there is still room for further improvement in our detection model.

Methods	VOC		MSCOCO
	07-mAP_{50}	12-mAP_{50}	mAP_{50}
PCL	43.5	40.6	19.4
MIST	54.9	52.1	24.3
CASD	56.8	53.6	26.4
CBL	57.4	-	27.6
NDI-MIL	56.8	53.9	26.2
Feng *et al.*	55.6	50.7	-
Ge *et al.*	-	43.2	16.3
Peng *et al.*	31.2	-	-
Zhang *et al.*	59.3	55.1	-
Ours	**68.6**	**63.0**	**30.5**

Table 8: Performance comparison among the state-of-the-art methods on PASCAL VOC 2007, 2012, and MSCOCO datasets. "-" indicates that the result cannot be obtained.

Computational Complexity

The computational complexity of IST depends on the number of iterations and the base model (e.g., DeepLabv3+). When we trained DeepLabv3+ on an NVIDIA RTX 2080 Ti GPU with a batch size of 8, using 46,731 images for 30 epochs, taking about 48.5 hours. Testing on 12,031 images took only 573.6 seconds. Compared to the cost of annotating large datasets, our method's training overhead is minimal.

Conclusion

In this work, we introduce a novel framework (i.e., IST-CATS) for visual task learning. It employs our proposed CATS to automatically generate synthetic images with annotations (i.e., bounding-box annotations and pixel-level annotations) and uses IST with LabFilt to progressively learn segmentation and detection knowledge by harnessing synthetic data and unlabeled data. Extensive experimentation on the PASCAL VOC and MSCOCO datasets showcases the outstanding performance of our IST-CATS in both object detection and semantic segmentation. Comparative analysis against synthetic data and semi/weakly-supervised approaches unequivocally highlights the superior capabilities of our framework.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (Grant No. 62306237, Grant No. 62273275), the National Key Research and Development Program of China (No. 2022YFE0203800), and Xi'an Science and Technology Innovation and Qinchuangyuan Innovation Major Program (23ZDCYJSGG0009-2023).

References

Bosquet, B.; Cores, D.; Seidenari, L.; Brea, V. M.; Mucientes, M.; and Del Bimbo, A. 2023. A full data augmentation pipeline for small object detection based on generative adversarial networks. *Pattern Recognition*, 133: 108998.

Chang, H.; Zhang, H.; Barber, J.; Maschinot, A.; Lezama, J.; Jiang, L.; Yang, M.-H.; Murphy, K.; Freeman, W. T.; Rubinstein, M.; et al. 2023. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2014. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*.

Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 801–818.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, 248–255. Ieee.

DeVries, T.; and Taylor, G. W. 2017. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88: 303–338.

Feng, Y.; Zeng, H.; Li, S.; Liu, Q.; and Wang, Y. 2024. Refining and reweighting pseudo labels for weakly supervised object detection. *Neurocomputing*, 127387.

Feng, Z.; Zhou, Q.; Gu, Q.; Tan, X.; Cheng, G.; Lu, X.; Shi, J.; and Ma, L. 2022. Dmt: Dynamic mutual training for semi-supervised learning. *Pattern Recognition*, 108777.

Ge, Y.; Xu, J.; Zhao, B. N.; Joshi, N.; Itti, L.; and Vineet, V. 2023. Beyond generation: Harnessing text to image models for object detection and segmentation. *arXiv preprint arXiv:2309.05956*.

Ghiasi, G.; Cui, Y.; Srinivas, A.; Qian, R.; Lin, T.-Y.; Cubuk, E. D.; Le, Q. V.; and Zoph, B. 2021. Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2918–2928.

Goyal, M.; Rajpura, P.; Bojinov, H.; and Hegde, R. 2018. Dataset augmentation with synthetic images improves semantic segmentation. In *Computer Vision, Pattern Recognition, Image Processing, and Graphics: 6th National Conference, NCVPRIPG 2017, Mandi, India, December 16-19, 2017, Revised Selected Papers 6*, 348–359. Springer.

Huang, Z.; Zou, Y.; Kumar, B.; and Huang, D. 2020. Comprehensive attention self-distillation for weakly-supervised object detection. *Advances in neural information processing systems*, 33: 16797–16807.

Jiang, Z.; Zhou, K.; Liu, Z.; Li, L.; Chen, R.; Choi, S.-H.; and Hu, X. 2021. An information fusion approach to learning with instance-dependent label noise. In *International Conference on Learning Representations*.

Jocher, G.; Stoken, A.; Borovec, J.; Christopher, S.; and Laughing, L. C. 2021. Ultralytics/yolov5: V4. 0-Nn. silu () Activations Weights & Biases Logging Pytorch Hub Integration. *Zenodo*.

Kundu, R.; Paul, S.; Lal, R.; and Roy-Chowdhury, A. K. 2023. Towards Granularity-adjusted Pixel-level Semantic Annotation. *arXiv preprint arXiv:2312.02420*.

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 *ECCV*, 740–755. Springer.

Liu, S.; Liu, K.; Zhu, W.; Shen, Y.; and Fernandez-Granda, C. 2022. Adaptive early-learning correction for segmentation from noisy annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2606–2616.

Ma, C.; Yang, Y.; Ju, C.; Zhang, F.; Liu, J.; Wang, Y.; Zhang, Y.; and Wang, Y. 2023. Diffusionseg: Adapting diffusion towards unsupervised object discovery. *arXiv preprint arXiv:2303.09813*.

Ma, X.; Huang, H.; Wang, Y.; Romano, S.; Erfani, S.; and Bailey, J. 2020. Normalized loss functions for deep learning with noisy labels. In *International conference on machine learning*, 6543–6553. PMLR.

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.

Peng, X.; Sun, B.; Ali, K.; and Saenko, K. 2015. Learning deep object detectors from 3d models. In *Proceedings of the IEEE International Conference on Computer Vision*, 1278–1286.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Ren, Z.; Yu, Z.; Yang, X.; Liu, M.-Y.; Lee, Y. J.; Schwing, A. G.; and Kautz, J. 2020. Instance-Aware, Context-Focused, and Memory-Efficient Weakly Supervised Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10598–10607.

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.

Shehzadi, T.; Hashmi, K. A.; Stricker, D.; and Afzal, M. Z. 2024. Sparse semi-detr: Sparse learnable queries

for semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5840–5850.

Shorten, C.; and Khoshgoftaar, T. M. 2019. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1): 1–48.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sun, B.; Yang, Y.; Zhang, L.; Cheng, M.-M.; and Hou, Q. 2024. Corrmatch: Label propagation via correlation matching for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3097–3107.

Tang, P.; Wang, X.; Bai, S.; Shen, W.; Bai, X.; Liu, W.; and Yuille, A. 2018. Pcl: Proposal cluster learning for weakly supervised object detection. *IEEE transactions on pattern analysis and machine intelligence*, 42(1): 176–191.

Wang, G.; Zhang, X.; Peng, Z.; Zhang, T.; Tang, X.; Zhou, H.; and Jiao, L. 2024. Negative Deterministic Information-Based Multiple Instance Learning for Weakly Supervised Object Detection and Segmentation. *IEEE Transactions on Neural Networks and Learning Systems*.

Wang, X.; Girdhar, R.; Yu, S. X.; and Misra, I. 2023. Cut and learn for unsupervised object detection and instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3124–3134.

Wu, W.; Zhao, Y.; Chen, H.; Gu, Y.; Zhao, R.; He, Y.; Zhou, H.; Shou, M. Z.; and Shen, C. 2023a. Datasetdm: Synthesizing data with perception annotations using diffusion models. *Advances in Neural Information Processing Systems*, 36: 54683–54695.

Wu, W.; Zhao, Y.; Shou, M. Z.; Zhou, H.; and Shen, C. 2023b. 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.

Wu, Z.; Wang, X.; Gonzalez, J. E.; Goldstein, T.; and Davis, L. S. 2019. Ace: Adapting to changing environments for semantic segmentation. In *ICCV*, 2121–2130.

Yang, L.; Liu, J.; Hong, S.; Zhang, Z.; Huang, Z.; Cai, Z.; Zhang, W.; and Cui, B. 2024a. Improving diffusion-based image synthesis with context prediction. *Advances in Neural Information Processing Systems*, 36.

Yang, L.; Qi, L.; Feng, L.; Zhang, W.; and Shi, Y. 2023. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7236–7246.

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

Yang, L.; Zhuo, W.; Qi, L.; Shi, Y.; and Gao, Y. 2022. St++: Make self-training work better for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4268–4277.

Yin, Y.; Deng, J.; Zhou, W.; Li, L.; and Li, H. 2023. Cyclic-Bootstrap Labeling for Weakly Supervised Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7008–7018.

Yoon, S.-H.; Kwon, H.; Kim, H.; and Yoon, K.-J. 2024. Class Tokens Infusion for Weakly Supervised Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3595–3605.

Yoshihashi, R.; Otsuka, Y.; Tanaka, T.; et al. 2023. Attention as annotation: Generating images and pseudo-masks for weakly supervised semantic segmentation with diffusion. *arXiv preprint arXiv:2309.01369*.

Zhang, X.; Zhao, C.; Luo, H.; Zhao, W.; Zhong, S.; Tang, L.; Peng, J.; and Fan, J. 2022. Automatic learning for object detection. *Neurocomputing*, 484: 260–272.

Zhang, Y.; Wu, Z.; Zhou, Z.; and Wang, Y. 2018. Synthesizing Training Images for Semantic Segmentation. In *Image and Graphics Technologies and Applications: 13th Conference on Image and Graphics Technologies and Applications, IGTA 2018, Beijing, China, April 8–10, 2018, Revised Selected Papers 13*, 220–227. Springer.

Zhao, H.; Shi, J.; Qi, X.; Wang, X.; and Jia, J. 2017. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2881–2890.

Zheng, Z.; and Yang, Y. 2021. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *IJCV*, 1–15.

Zoph, B.; Ghiasi, G.; Lin, T.-Y.; Cui, Y.; Liu, H.; Cubuk, E. D.; and Le, Q. V. 2020. Rethinking pre-training and self-training. *arXiv preprint arXiv:2006.06882*.