DiverGen: Improving Instance Segmentation by Learning Wider Data Distribution with More Diverse Generative Data

Chengxiang Fan[1*] Muzhi Zhu[1*] Hao Chen[1†] Yang Liu[1] Weijia Wu[1] Huaqi Zhang[2] Chunhua Shen[1†]

[1] Zhejiang University, China [2] vivo Mobile Communication Co.

Abstract

Instance segmentation is data-hungry, and as model capacity increases, data scale becomes crucial for improving the accuracy. Most instance segmentation datasets today require costly manual annotation, limiting their data scale. Models trained on such data are prone to overfitting on the training set, especially for those rare categories. While recent works have delved into exploiting generative models to create synthetic datasets for data augmentation, these approaches do not efficiently harness the full potential of generative models.

To address these issues, we introduce a more efficient strategy to construct generative datasets for data augmentation, termed ***DiverGen****. Firstly, we provide an explanation of the role of generative data from the perspective of distribution discrepancy. We investigate the impact of different data on the distribution learned by the model. We argue that generative data can expand the data distribution that the model can learn, thus mitigating overfitting. Additionally, we find that the diversity of generative data is crucial for improving model performance and enhance it through various strategies, including category diversity, prompt diversity, and generative model diversity. With these strategies, we can scale the data to millions while maintaining the trend of model performance improvement. On the LVIS dataset, DiverGen significantly outperforms the strong model X-Paste, achieving* $+1.1$ *box AP and* $+1.1$ *mask AP across all categories, and* $+1.9$ *box AP and* $+2.5$ *mask AP for rare categories. Our codes are available at https://github.com/aim-uofa/DiverGen.*

1. Introduction

Instance segmentation [1, 3, 8] is one of the challenging tasks in computer vision, requiring the prediction of masks and categories for instances in an image, which serves as the foundation for numerous visual applications. As models' learning capabilities improve, the demand for training data increases. However, current datasets for instance segmentation heavily rely on manual annotation, which is time-consuming and costly, and the dataset scale cannot meet the training needs of models. Despite the recent emergence of the automatically annotated dataset SA-1B [11], it lacks category annotations, failing to meet the requirements of instance segmentation. Meanwhile, the ongoing development of the generative model has largely improved the controllability and realism of generated samples. For example, the recent text2image diffusion model [20, 22] can generate high-quality images corresponding to input prompts. Therefore, current methods [25, 26, 32] use generative models for data augmentation by generating datasets to supplement the training of models on real datasets and improve model performance. Although current methods have proposed various strategies to enable generative data to boost model performance, there are still some limitations: 1) Existing methods have not fully exploited the potential of generative models. First, some methods [32] not only use generative data but also need to crawl images from the internet, which is significantly challenging to obtain large-scale data. Meanwhile, the content of data crawled from the internet is uncontrollable and needs extra checking. Second, existing methods do not fully use the controllability of generative models. Current methods often adopt manually designed templates to construct prompts, limiting the potential output of generative models. 2) Existing methods [25, 26] often explain the role of generative data from the perspective of class imbalance or data scarcity, without considering the discrepancy between real-world data and generative data. Moreover, these methods typically show improved model performance only in scenarios with a limited number of real samples, and the effectiveness of generative data on existing large-scale real datasets, like LVIS [7], is not thoroughly investigated.

In this paper, we first explore the role of generative data from the perspective of distribution discrepancy, addressing two main questions: 1) *Why does generative data augmentation enhance model performance?* 2) *What types of generative data are beneficial for improving model perfor-*

*Equal contribution.
†Correspondence should be addressed to HC and CS.

mance? First, we find that there exist discrepancies between the model learned distribution of the limited real training data and the distribution of real-world data. We visualize the data and find that compared to the real-world data, generative data can expand the data distribution that the model can learn. Furthermore, we find that the role of adding generative data is to alleviate the bias of the real training data, effectively mitigating overfitting the training data. Second, we find that there are also discrepancies between the distribution of the generative data and the real-world data distribution. If these discrepancies are not handled properly, the full potential of the generative model cannot be utilized. By conducting several experiments, we find that using diverse generative data enables models to better adapt to these discrepancies, improving model performance.

Based on the above analysis, we propose an efficient strategy for enhancing data diversity, namely, *Generative Data Diversity Enhancement*. We design various diversity enhancement strategies to increase data diversity from the perspectives of *category diversity, prompt diversity, and generative model diversity*. For category diversity, we observe that models trained with generative data covering all categories adapt better to distribution discrepancy than models trained with partial categories. Therefore, we introduce not only categories from LVIS [7] but also extra categories from ImageNet-1K [21] to enhance category diversity in data generation, thereby reinforcing the model's adaptability to distribution discrepancy. For prompt diversity, we find that as the scale of the generative dataset increases, manually designed prompts cannot scale up to the corresponding level, limiting the diversity of output images from the generative model. Thus, we design a set of diverse prompt generation strategies to use large language models, like ChatGPT, for prompt generation, requiring the large language models to output maximally diverse prompts under constraints. By combining manually designed prompts and ChatGPT designed prompts, we effectively enrich prompt diversity and further improve generative data diversity. For generative model diversity, we find that data from different generative models also exhibit distribution discrepancies. Exposing models to data from different generative models during training can enhance adaptability to different distributions. Therefore, we employ Stable Diffusion [20] and DeepFloyd-IF [22] to generate images for all categories separately and mix the two types of data during training to increase data diversity.

At the same time, we optimize the data generation workflow and propose a four-stage generative pipeline consisting of instance generation, instance annotation, instance filtration, and instance augmentation. In the instance generation stage, we employ our proposed Generative Data Diversity Enhancement to enhance data diversity, producing diverse raw data. In the instance annotation stage, we introduce an annotation strategy called SAM-background. This strategy obtains high-quality annotations by using background points as input prompts for SAM [11], obtaining the annotations of raw data. In the instance filtration stage, we introduce a metric called CLIP inter-similarity. Utilizing the CLIP [19] image encoder, we extract embeddings from generative and real data, and then compute their similarity. A lower similarity indicates lower data quality. After filtration, we obtain the final generative dataset. In the instance augmentation stage, we use the instance paste strategy [32] to increase model learning efficiency on generative data.

Experiments demonstrate that our designed data diversity strategies can effectively improve model performance and maintain the trend of performance gains as the data scale increases to the million level, which enables large-scale generative data for data augmentation. On the LVIS dataset, DiverGen significantly outperforms the strong model X-Paste [32], achieving $+\mathbf{1.1}$ box AP [7] and $+\mathbf{1.1}$ mask AP across all categories, and $+\mathbf{1.9}$ box AP and $+\mathbf{2.5}$ mask AP for rare categories.

In summary, our main contributions are as follows:

- We explain the role of generative data from the perspective of distribution discrepancy. We find that generative data can expand the data distribution that the model can learn, mitigating overfitting the training set and the diversity of generative data is crucial for improving model performance.
- We propose the Generative Data Diversity Enhancement strategy to increase data diversity from the aspects of category diversity, prompt diversity, and generative model diversity. By enhancing data diversity, we can scale the data to millions while maintaining the trend of model performance improvement.
- We optimize the data generation pipeline. We propose an annotation strategy SAM-background to obtain higher-quality annotations. We also introduce a filtration metric called CLIP inter-similarity to filter data and further improve the quality of the generative dataset.

2. Related Work

Instance segmentation. Instance segmentation is an important task in the field of computer vision and has been extensively studied. Unlike semantic segmentation, instance segmentation not only classifies the pixels at a pixel level but also distinguishes different instances of the same category. Previously, the focus of instance segmentation research has primarily been on the design of model structures. Mask-RCNN [8] unifies the tasks of object detection and instance segmentation. Subsequently, Mask2Former [3] further unified the tasks of semantic segmentation and instance segmentation by leveraging the structure of DETR [1].

Orthogonal to these studies focusing on model architecture, our work primarily investigates how to better utilize generated data for this task. We focus on the challenging

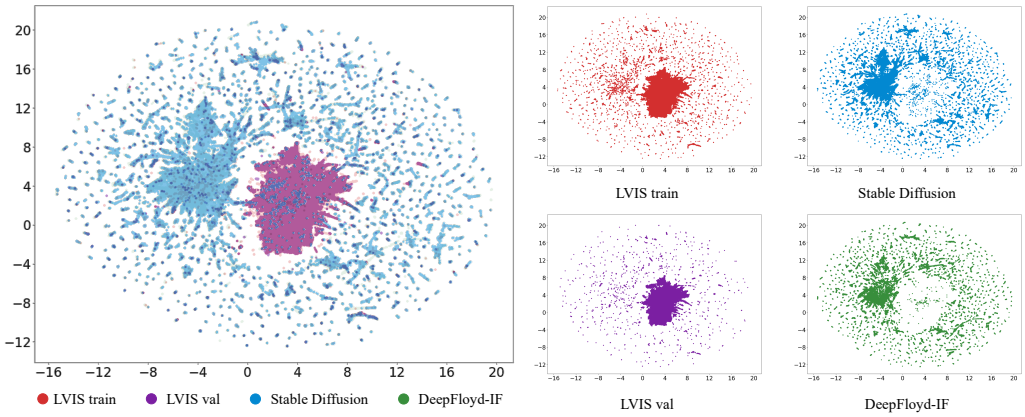

Figure 1. **Visualization of data distributions on different sources.** Compared to real-world data (LVIS train and LVIS val), generative data (Stable Diffusion and IF) can expand the data distribution that the model can learn.

long-tail dataset LVIS [7] because it is only the long-tailed categories that face the issue of limited real data and require generative images for augmentation, making it more practically meaningful.

Generative data augmentation. The use of generative models to synthesize training data for assisting perception tasks such as classification [5, 30], detection [2, 32], segmentation [13, 25, 26], etc. has received widespread attention from researchers. In the field of segmentation, early works [12, 31] utilize generative adversarial networks (GANs) to synthesize additional training samples. With the rise of diffusion models, there have been numerous efforts [13, 25, 26, 28, 32] to utilize text2image diffusion models, such as Stable Diffusion [20], to boost the segmentation performance. Li et al. [13] combine the Stable Diffusion model with a novel grounding module and establish an automatic pipeline for constructing a segmentation dataset. DiffuMask [26] exploits the potential of cross-attention maps between text and images to synthesize accurate semantic labels. More recently, FreeMask [28] uses a mask-to-image generation model to generate images conditioned on the provided semantic masks. However, the aforementioned work is only applicable to semantic segmentation. The most relevant work to ours is X-Paste [32], which promotes instance segmentation through copy-pasting the generative images and a filter strategy based on CLIP [19].

In summary, most methods only demonstrate significant advantages when training data is extremely limited. They consider generating data as a means to compensate for data scarcity or class imbalance. However, in this work, we take a further step to examine and analyze this problem from the perspective of data distribution. We propose a pipeline that enhances diversity from multiple levels to alleviate the impact of data distribution discrepancies. This provides new insights and inspirations for further advancements in this field.

3. Our Proposed DiverGen

3.1. Analysis of Data Distribution

Existing methods [26, 27, 32] often attribute the role of generative data to addressing class imbalance or data scarcity. In this paper, we provide an explanation for two main questions from the perspective of distribution discrepancy.

Why does generative data augmentation enhance model performance? We argue that there exist discrepancies between the model learned distribution of the limited real training data and the distribution of real-world data. The role of adding generative data is to alleviate the bias of the real training data, effectively mitigating overfitting the training data.

First, to intuitively understand the discrepancies between different data sources, we use CLIP [19] image encoder to extract the embeddings of images from different data sources, and then use UMAP [17] to reduce dimensions for visualization. Visualization of data distributions on different sources is shown in Figure 1. Real-world data (LVIS [7] train and LVIS val) cluster near the center, while generative data (Stable Diffusion [20] and IF [22]) are more dispersed, indicating that generative data can expand the data distribution that the model can learn.

Then, to characterize the distribution learned by the model, we employ the free energy formulation used by Joseph et al. [9]. This formulation transforms the logits outputted by the classification head into an energy function. The formulation is shown below:

$$F(\boldsymbol{q}; h) = -\tau \log \sum_{c=1}^{n} \exp\left(\frac{h_c(\boldsymbol{q})}{\tau}\right). \quad (1)$$

Here, $\boldsymbol{q}$ is the feature of instance, $h_c(\boldsymbol{q})$ is the c^{th} logit outputted by classification head $h(.)$, n is the number of categories and τ is the temperature parameter. We train

one model using only the LVIS train set (θ_{train}), and another model using LVIS train with generative data (θ_{gen}). Both models are evaluated on the LVIS val set and we use instances that are successfully matched by both models to obtain energy values. Additionally, we train another model using LVIS val (θ_{val}), treating it as representative of real-world data distribution. Then, we further fit Gaussian distributions to the histograms of energy values to obtain the mean μ and standard deviation σ for each model and compute the KL divergence [10] between them. $D_{KL}(p_{\theta_{\text{train}}} \| p_{\theta_{\text{val}}})$ is 0.063, and $D_{KL}(p_{\theta_{\text{gen}}} \| p_{\theta_{\text{val}}})$ is 0.019. The latter is lower, indicating that using generative data mitigates the bias of limited real training data.

Moreover, we also analyze the role of generative data from a metric perspective. We randomly select up to five images per category to form a minitrain set and then conduct inferences using θ_{train} and θ_{gen}. Then, we define a metric, termed train-val gap (TVG), which is formulated as follows:

$$\text{TVG}_w^k = \text{AP}_w^k minitrain - \text{AP}_w^k val. \quad (2)$$

Here, TVG_w^k is train-val gap of w category on task k, $\text{AP}_w^k d$ is AP [7] of w category on k obtained on dataset d, $w \in \{f, c, r\}$, with f, c, r standing for frequent, common, rare [7] respectively, and $k \in \{box, mask\}$, with box, $mask$ referring to the object detection and instance segmentation. The train-val gap serves as a measure of the disparity in the model's performance between the training and validation sets. A larger gap indicates a higher degree of overfitting the training set. The results, as presented in Table 1, show that the metrics for the rare categories consistently surpass those of frequent and common. This observation suggests that the model tends to overfit more on the rare categories that have fewer examples. With the augmentation of generative data, all TVG of θ_{gen} are lower than θ_{train}, showing that adding generative data can effectively alleviate overfitting the training data.

Data Source	TVG_f^{box}	TVG_f^{mask}	TVG_c^{box}	TVG_c^{mask}	TVG_r^{box}	TVG_r^{mask}
LVIS	13.16	10.71	21.80	16.80	39.59	31.68
LVIS + Gen	9.64	8.38	15.64	12.69	29.39	22.49

Table 1. **Results of train-val gap on different data sources.** With the augmentation of generative data, all TVG of LVIS are lower than LVIS + Gen, showing that adding generative data can effectively alleviate overfitting to the training data.

What types of generative data are beneficial for improving model performance? We argue that there are also discrepancies between the distribution of the generative data and the real-world data distribution. If these discrepancies are not properly addressed, the full potential of the generative model cannot be attained.

We divide the generative data into 'frequent', 'common', and 'rare' [7] groups, and train three models using each group of data as instance paste source. The inference results are shown in Table 2. We find that the metrics on the corresponding category subset are lowest when training with only one group of data. We consider model performance to be primarily influenced by the quality and diversity of data. Given that the quality of generative data is relatively consistent, we contend insufficient diversity in the data can mislead the distribution that the model can learn and a more comprehensive understanding is obtained by the model from a diverse set of data. Therefore, we believe that *using diverse generative data enables models to better adapt to these discrepancies*, improving model performance.

# Gen Category	AP_f^{box}	AP_f^{mask}	AP_c^{box}	AP_c^{mask}	AP_r^{box}	AP_r^{mask}
none	50.14	43.84	47.54	43.12	41.39	36.83
f	50.81	44.24	47.96	43.51	41.51	37.92
c	51.86	45.22	47.69	42.79	42.32	37.30
r	51.46	44.90	48.24	43.51	32.67	29.04
all	52.10	45.45	50.29	44.87	46.03	41.86

Table 2. **Results of different category data subset for training.** The metrics on the corresponding category subset are lowest when training with only one group of data, showing insufficient diversity in the data can mislead the distribution that the model can learn. Blue font means the lowest value in models using generative data.

3.2. Generative Data Diversity Enhancement

Through the analysis above, we find that the diversity of generative data is crucial for improving model performance. Therefore, we design a series of strategies to enhance data diversity at three levels: category diversity, prompt diversity, and generative model diversity, which help the model to better adapt to the distribution discrepancy between generative data and real data.

Category diversity. The above experiments show that including data from partial categories results in lower performance than incorporating data from all categories. We believe that, akin to human learning, the model can learn features beneficial to the current category from some other categories. Therefore, we consider increasing the diversity of data by adding extra categories. First, we select some extra categories besides LVIS from ImageNet-1K [21] categories based on WordNet [4] similarity. Then, the generative data from LVIS and extra categories are mixed for training, requiring the model to learn to distinguish all categories. Finally, we truncate the parameters in the classification head corresponding to the extra categories during inference, ensuring that the inferred category range remains within LVIS.

Prompt diversity. The output images of the text2image generative model typically rely on the input prompts. Existing methods [32] usually generate prompts by manually designing templates, such as "a photo of a single {*category_name*}." When the data scale is small, designing prompts manually is convenient and fast. However, when generating a large scale

of data, it is challenging to scale the number of manually designed prompts correspondingly. Intuitively, it is essential to diversify the prompts to enhance data diversity. To easily generate a large number of prompts, we choose large language model, like ChatGPT, to enhance the prompt diversity. We have three requirements for the large language model: 1) each prompt should be as different as possible; 2) each prompt should ensure that there is only one object in the image; 3) prompts should describe different attributes of the category. For example, if the category is food, prompts should cover attributes like color, brand, size, freshness, packaging type, packaging color, etc. Limited by the inference cost of ChatGPT, we use the manually designed prompts as the base and only use ChatGPT to enhance the prompt diversity for a subset of categories. Moreover, we also leverage the controllability of the generative model, adding the constraint "in a white background" after each prompt to make the background of output images simple and clear, which reduces the difficulty of mask annotation.

Generative model diversity. The quality and style of output images vary across generative models, and the data distribution learned solely from one generative model's data is limited. Therefore, we introduce multiple generative models to enhance the diversity of data, allowing the model to learn from wider data distributions. We selected two commonly used generative models, Stable Diffusion [20] (SD) and DeepFloyd-IF [22] (IF). We use Stable Diffusion V1.5, generating images with a resolution of 512 $\times$ 512, and use images output from Stage II of IF with a resolution of 256 $\times$ 256. For each category in LVIS, we generated 1k images using two models separately. Examples from different generative models are shown in Figure 2.

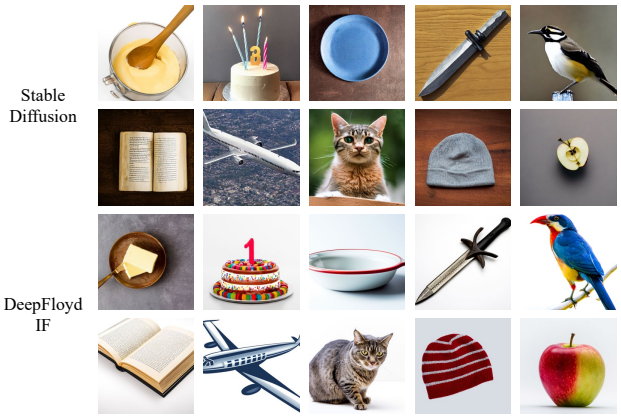

Figure 2. **Examples of various generative models.** The samples generated by different generative models vary, even within the same category.

3.3. Generative Pipeline

The generative pipeline of DiverGen is built upon X-Paste [32]. It can be divided into four stages: instance generation, instance annotation, instance filtration and instance augmentation. The overview of DiverGen is illustrated in Figure 3.

Instance generation. Instance generation is a crucial stage for enhancing data diversity. In this stage, we employ our proposed Generative Data Diversity Enhancement (GDDE), as mentioned in Sec 3.2. In category diversity enhancement, we utilize the category information from LVIS [7] categories and extra categories selected from ImageNet-1K [21]. In prompt diversity enhancement, we utilize manually designed prompts and ChatGPT designed prompts to enhance prompt diversity. In model diversity enhancement, we employ two generative models, SD and IF.

Instance annotation. We employ SAM [11] as our annotation model. SAM is a class-agnostic promptable segmenter that outputs corresponding masks based on input prompts, such as points, boxes, etc. In instance generation, leveraging the controllability of the generative model, the generative images have two characteristics: 1) each image predominantly contains only one foreground object; 2) the background of the images is relatively simple. Therefore, we introduce a SAM-background (SAM-bg) annotation strategy. SAM-bg takes the four corner points of an image as input prompts for SAM to obtain the background mask, then inverts the background mask as the mask of the foreground object. Due to the conditional constraints during the instance generation stage, this strategy is simple but effective in producing high-quality masks.

Instance filtration. In the instance filtration stage, X-Paste utilizes the CLIP score (similarity between images and text) as the metric for image filtering. However, we observe that the CLIP score is ineffective in filtering low-quality images. In contrast to the similarity between images and text, we think the similarity between images can better filter out low-quality images. Therefore, we propose a new metric called CLIP inter-similarity. We use the image encoder of CLIP [19] to extract image embeddings for objects in the training set and generative images, then calculate the similarity between them. If the similarity is too low, it indicates a significant disparity between the generative and real images, suggesting that it is probably a poor-quality image and needs to be filtered.

Instance augmentation. We use the augmentation strategy proposed by X-Paste [32] but do not use the data retrieved from the network or the instances in LVIS [7] training set as the paste data source, only use the generative data as the paste data source.

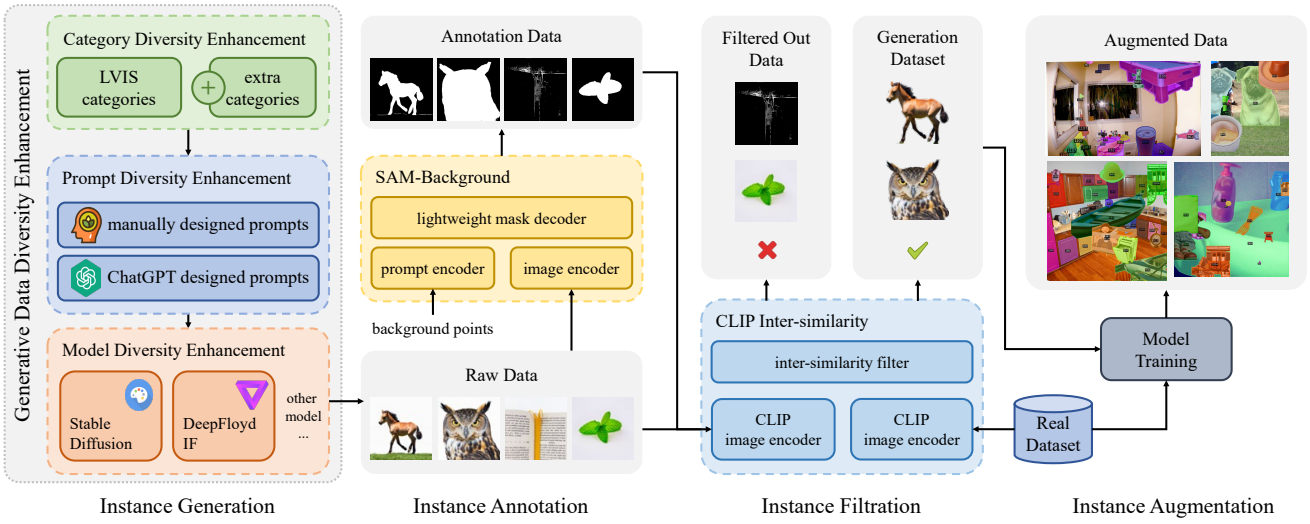

Figure 3. **Overview of the DiverGen pipeline.** In instance generation, we enhance data diversity at three levels: category diversity, prompt diversity, and generative model diversity. Next, we use SAM-background to obtain high-quality masks. Then, we use CLIP inter-similarity to filter out low-quality data. At last, we use the instance paste strategy to increase model learning efficiency on generative data.

4. Experiments

4.1. Settings

Datasets. We choose LVIS [7] for our experiments. LVIS is a large-scale instance segmentation dataset, containing 164k images with approximately two million high-quality annotations of instance segmentation and object detection. LVIS dataset uses images from COCO 2017 [14] dataset, but redefines the train/val/test splits, with around 100k images in the training set and around 20k images in the validation set. The annotations in LVIS cover 1,203 categories, with a typical long-tailed distribution of categories, so LVIS further divides the categories into frequent, common, and rare based on the frequency of each category in the dataset. We use the official LVIS training split and the validation split.

Evaluation metrics. The evaluation metrics are LVIS box average precision (AP^{box}) and mask average precision (AP^{mask}). We also provide the average precision of rare categories (AP_r^{box} and AP_r^{mask}). The maximum number of detections per image is 300.

Implementation details. We use CenterNet2 [33] as the baseline and Swin-L [15] as the backbone. In the training process, we initialize the parameters by the pre-trained Swin-L weights provided by Liu et al. [15]. The training size is 896 and the batch size is 16. The maximum training iterations is 180,000 with an initial learning rate of 0.0001. We use the instance paste strategy provided by Zhao et al. [32].

4.2. Main Results

Data diversity is more important than quantity. To investigate the impact of different scales of generative data, we use generative data of varying scales as paste data sources. We construct three datasets using only DeepFloyd-IF [22] with manually designed prompts, all containing original LVIS 1,203 categories, but with per-category quantities of 0.25k, 0.5k, and 1k, resulting in total dataset scales of 300k, 600k, and 1,200k. As shown in Table 3, we find that using generative data improves model performance compared to the baseline. However, as the dataset scale increases, the model performance initially improves but then declines. The model performance using 1,200k data is lower than that using 600k data. Due to the limited number of manually designed prompts, the generative model produces similar data, as shown in Figure 4a. Consequently, the model can not gain benefits from more data. However, when using our proposed Generative Data Diversity Enhancement (GDDE), due to the increased data diversity, the model trained with 1,200k images achieves better results than using 600k images, with an improvement of 1.21 box AP and 1.04 mask AP. Moreover, when using the same data scale of 600k, the mask AP increased by 0.64 AP and the box AP increased by 0.55 AP when using GDDE compared to not using it. The results demonstrate that data diversity is more important than quantity. When the scale of data is small, increasing the quantity of data can improve model performance, which we consider is an indirect way of increasing data diversity. However, this simplistic approach of solely increasing quantity to increase diversity has an upper limit. When it reaches this limit, explicit data diversity enhancement strategies become necessary to maintain the trend of model performance improvement.

Comparision with previous methods. We compare Di-

(a) Images of manually designed prompts.

(b) Images of ChatGPT designed prompts.

Figure 4. **Examples of generative data using different prompts.** By using prompts designed by ChatGPT, the diversity of generated images in terms of shapes, textures, etc. can be significantly improved.

# Gen Data	GDDE	AP^{box}	AP^{mask}	AP_r^{box}	AP_r^{mask}
0		47.50	42.32	41.39	36.83
300k		49.65	44.01	45.68	41.11
600k		50.03	44.44	47.15	41.96
1200k		49.44	43.75	42.96	37.91
600k	✓	50.67	44.99	48.52	43.63
1200k	✓	**51.24**	**45.48**	**50.07**	**45.85**

Table 3. **Results of different scales of generative data.** When using the same data scale, models using our proposed GDDE can achieve higher performance than those without it, showing that data diversity is more important than quantity.

verGen with previous data-augmentation related methods in Table 4. Compared to the baseline CenterNet2 [33], our method significantly improves, increasing box AP by +3.7 and mask AP by +3.2. Regarding rare categories, our method surpasses the baseline with +8.7 in box AP and +9.0 in mask AP. Compared to the previous strong model X-Paste [32], we outperform it with +1.1 in box AP and +1.1 in mask AP of all categories, and +1.9 in box AP and +2.5 in mask AP of rare categories. It is worth mentioning that, X-Paste utilizes both generative data and web-retrieved data as paste data sources during training, while our method exclusively uses generative data as the paste data source. We achieve this by designing diversity enhancement strategies, further unlocking the potential of generative models.

Method	Backbone	AP^{box}	AP^{mask}	AP_r^{box}	AP_r^{mask}
Copy-Paste [6]	EfficientNet-B7	41.6	38.1	-	32.1
Tan et al. [24]	ResNeSt-269	-	41.5	-	30.0
Detic [34]	Swin-B	46.9	41.7	45.9	41.7
CenterNet2 [33]	Swin-L	47.5	42.3	41.4	36.8
X-Paste [32]	Swin-L	50.1	44.4	48.2	43.3
DiverGen (Ours)	Swin-L	**51.2** (+1.1)	**45.5** (+1.1)	**50.1** (+1.9)	**45.8** (+2.5)

Table 4. **Comparison with previous methods on LVIS val set.**

4.3. Ablation Studies

We analyze the effects of the proposed strategies in DiverGen through a series of ablation studies using the Swin-L [15] backbone.

Effect of category diversity. We select 50, 250, and 566 extra categories from ImagNet-1K [21], and generate 0.5k images for each category, which are added to the baseline. The baseline only uses 1,203 categories of LIVS [7] to generate data. We show the results in Table 5. Generally, increasing the number of extra categories initially improves then declines model performance, peaking at 250 extra categories. The trend suggests that using extra categories to enhance category diversity can improve the model's generalization capabilities, but too many extra categories may mislead the model, leading to a decrease in performance.

# Extra Category	AP^{box}	AP^{mask}	AP_r^{box}	AP_r^{mask}
0	49.44	43.75	42.96	37.91
50	49.92	44.17	44.94	39.86
250	**50.59**	**44.77**	**47.99**	**42.91**
566	50.35	44.63	47.68	42.53

Table 5. **Ablation of the number of extra categories during training.** Using extra categories to enhance category diversity can improve the model's generalization capabilities, but too many extra categories may mislead the model, leading to a decrease in performance.

Effect of prompt diversity. We select a subset of categories and use ChatGPT to generate 32 and 128 prompts for each category, with each prompt being used to generate 8 and 2 images, respectively, ensuring that the image count for each category is 0.25k. The baseline uses only one prompt per category to generate 0.25k images. The regenerated images will replace the corresponding categories in the baseline to

ensure that the final data scale is consistent. The results are presented in Table 6. With the increase in prompt diversity, there is a continuous improvement in model performance, indicating that prompt diversity is indeed beneficial for enhancing model performance.

# Prompt	AP^{box}	AP^{mask}	AP_r^{box}	AP_r^{mask}
1	49.65	44.01	45.68	41.11
32	50.03	44.39	45.83	41.32
128	50.27	44.50	46.49	41.25

Table 6. **Ablation of the number of prompts used to generate data.** With the increase in prompt diversity, there is a continuous improvement in model performance, indicating that prompt diversity is indeed beneficial for enhancing model performance.

Effect of generative model diversity. We choose two commonly used generative models, Stable Diffusion [20] (SD) and DeepFloyd-IF [22] (IF). We generate 1k images per category for each generative model, totaling 1,200k. When using a mixed dataset (SD + IF), we take 600k from SD and 600k from IF per category, respectively, to ensure the total dataset scale is consistent. The baseline does not use any generative data (none). As shown in Table 7, using data generated by either SD or IF alone can improve performance, further mixing the generative data of both leads to significant performance gains. This demonstrates that increasing model diversity is beneficial for improving model performance.

Model	AP^{box}	AP^{mask}	AP_r^{box}	AP_r^{mask}
none	47.50	42.32	41.39	36.83
SD [20]	48.13	42.82	43.68	39.15
IF [22]	49.44	43.75	42.96	37.91
SD + IF	**50.78**	**45.27**	**48.94**	**44.35**

Table 7. **Ablation of different generative models.** Increasing model diversity is beneficial for improving model performance.

Effect of annotation strategy. X-Paste [32] uses four models (U2Net [18], SelfReformer [29], UFO [23] and CLIPseg [16]) to generate masks and selects the one with the highest CLIP score. We compare our proposed annotation strategy (SAM-bg) to that proposed by X-Paste (max CLIP). In Table 8, SAM-bg outperforms max CLIP strategy across all metrics, indicating that our proposed strategy can produce better annotations, improving model performance. As shown in Figure 5, SAM-bg unlocks the potential capability of SAM, obtaining precise and refined masks.

Effect of CLIP inter-similarity. We compare our proposed CLIP inter-similarity to CLIP score [32]. The results are shown in Table 9. The performance of data filtered by CLIP inter-similarity is higher than that of CLIP score, demonstrating that CLIP inter-similarity can filter low-quality images more effectively.

Figure 5. **Examples of object mask of different annotation strategies.** SAM-bg can obtain more complete and delicate masks.

Strategy	AP^{box}	AP^{mask}	AP_r^{box}	AP_r^{mask}
max CLIP [32]	49.10	43.45	42.75	37.55
SAM-bg	**49.44**	**43.75**	**42.96**	**37.91**

Table 8. **Ablation of different annotation strategies.** Our proposed SAM-bg can produce better annotations, improving model performance.

Strategy	AP^{box}	AP^{mask}	AP_r^{box}	AP_r^{mask}
none	49.44	43.75	42.96	37.91
CLIP score [32]	49.84	44.27	44.83	40.82
CLIP inter-similarity	**50.07**	**44.44**	**45.53**	**41.16**

Table 9. **Ablation of the different filtration strategies.** Our proposed CLIP inter-similarity can filter low-quality images more effectively.

5. Conclusions

In this paper, we explain the role of generative data augmentation from the perspective of data distribution discrepancies and find that generative data can expand the data distribution that the model can learn, mitigating overfitting the training set. Furthermore, we find that data diversity of generative data is crucial for improving model performance. Therefore, we design an efficient data diversity enhancement strategy, Generative Data Diversity Enhancement. We design various diversity enhancement strategies to increase data diversity from the aspects of category diversity, prompt diversity, and generative model diversity. Finally, we optimize the data generative pipeline by designing the annotation strategy SAM-background to obtain higher quality annotations and introducing the metric CLIP inter-similarity to filter data, which further improves the quality of the generative dataset. Through these designed strategies, our proposed method significantly outperforms the existing strong models. We hope DiverGen can provide new insights and inspirations for future research on the effectiveness and efficiency of generative data augmentation.

Acknowledgments

This work was in part supported by National Key R&D Program of China (No. 2022ZD0118700).

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proc. Eur. Conf. Comp. Vis.* Springer, 2020. 1, 2

[2] Kai Chen, Enze Xie, Zhe Chen, Lanqing Hong, Zhenguo Li, and Dit-Yan Yeung. Integrating geometric control into text-to-image diffusion models for high-quality detection data generation via text prompt. *arXiv: Comp. Res. Repository*, 2023. 3

[3] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 1290–1299, 2022. 1, 2

[4] Christiane Fellbaum. Wordnet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer, 2010. 4

[5] Chun-Mei Feng, Kai Yu, Yong Liu, Salman Khan, and Wangmeng Zuo. Diverse data augmentation with diffusions for effective test-time prompt tuning. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 2704–2714, 2023. 3

[6] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D. Cubuk, Quoc V. Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 2918–2928, 2021. 7

[7] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 5356–5364, 2019. 1, 2, 3, 4, 5, 6, 7

[8] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 2961–2969, 2017. 1, 2

[9] KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 5830–5840, 2021. 3

[10] James M Joyce. Kullback-leibler divergence. In *International Encyclopedia of Statistical Science*, pages 720–722. Springer, 2011. 4

[11] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander Berg, Wan-Yen Lo, et al. Segment anything. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 4015–4026, 2023. 1, 2, 5

[12] Daiqing Li, Huan Ling, Seung Wook Kim, Karsten Kreis, Sanja Fidler, and Antonio Torralba. Bigdatasetgan: Synthesizing imagenet with pixel-wise annotations. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 21330–21340, 2022. 3

[13] Ziyi Li, Qinye Zhou, Xiaoyun Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Open-vocabulary object segmentation with diffusion models. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 7667–7676, 2023. 3

[14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proc. Eur. Conf. Comp. Vis.*, pages 740–755. Springer, 2014. 6

[15] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 10012–10022, 2021. 6, 7

[16] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 7086–7096, 2022. 8

[17] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv: Comp. Res. Repository*, 2018. 3

[18] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R Zaiane, and Martin Jagersand. U2-net: Going deeper with nested u-structure for salient object detection. *Pattern Recognition*, 106:107404, 2020. 8

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proc. Int. Conf. Mach. Learn.*, pages 8748–8763. PMLR, 2021. 2, 3, 5

[20] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 10684–10695, 2022. 1, 2, 3, 5, 8

[21] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vision*, 115: 211–252, 2015. 2, 4, 5, 7

[22] Alex Shonenkov, Misha Konstantinov, Daria Bakshandaeva, Christoph Schuhmann, Ksenia Ivanova, and Nadiia Klokova. Deepfloyd-if, 2023. 1, 2, 3, 5, 6, 8

[23] Yukun Su, Jingliang Deng, Ruizhou Sun, Guosheng Lin, Hanjing Su, and Qingyao Wu. A unified transformer framework for group-based segmentation: Co-segmentation, co-saliency detection and video salient object detection. *IEEE Trans. Multimedia*, 2023. 8

[24] Jingru Tan, Gang Zhang, Hanming Deng, Changbao Wang, Lewei Lu, Quanquan Li, and Jifeng Dai. 1st place solution of LVIS challenge 2020: A good box is not a guarantee of a good mask. *arXiv: Comp. Res. Repository*, 2020. 7

[25] Weijia Wu, Yuzhong Zhao, Hao Chen, Yuchao Gu, Rui Zhao, Yefei He, Hong Zhou, Mike Zheng Shou, and Chunhua Shen. DatasetDM: Synthesizing data with perception annotations using diffusion models. *Proc. Advances in Neural Inf. Process. Syst.*, 2023. 1, 3

[26] Weijia Wu, Yuzhong Zhao, Mike Zheng Shou, Hong Zhou, and Chunhua Shen. Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. *Proc. IEEE Int. Conf. Comp. Vis.*, 2023. 1, 3

[27] Jiahao Xie, Wei Li, Xiangtai Li, Ziwei Liu, Yew Soon Ong, and Chen Change Loy. Mosaicfusion: Diffusion models as data augmenters for large vocabulary instance segmentation. *arXiv: Comp. Res. Repository*, 2023. 3

[28] Lihe Yang, Xiaogang Xu, Bingyi Kang, Yinghuan Shi, and Hengshuang Zhao. FreeMask: Synthetic images with dense annotations make stronger segmentation models. *Proc. Advances in Neural Inf. Process. Syst.*, 2023. 3

[29] Yi Ke Yun and Weisi Lin. Selfreformer: Self-refined network with transformer for salient object detection. *arXiv: Comp. Res. Repository*, 2022. 8

[30] Renrui Zhang, Xiangfei Hu, Bohao Li, Siyuan Huang, Hanqiu Deng, Yu Qiao, Peng Gao, and Hongsheng Li. Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 15211–15222, 2023. 3

[31] Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. Datasetgan: Efficient labeled data factory with minimal human effort. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 10145–10155, 2021. 3

[32] Hanqing Zhao, Dianmo Sheng, Jianmin Bao, Dongdong Chen, Dong Chen, Fang Wen, Lu Yuan, Ce Liu, Wenbo Zhou, Qi Chu, Weiming Zhang, and Nenghai Yu. X-paste: Revisiting scalable copy-paste for instance segmentation using CLIP and stablediffusion. *Proc. Int. Conf. Mach. Learn.*, 2023. 1, 2, 3, 4, 5, 6, 7, 8

[33] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Probabilistic two-stage detection. *arXiv: Comp. Res. Repository*, 2021. 6, 7

[34] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *Proc. Eur. Conf. Comp. Vis.*, pages 350–368. Springer, 2022. 7