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]

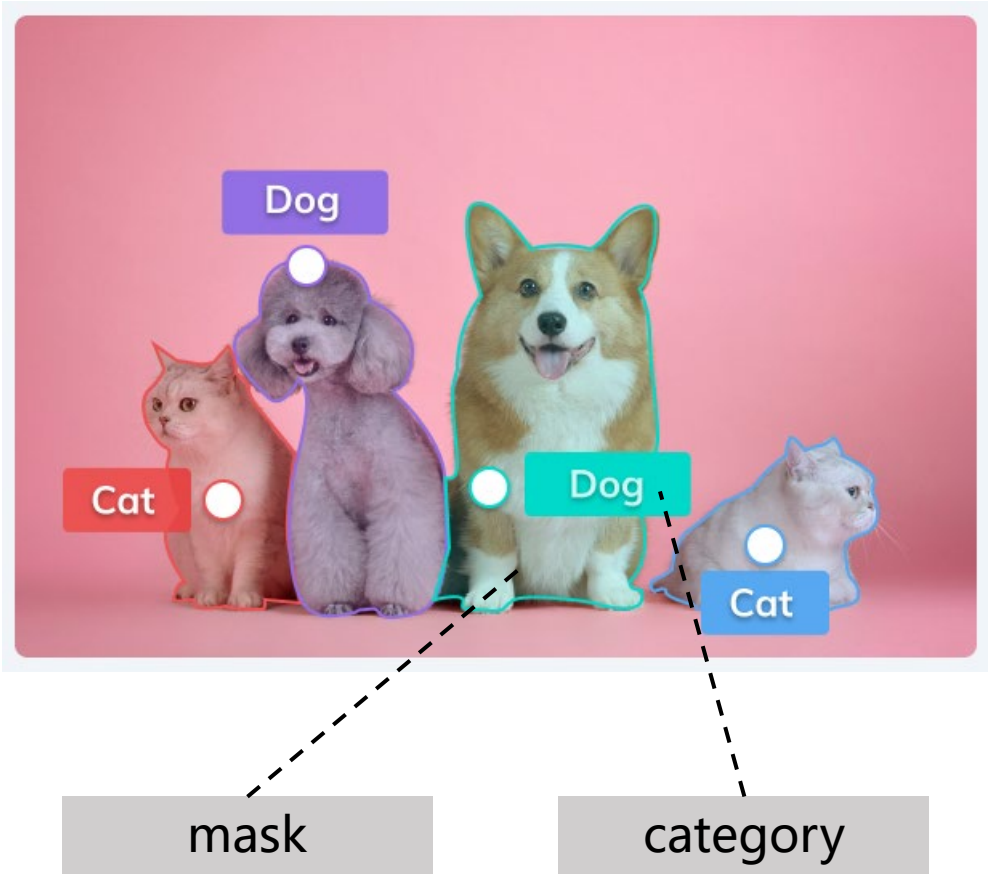
Weijia Wu[1] Huaqi Zhang[2] Chunhua Shen[1†]

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

Introduction - Task Definition

instance segmentation is one of the challenging tasks in computer vision

require the prediction of **masks** and **categories** for **instances** in an image



Introduction - LVIS

LVIS is a **large-scale** instance segmentation dataset

- 164k images
- 2M **high-quality** annotations
- 1203 categories
- **long-tailed** distribution (“natural setting”)
- frequent, common, and rare category group

samples from LVIS dataset

Background

- instance segmentation is data-hungry
- 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

the Need for Scalable Data Acquisition

+ the ongoing development of the generative model has largely improved the controllability and realism of generated samples

current methods use generative models for data augmentation by generating datasets to supplement the training of models on real datasets and improve model performance

Background

Existing methods of generative data augmentation:

1) not fully exploit the potential of generative models

crawl images from the internet → challenging, uncontrollable

only manually designed prompts → limiting the potential output

2) not consider the discrepancy between real-world data and generative data

explain the role of generative data from the perspective of class imbalance or data scarcity

typically show improve model performance only in scenarios with a limited number of real samples

Analysis of Data Distribution

explore the role of generative data from the perspective of distribution discrepancy

two main questions:

1) **Why does generative data augmentation enhance model performance?**
2) **What types of generative data are beneficial for improving model performance?**

Analysis of Data Distribution

Why does generative data augmentation enhance model performance?

the role of adding generative data is to alleviate the bias of the real training data, effectively mitigating overfitting the training data

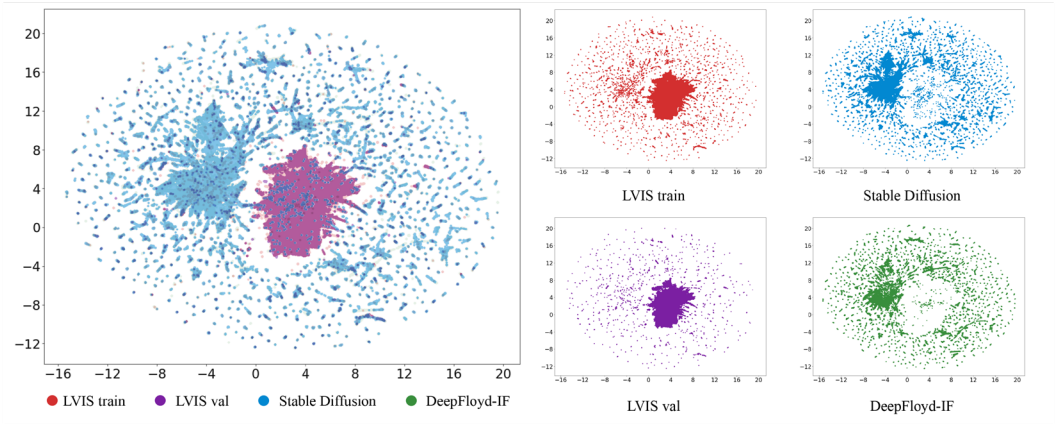


Figure 1. **Visualization of data distributions on different sources.** Compared to real-world data (LVIS train and LVIS val), generative data (Stable Diffusion and IF) can expand the data distribution that the model can learn.

$$\mathrm{TVG}_w^k = \mathrm{AP}_w^k minitrain - \mathrm{AP}_w^k val.$$

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.

Analysis of Data Distribution

What types of generative data are beneficial for improving model performance?

1) insufficient diversity in the data can mislead the distribution that the model can learn

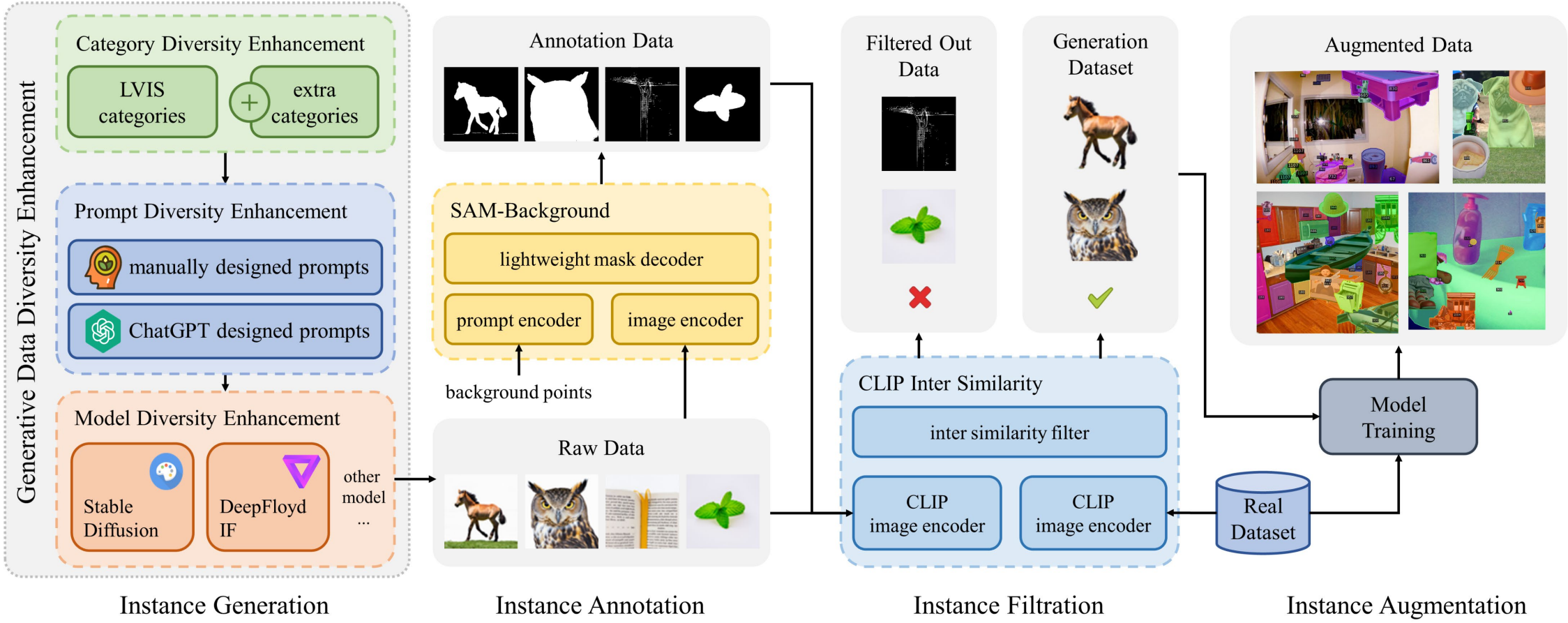
2) using diverse generative data enables models to better adapt to these discrepancies, improving model performance

# 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.

Method - DiverGen

Overview of the DiverGen pipeline



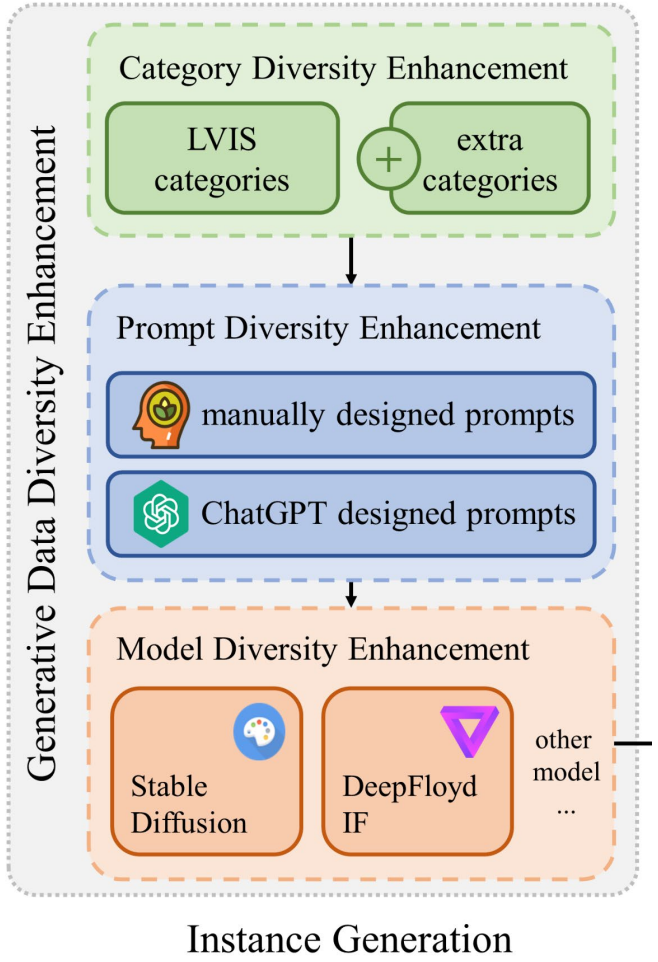
DiverGen - GDDE

Generative Data Diversity Enhancement

enhance data diversity at three levels:

- category diversity
- prompt diversity
- generative model diversity

help the model to better adapt to the distribution discrepancy between generative data and real data



DiverGen - Category Diversity

including data from partial categories results in lower performance than incorporating data from all categories

akin to human learning, the model can learn features beneficial to the current category from some other categories

increase the diversity of data by adding extra categories

Implementation Details:

LVIS categories + selected ImageNet-1K categories (based on WordNet similarity)

training phase: mix, requiring the model to learn to distinguish all categories

inference phase: truncate the parameters in the classification head

DiverGen - Prompt Diversity

manually designed template (challenging to scale)

a photo of a single {category_name}

ChatGPT designed prompt

three requirements:

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

controllability of the generative model

add "in a white background" after each prompt → simple and clear background reduces the difficulty of mask annotation

DiverGen – Example of ChatGPT

Human Input:

Please write 20 prompts for stable diffusion to generate images of cream pitcher.
Each prompt needs to be different, the focus of the description should be the different attributes of the category, such as if the category is animal, the prompts should describe the animal's color, size, age, species, etc., if it is food, the prompts should describe the color of food, brand, size, freshness, packaging type, packaging color, etc.
Each prompt should ensure that there is only one entity in the generated image. The prompt should start with words like "a photo of ", "a real photo of", "an image of ", "a photography of ", "an illustration of ", etc.
The output result should not contain anything else, just a prompt list in the format:
1. promptA
2. promptB
3. promptC
...
The first sentence of the output is "this is the list:".

DiverGen – Example of ChatGPT

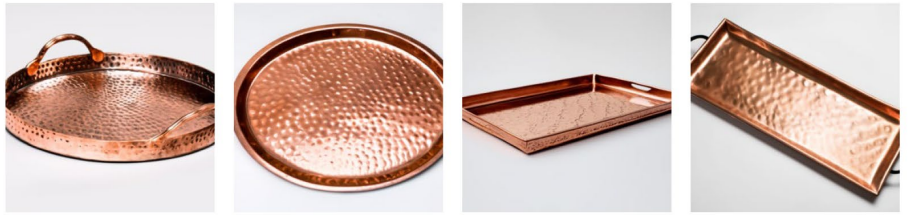
ChatGPT Response:

This is the list:

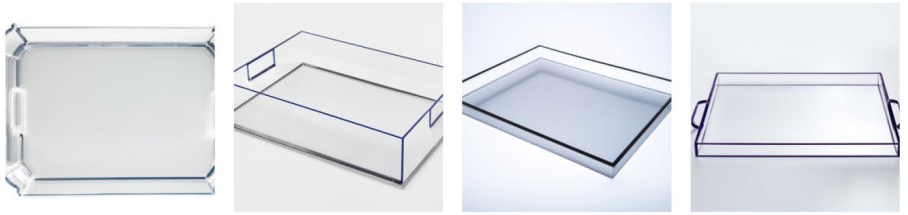
1. A photo of a small cream pitcher with a glossy white finish.
2. An image of a vintage cream pitcher with delicate floral patterns.
3. A real photo of a large cream pitcher made of fine bone china.
4. A photography of a modern cream pitcher with a sleek silver handle.
5. An illustration of a cream pitcher in a vibrant shade of turquoise.
6. A photo of a small cream pitcher with a rustic, weathered appearance.
7. An image of a cream pitcher adorned with intricate gold filigree.
8. A real photo of a tall cream pitcher with a matte black finish.
9. A photography of a cream pitcher in a classic white color with a scalloped edge.
10. An illustration of a cream pitcher shaped like a whimsical cow.
11. A photo of a vintage cream pitcher with a charming polka dot design.
12. An image of a cream pitcher made of clear glass with a frosted bottom.
13. A real photo of a cream pitcher in a unique hourglass shape.
14. A photography of a cream pitcher with a playful print of colorful flowers.
15. An illustration of a cream pitcher with a hand-painted blue and white pattern.
16. A photo of a small cream pitcher made of delicate porcelain.
17. An image of a cream pitcher with a textured surface resembling tree bark.
18. A real photo of a cream pitcher in a vibrant shade of coral pink.
19. A photography of a cream pitcher with a gold-trimmed spout and handle.
20. An illustration of a cream pitcher shaped like a graceful swan.

DiverGen – Samples

A photo of a copper tray with a hammered texture, in a white background



A real photo of a large tray made of clear acrylic material, in a white background



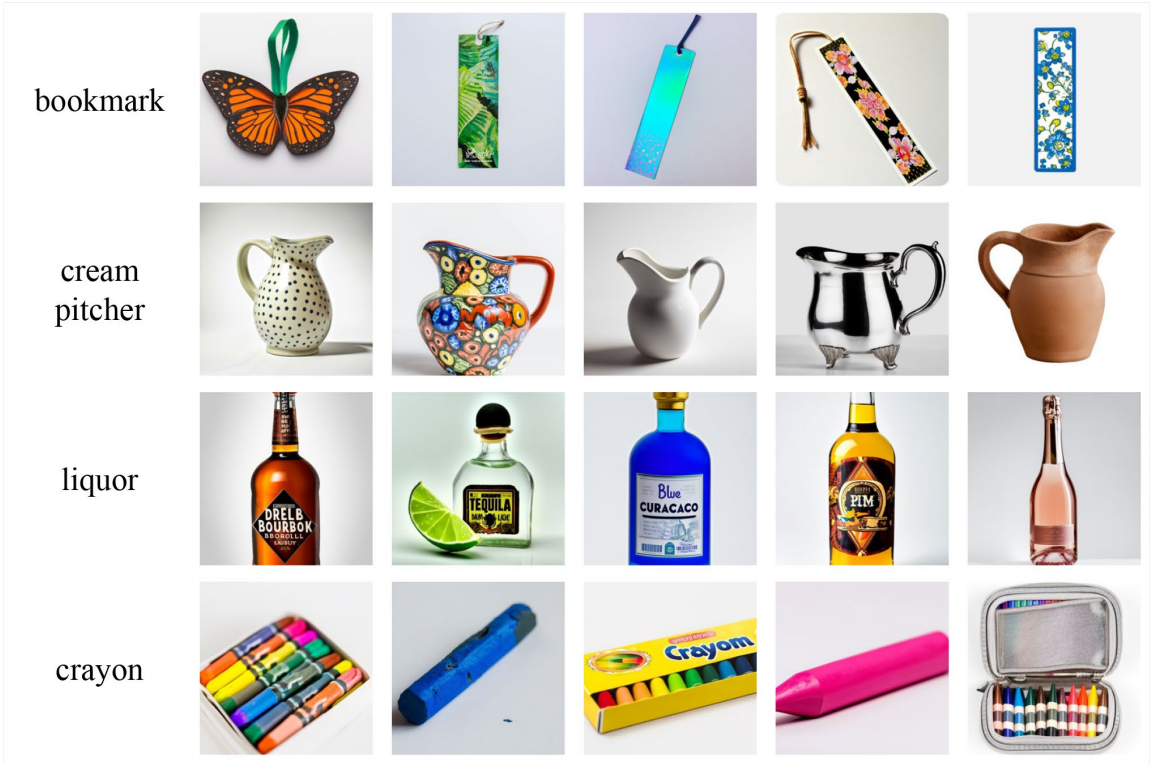
A photo of a large tray with vibrant floral designs, in a white background

An image of a wooden tray with intricate carvings, in a white background

A photo of a small ceramic tray in a vibrant turquoise color, in a white background

DiverGen – Comparison

(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.

DiverGen - 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

introduce multiple generative models to enhance the diversity of data, allowing the model to learn from wider data distributions

Stable Diffusion: 512 × 512

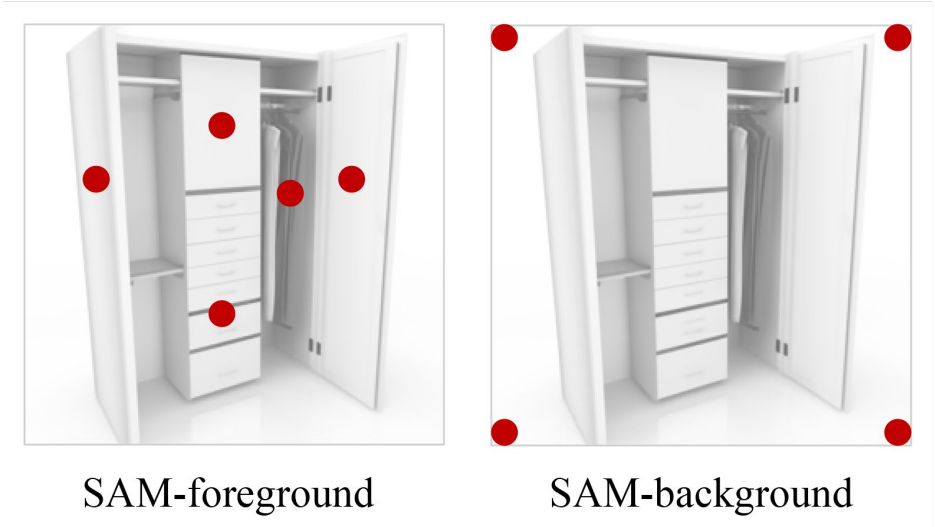
DeepFloyd-IF: 256 × 256

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

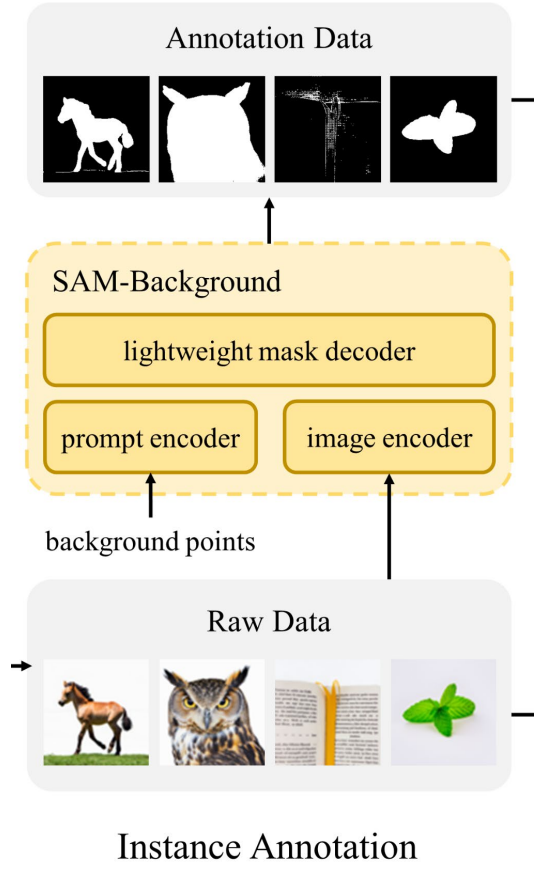
DiverGen - SAM-background

step1: take the four corner points of an image as input prompts for SAM to obtain the background mask

step2: invert the background mask as the mask of the foreground object



SAM-foreground SAM-background

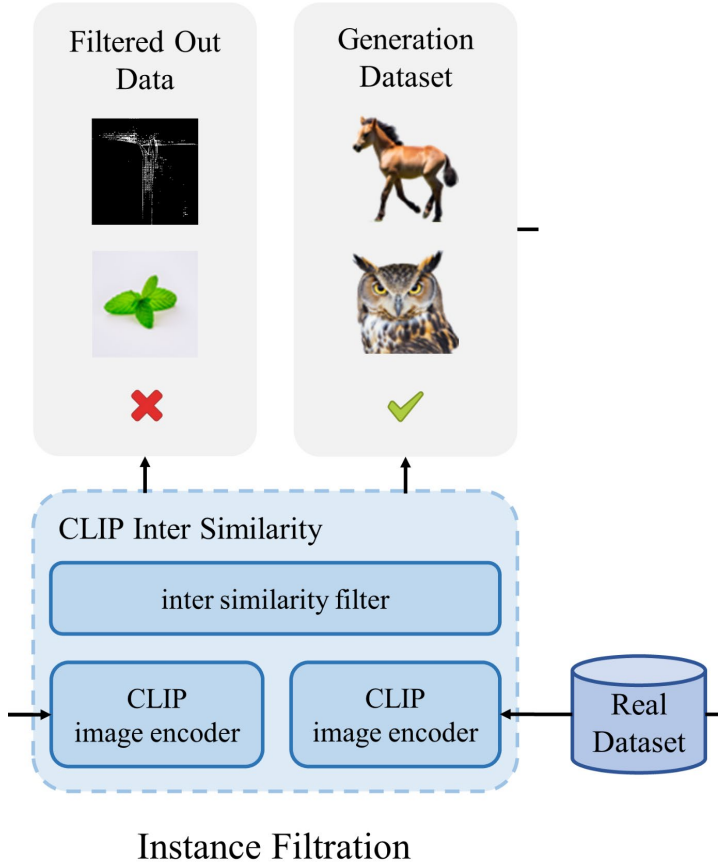


Instance Annotation

DiverGen - CLIP Inter-similarity

step1: use the image encoder of CLIP to extract image embeddings for objects in the training set and generative images

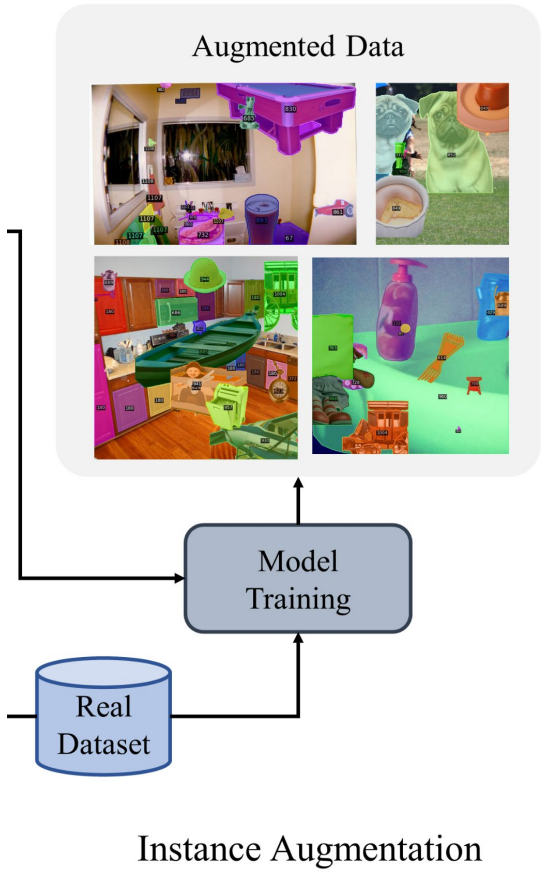
step2: calculate the similarity between them



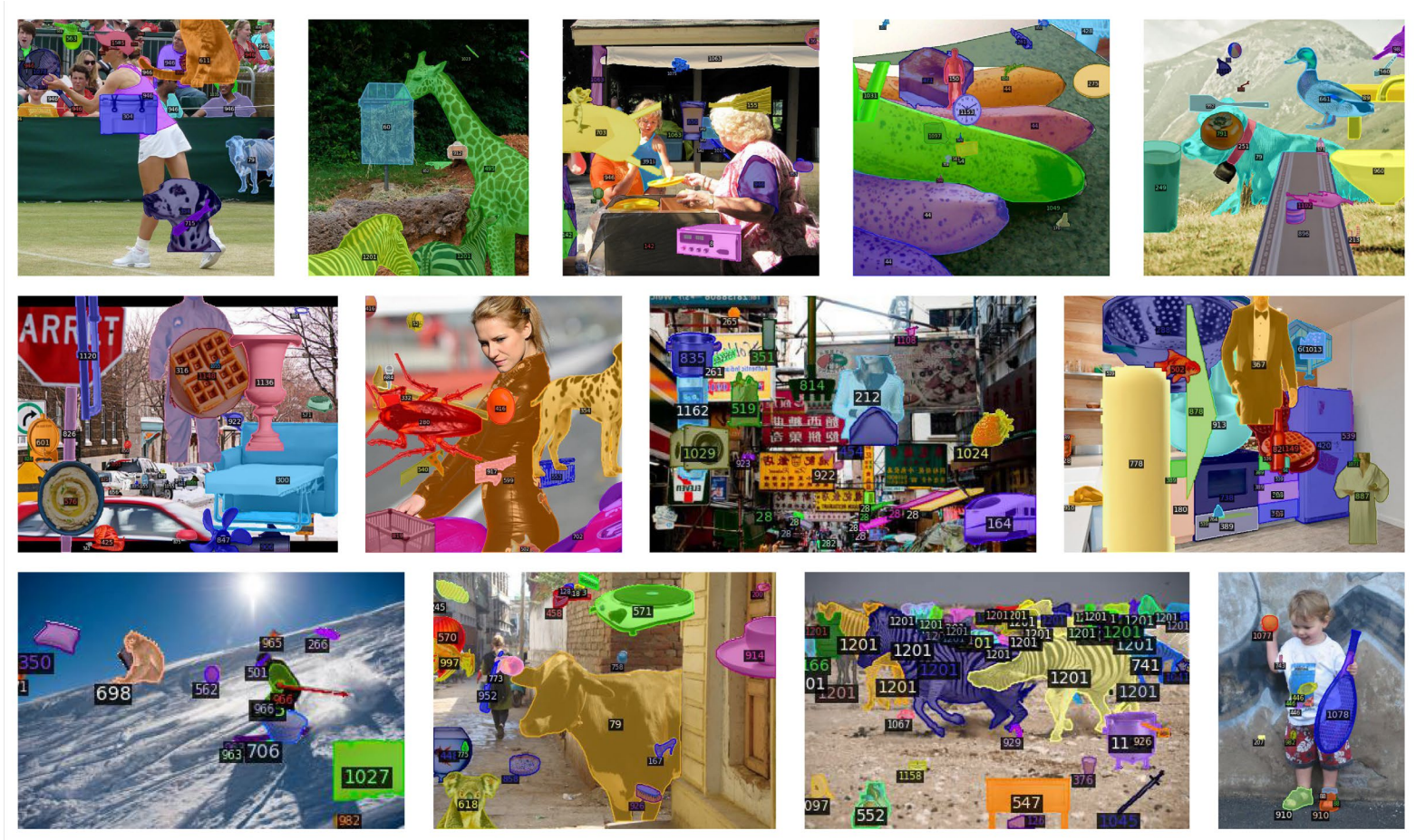
Instance Filtration

DiverGen - Instance Augmentation

use the instance paste strategy to increase model learning efficiency on generative data



Instance Augmentation



samples of augmented data

Experiments - Settings

Datasets

official LVIS training split + 1,200k generative data

official LVIS validation split

Evaluation Metrics

LVIS box average precision and mask average precision

maximum number of detections per image: 300

Implementation Details

baseline: CenterNet2

backbone: Swin-L

training size: 896

batch size: 16

Main Results

- **Data diversity is more important than quantity**

without GDDE: initially improves but then declines, 1,200k is lower than 600k

with GDDE: 1,200k is higher than 600k, 600k is higher than 600k without GDDE

# 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.

Main Results

- **DiverGen significantly outperforms previous methods**

compared to the previous strong model X-Paste, 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

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.**

Ablation Studies - Effect of Category Diversity

- 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.

Ablation Studies - Effect of Prompt Diversity

- 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.

Ablation Studies - Effect of Generative Model Diversity

- 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.

Ablation Studies - Effect of Annotation Strategy

- SAM-bg outperforms max CLIP strategy across all metrics, indicating that our proposed strategy can produce **better annotations**, improving model performance

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.

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

Ablation Studies - Effect of CLIP Inter-similarity

- CLIP inter-similarity can filter low-quality images **more effectively**

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.

Conclusions

- 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.

vivo

Thanks

References

- Gupta, Agrim, Piotr Dollar, and Ross Girshick. "Lvis: A dataset for large vocabulary instance segmentation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.
- Wu, Weijia, et al. "Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.
- Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.
- Kirillov, Alexander, et al. "Segment anything." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.
- Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PMLR, 2021.
- Ghiasi, Golnaz, et al. "Simple copy-paste is a strong data augmentation method for instance segmentation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.
- Tan, Jingru, et al. "1st place solution of lvis challenge 2020: A good box is not a guarantee of a good mask." arXiv preprint arXiv:2009.01559 (2020).
- Zhou, Xingyi, et al. "Detecting twenty-thousand classes using image-level supervision." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022.
- Zhou, Xingyi, Vladlen Koltun, and Philipp Krähenbühl. "Probabilistic two-stage detection." arXiv preprint arXiv:2103.07461 (2021).
- Zhao, Hanqing, et al. "X-Paste: revisiting scalable copy-paste for instance segmentation using CLIP and stablediffusion." International Conference on Machine Learning. PMLR, 2023.
- https://github.com/deep-floyd/IF