X-Paste: Revisiting Scalable Copy-Paste for Instance Segmentation using CLIP and StableDiffusion

Hanqing Zhao [*1] **Dianmo Sheng** [*1] **Jianmin Bao** [2] **Dongdong Chen** [2] **Dong Chen** [2] **Fang Wen** [2] **Lu Yuan** [2] **Ce Liu** [2] **Wenbo Zhou** [1] **Qi Chu** [1] **Weiming Zhang** [1] **Nenghai Yu** [1]

Abstract

Copy-Paste is a simple and effective data augmentation strategy for instance segmentation. By randomly pasting object instances onto new background images, it creates new training data for free and significantly boosts the segmentation performance, especially for rare object categories. Although diverse, high-quality object instances used in Copy-Paste result in more performance gain, previous works utilize object instances either from human-annotated instance segmentation datasets or rendered from 3D object models, and both approaches are too expensive to scale up to obtain good diversity. In this paper, we revisit Copy-Paste at scale with the power of newly emerged zero-shot recognition models (e.g., CLIP) and text2image models (e.g., StableDiffusion). We demonstrate for the first time that using a text2image model to generate images or zero-shot recognition model to filter noisily crawled images for different object categories is a feasible way to make Copy-Paste truly scalable. To make such success happen, we design a data acquisition and processing framework, dubbed "X-Paste", upon which a systematic study is conducted. On the LVIS dataset, X-Paste provides impressive improvements over the strong baseline CenterNet2 with Swin-L as the backbone. Specifically, it archives **+2.6** box AP and **+2.1** mask AP gains on all classes and even more significant gains with **+6.8** box AP **+6.5** mask AP on long-tail classes. Our code and models are available at https://github.com/yoctta/XPaste.

[*]Equal contribution [1]University of Science and Technology of China [2]Microsoft. Correspondence to: Jianmin Bao <jianmin.bao@microsoft.com>, Wenbo Zhou <welbeckz@ustc.edu.cn>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

1. Introduction

Instance segmentation (Dai et al., 2016; He et al., 2017; Hafiz & Bhat, 2020) is a fundamental task in computer vision with very broad applications. In order to get plausible performance for one specific category, most existing methods (Li et al., 2022; Dong et al., 2022a; Liu et al., 2022; He et al., 2017) rely on a large number of images annotated for this category, which is not only expensive but also time-consuming. This also makes expanding the object category coverage extremely hard. In fact, the real-world images often follow a long-tail distribution, so collecting enough images for some rare categories itself is already very difficult. Therefore, it is of great value to study how to augment and create training data in an efficient and scalable way.

As one simple yet effective data augmentation strategy, Copy-Paste (Dwibedi et al., 2017; Ghiasi et al., 2021a; Dvornik et al., 2018) has been extensively studied to improve data efficiency. By randomly pasting object instances onto background images, it can generate a combinatorial number of training data for free and boost the instance segmentation model performance, especially for rare categories. Intuitively, if we can utilize more diverse object instances in Copy-Paste, more performance gain can be achieved. However, the object instances used in existing methods are either from the instance segmentation dataset itself (Ghiasi et al., 2021a; Dvornik et al., 2018) or rendered from external 3D models (Dwibedi et al., 2017). In this paper, we argue that both these two manners are not scalable and have not exploited the full potential of Copy-Paste. More specifically, for the former manner, on one hand, collecting human-annotated data is not scalable. On the other hand, the limited instance number of rare categories also results in limited diversity of Copy-Paste data. In the latter manner, building/collecting 3D models itself is super difficult and not scalable.

In this paper, we propose a new object instance acquisition and processing framework, dubbed "X-Paste". X-Paste is built upon Copy-Paste (Ghiasi et al., 2021a) to train the instance segmentation model but aims to make Copy-Paste more scalable, i.e., *obtain large-scale object instances with high-quality masks ("oxygen of Copy-Paste") for unlimited*

categories in an efficient and automatic way. X-Paste consists of four major modules: Object Instance Acquisition, Instance Mask Generation, Instance Filtering, and Instance Composition with Background Images.

To get large-scale images for different categories in a scalable way, the core idea of the Object Instance Acquisition module is to take full advantage of the newly emerged powerful zero-shot text2image generative model like DALLE (Ramesh et al., 2022) and StableDiffusion(Rombach et al., 2022) (used because of the accessibility), or zero-shot recognition model like CLIP (Radford et al., 2021) and Florence (Yuan et al., 2021), both of which are trained on web-scale image-text pairs. For the text2image generative models, by feeding different text prompts of one specific category, they can generate very diverse images with different appearances, viewpoints and styles. In contrast, previous methods (Dwibedi et al., 2017; Ghiasi et al., 2021a; Dvornik et al., 2018) "Copying" images from training datasets or rendering from limited 3D models suffer worse diversity. For the zero-shot recognition model, even though it cannot generate images like text-to-image model, it enables us to filter high-quality data for any category from large-scale web-retrieved images, which is unachievable for previous close-set (limited category coverage) image recognition models. Moreover, our study shows that these two data acquisition methods can be combined together to get better performance.

Given the generated or retrieved images, the following Instance Mask Generation module is designed to get the instance masks, which are required in Copy-Paste for composition. We observe that the generated or web-retrieved images for different categories are usually object-centric. Especially for generated images, they are often with simple background. So they are relatively easier to be segmented with off-the-shelf salient object segmentation models (Qin et al., 2020; Yun & Lin, 2022; Su et al., 2022). One advantage of these models is that they can estimate the mask for any category, i.e., category-agnostic, making it possible for generating instances under open-vocabulary setting. Note that, in order to get precise segmentation mask for web-retrieved images, we conduct the background analysis and remove the images with complex backgrounds.

After getting the instance masks, X-Paste performs Instance Filtering to remove samples with wrong categories or imprecise mask segmentation. We further utilize the CLIP model to calculate the similarity between the segmented results and the given category, and then remove samples with low similarity. After that, we can have a great number of object instances for different categories. Finally, the Instance Composition module generates the training data to train the segmentation model by pasting the object instances onto different background images. We study various composition strategies and find that assigning a random scale and location for these instances can get significant performance boost. We also find that advanced blending methods like Poisson blending show no more gains.

Our X-Paste presents an important attempt that utilizes the high-quality zero-shot text2image model and recognition model to help image understanding tasks, even though it may look very intuitive and straightforward. Its great scalability can help make large or even open vocabulary instance segmentation happen, by alleviating the pressure of human annotation for instance segmentation. It can act as one "plug and play" component for any instance segmentation framework without any architecture change or inference overhead. To the best of our knowledge, this is the first work showing how to make Copy-Paste at scale and further explore its potential.

We perform extensive experiments to validate the superiority of X-Paste, On COCO dataset, X-Paste achieves 57.0 box mAP and 48.6 mask mAP, outperforming baseline by +1.7 mAP and +0.9 mAP. On LVIS dataset, we achieve 44.4 mask mAP for all objects and 43.3 mask mAP for rare objects. yield +2.1 and +6.5 gains over baseline. Our method can also benefit open-vocabulary setting, our model achieves 31.8 mask mAP for all objects and 21.4 mask mAP for rare objects, +1.6, and +5.0 mAP over baseline. Meanwhile, we perform a comprehensive analysis of our framework X-Paste the demonstrate the effectiveness of each step.

2. Related Works

Instance Detection and Segmentation. Instance detection and segmentation (He et al., 2017; Dai et al., 2021; Meng et al., 2022; Li et al., 2022) has been extensively studied in the past decades. Given an image, instance detection is to locate the exact position of object instance and determine the category of the instance, while instance segmentation further needs to get the fine-grained mask of the instance. Since Mask-RCNN (He et al., 2017), these two tasks are usually integrated into an unified two-stage framework, so we use instance segmentation to denote these two tasks. Recently, long-tail and open-vocabulary instance segmentation are drawing more attention, since they are more realistic settings in real applications.

LVIS dataset (Gupta et al., 2019) is one of the most challenging long-tail instance segmentation dataset. Due to the high imbalance and limited instances, current instance segmentation models often struggle to handle such rare categories well. Therefore, different techniques have been proposed, e.g., data re-sampling and loss re-weighting(Wang et al., 2020; Mahajan et al., 2018; Zang et al., 2021; Tan et al., 2021; Wang et al., 2021), score normalization (Pan et al., 2021), and data augmentation (Ghiasi et al., 2021a). In

this paper, we focus on Copy-Paste based augmentation and study how to make it scalable to further explore its potential. In this sense, our study is complementary to other non-Copy-Paste techniques.

Open-vocabulary instance detection (OVD) aims to detect target/novel class (unseen) objects not present in the training/base class (seen) vocabulary at inference. Therefore, it can be viewed as a hard corner of long-tail instance detection, where no instances exist for some categories. To improve the open-vocabulary performance, some existing methods leverage the large-scale pretrained vision-language models (e.g., CLIP) to transfer the rich knowledge into the classifier, including OVR-CNN (Zareian et al., 2021), ViLD (Gu et al., 2021), OpenSeg (Ghiasi et al., 2021b), RegionCLIP (Zhong et al., 2022), DetPro (Du et al., 2022) and PromptDet (Feng et al., 2022). And the latest work Detic (Zhou et al., 2022) proposes to utilize image classification data (i.e., ImageNet-22k(Deng et al., 2009)) to help train the classifier via image-level supervision. In the following experiments, we will show our X-Paste, as a simple "Plug and Play" augmentation strategy without the need of any model architecture or training strategy change, can already enable the vanilla instance segmentation model plausible open-vocabulary capability. Again, it is complementary to the aforementioned open vocabulary techniques.

Data Boosting for Instance Segmentation. Since most existing instance segmentation models are data hungry, a lot of efforts have been devoted to improve the performance from the data boosting perspective. Based on the boosted data type, they can be roughly divided into synthesis-based and retrieval-based.

For synthesis-based methods, Early works use graphics based renderings(Hinterstoisser et al., 2018; Su et al., 2015; Hodaň et al., 2019) or computer games(Richter et al., 2016; 2017) to generate high quality labelled data. However, these methods often suffer from the huge domain gap between real and pure synthetic data. Then composing real images are introduced in following works, including Copy-Paste (Dwibedi et al., 2017), Contextual Copy-Paste (Dvornik et al., 2018), and Instaboost (Fang et al., 2019). And Simple Copy-Paste (Ghiasi et al., 2021a) further indicates that simply pasting real segmented objects randomly onto background images already works very well, without the need of advanced strategies like contextual modeling. The main problem of existing Copy-Paste methods is that they are not scalable and suffer from limited instance diversity, since they get the object instances either from the instance dataset itself or rendered from 3D models.

Retrieval is another way to collect large-scale data. with the development of large-scale image-text dataset LAION(Schuhmann et al., 2022) and image classification dataset ImageNet-22k(Deng et al., 2009), some approaches (Feng et al., 2022; Zhou et al., 2022; Hong et al., 2017; Jin et al., 2017; Shen et al., 2018; Wei et al., 2016) study how to leverage such data to help instance segmentation. But since these data themselves do not have the precise instance annotation, some dedicated designs are needed, e.g., pseudo labeling (Zhong et al., 2022) or freezing the location branch and only train the classifier with the image-level supervision in (Zhou et al., 2022).

The key difference between our X-Paste and the above efforts is that we try to utilize the text2image model synthesized data and retrieval data under the same simple Copy-Paste framework and make it scalable, without the need of any algorithm change. Moreover, it is orthogonal to the efforts like leveraging the pretrained CLIP model for knowledge distillation or image-level supervision.

Zero-shot Recognition and Text2Image Generation. Vision-Language Pre-training (VLP) has recently made very encouraging breakthrough. By conducting contrastive learning upon web-scale image-text data, the representative works, including CLIP(Radford et al., 2021), MaskCLIP (Dong et al., 2022b), Florence (Yuan et al., 2021), ALIGN(Jia et al., 2021) and OmniVL(Wang et al., 2022), have shown great power in aligning the visual feature with textual feature, enabling the real zero-shot recognition rather than traditional close-set recognition. Also pretraining on the web-scale image-text data, another breakthrough is about zero-shot text2image generative models. Representative works include DALL-E (Ramesh et al., 2021; Nichol et al., 2021; Ramesh et al., 2022), CogView (Ding et al., 2021), Parti (Yu et al., 2022), Imagen (Saharia et al., 2022), VQ-Diffusion (Gu et al., 2022), StableDiffusion(Rombach et al., 2022) and Frido(Fan et al., 2023). By feeding free-form text, these models can generate high-fidelity images that match the text input condition. Intuitively, such zero-shot text2image models are natural data generators, but very few prior works have used them to help image understanding training. In this paper, we present the first attempt that utilizes the powerful zero-shot recognition and text2image models to revisit Copy-Paste at scale and help instance segmentation.

3. X-Paste

As described above, previous Copy-Paste methods (Dwibedi et al., 2017; Ghiasi et al., 2021a; Dvornik et al., 2018) utilize instances from internal instance segmentation dataset itself or rendered from 3D CG models, making them hard to scale up and exploit the full potentials of Copy-Paste. In this paper, we study how to make Copy-Paste scalable by acquiring external object instances in an automatic and efficient way. We propose to leverage generated images from zero-shot text2image model(i.e., StableDiffusion (Rombach et al., 2022)) or web-retrieved images filtered by zero-shot image

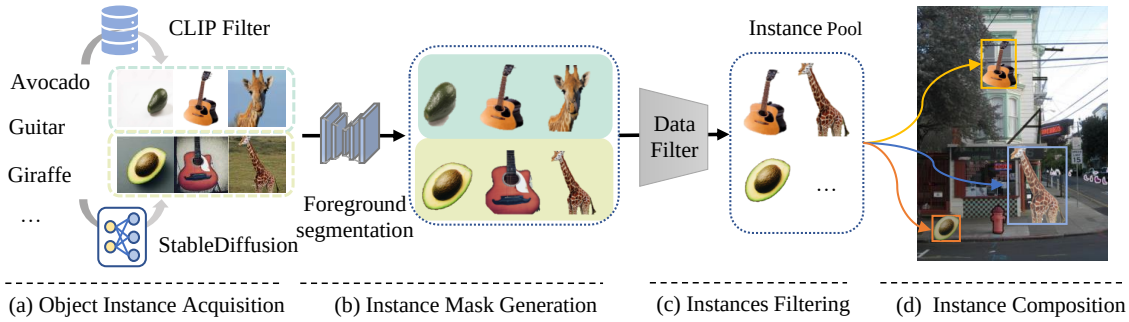

Figure 1. Overview of the X-Paste pipeline. X-Paste first obtains filtered real images from CLIP and generated images from StableDiffusion. Then Instance Mask Generation module gets the pseudo mask for all images. After that, the Instance Filtering module removes these samples with inaccurate mask segmentation. Finally, we paste these resulting instances into a background image for training.

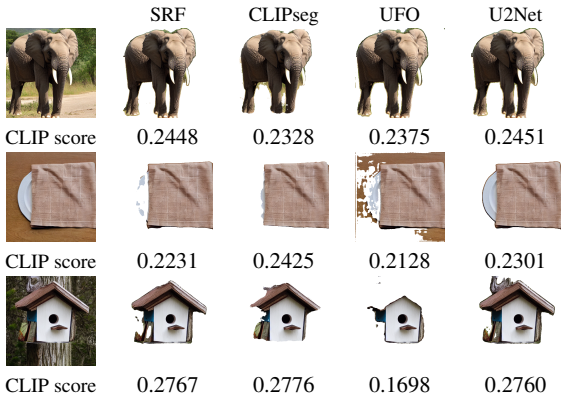

Figure 2. Predicted foregrounds of four segmentation methods and their CLIP score with the given text.

classification model(e.g.,, CLIP (Radford et al., 2021)). In order to use such images under the same simple Copy-Paste framework, we propose a complete data acquisition and processing framework "X-Paste". As shown in Figure 1, the overall framework is very simple and straightforward, consisting of four major modules: Object Instance Acquisition, Instance Mask Generation, Instance Filtering, and Instance Composition. In the following sections, we will elaborate each module in detail.

3.1. Object Instance Acquisition

To collect images of a specific category with diverse appearances, viewpoints, and styles, we provide two efficient and effective solutions based on recently emerged large-scale vision-language models (i.e., text2image model and image-text contrastive model), both of which show strong zero-shot capability and make data acquisition for any category scalable. In detail, as the first solution, we propose to directly apply the powerful zero-shot text2image generative model (Rombach et al., 2022) to generate images for the interested categories. We obtain images of a specific category by feeding a text prompt like "a photo of a single [*Category Name*]". The word "single" here is to encourage the model to generate a single instance in the image since the generative model may generate multiple instances for many categories. We find that these generated images are photorealistic, super diverse, and semantically match with the given category, as such text2image models are trained with web-scale data and show great combinatorial capability. In this paper, we use StableDiffusion V1.4 with PLMS sampler by default, as it is publicly open-sourced.

As the second scalable solution, we propose to crawl real images from the internet as the object instance source. Since web-crawled images are often very noisy, directly using all of them without filtering will destroy the model. Before the emergence of latest zero-shot image-text contrastive models, filtering images for different categories is not a trivial task, since most recognition models are trained on close-set categories. To make such recognition models perform well for any category, large-scale category-specific data needs to be collected, thus making it a chicken-egg problem. By conducting contrastive learning on web-scale image-text data, the latest vision-language alignment models like CLIP (Radford et al., 2021), Florence (Yuan et al., 2021) make zero-shot image-text similarity matching possible. Therefore, we apply the CLIP model to calculate the semantic similarity between each category and the corresponding crawled images, and only keep the images with high semantic similarity. Moreover, to help the following Instance Mask Generation module generate high-quality instance masks, we further conduct the background analysis and only keep the images with simple backgrounds. In detail, we simply calculate the color histogram of each image and find the dominant color. Then we only keep the images in which over 40% image pixels are close to the dominant color (color difference smaller than 5).

3.2. Instance Mask Generation

Since Copy-Paste needs the instance mask during composition, given the generated or crawled images, another chal-

lenge is to get the precise object instance masks. By going through the acquired images, we observe most of them are object-centric. This motivates us to use the class-agnostic foreground segmentation models to generate the instance mask. In detail, we have investigated three types of foreground segmentation algorithms: salient object segmentation, co-saliency segmentation and text-guided segmentation. We use prevalent methods U2Net (Qin et al., 2020) and SelfReformer (abbreviated as SRF) (Yun & Lin, 2022) for salient object segmentation, UFO (Su et al., 2022) for co-saliency segmentation with a batch of images from each category, and CLIPseg (Lüddecke & Ecker, 2022) to segment the object with *Category Name* as the text prompt. Since CLIPseg (Lüddecke & Ecker, 2022) produces very coarse segment masks, we convert the generated masks into the tri-map and use an image matting method (Park et al., 2022) to further refine the mask.

We observe all these segmentation methods will fail in some cases, as shown in Figure 2. So we propose a CLIP-guided selection strategy to select one instance mask as the pseudo annotation from these four methods for each image. Specifically, we calculate the cross-modal similarity between the image of segmented object in blank background and its category name. Our motivation is that better segmentation will get higher semantic similarity. To illustrate this motivation, we visualize the CLIP score in Figure 2.

3.3. Instance Filtering

To further remove object instances with low-quality masks, we apply several strategies in the Instance Filtering module. First, we filter out masks with areas less than 5% or over 95% of the whole image because these instances are highly possibly segmented incorrectly. Second, we select the instances with high semantic relevance to the category with pre-computed CLIP scores. Considering the different sensitivity of CLIP score for each category, we set the category-specific threshold $thres_i = min(t, max(\mathbf{C}_i) - d)$, where t is the predefined CLIP score threshold, d is a subtractive threshold for those classes with low image-text similarity, default as 0.01, $\mathbf{C}_i$ is the CLIP score set of instances from category i. Since Copy-Paste is especially helpful for rare categories in a long-tail dataset, it is more important to generate enough diverse instances of rare categories. Therefore, we are very curious whether generative models can generate high-quality instances for such rare categories. To verify this, we visualize the distribution of CLIP scores of generated images in Figure 3 for different categories in the LVIS dataset. We find that StableDiffusion achieves similar CLIP score distribution in generating rare categories, showing it does not suffer serious data imbalance issues and performs similarly well for rare categories.

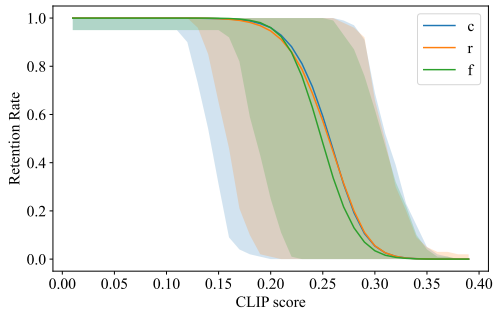

Figure 3. We show the retention rate of instances generated by StableDiffusion filter varying with CLIP score threshold. 'r', 'c', and 'f' indicates "rare,"common" and "frequent" categories in LVIS dataset, the shadow areas represent the upper and lower bound of per-category CLIP score distribution, the lines represent entire CLIP score distribution.

3.4. Instance Composition

When applying X-Paste to compose training images, we sample instances with a class-balanced sampling strategy and paste them at random locations of the background images. The instances are resized to proper scales based on the resolution of the background image and a scale factor (percentage of object area in the background image) depending on data distribution. The background images can be either images with annotations or plain background images, we only use the training data for versatility. When the instance occludes an object in the background image, we remove fully occluded objects and update mask and bounding box annotations accordingly.

Class-balanced Sampling: We randomly choose the number of instances $N_i \in [1, N_{max}]$ ($N_{max} = 20$ by default) for each training image, then we sample N_i categories and one instance for each category in a repeatable way.

Instance Composition: We calculate the mean μ_C and standard variance σ_C of object scale (square root of mask-area-divided-by-image-area) for each class C in training set. When pasting instance I with class C, we sample scale S_r from Gaussian distribution $N(\mu_C, \sigma_C^2)$ and paste the instance with scale $S_r^2 HW$ on background image (H, W indicate the image height and width), where the coordinates of the center point of the bounding box are randomly selected within the whole image. In Figure 4, we show some training samples generated by X-Paste. These generated images are different from synthesized images with the previous methods (Fang et al., 2019; Dvornik et al., 2018), which aims to generate realistic images for training. Our generated images are not real since the scale and relative relationship of the objects may not be reasonable. Still, these generated images can be used as good training data for improving performance.

Figure 4. Visualization of training samples synthesized by X-Paste.

4. Experiments

4.1. Settings

Datasets. We conduct experiments of object detection and instance segmentation on LVIS (Gupta et al., 2019) and MS-COCO (Lin et al., 2014) datasets. LVIS dataset contains 100k training images, and 20k validation images. It has 1203 categories with a long-tailed distribution of instances in each category. These categories can be divided into common, frequent, and rare categories according to the instances number of each category. And the number of the common, frequent, and rare categories is 461, 405, and 337. In the open-vocabulary instance detection setting, common and frequent categories are for training and rare categories serve as the novel categories for testing. MS-COCO dataset contains 118K training, 5K validation, and 20K test-dev images. We use the official split for training.

Learning Framework. We utilize the typical object detection and instance segmentation framework CenterNet2 (Zhou et al., 2021) in Detectron2 (Wu et al., 2019). Our basic setting uses Resnet 50 (He et al., 2016) pretrained on ImageNet-22k as backbone. The training configurations are set as follows: training resolution is set to 640, the batch size is 32, and 4× schedule (48 epochs). We test the performance with a single-scale strategy. We report the box AP and mask AP on all categories(donated as AP^{box} and AP^{mask}) as well as on the rate categories(donated as AP_r^{box} and AP_r^{mask}).

Baseline Settings of X-Paste. We design a baseline setting to ablate X-Paste. To make sure each category contains enough instances, we use the CLIP model (Radford et al., 2021) to filter 1k real images per category. Besides, we apply the StableDiffusion model (Rombach et al., 2022) to generate 1k images for each category. For StableDiffusion, the diffusion steps is set to 200 with the classifier-free scale set to 5.0. We adopt the methods(donated as *max CLIP*) described in Sec 3.2 as the default setting of instance mask generation. For Instance Filtering, we set the CLIP threshold as 0.21 to filter all the obtained instances. We keep 150k generated instances from StableDiffusion and 150k real instances from the CLIP model. During the Instance Composition module, the number of instances pasted to each background image is set to 20 for training.

# Gen	# Real	AP^{box}	AP^{mask}	AP_r^{box}	AP_r^{mask}
0	0	34.5	30.8	24.0	21.6
100k	100k	36.3	32.3	28.0	25.2
150k	150k	36.6	32.7	28.5	26.5
300k	300k	**36.7**	**32.9**	**29.9**	**27.6**

Table 1. Results of using different scales of filtered instances in X-Paste. # Gen and # Real denotes the number of generated images from StableDiffusion and real images filtered by the CLIP model.

# Gen	# Real	AP^{box}	AP^{mask}	AP_r^{box}	AP_r^{mask}
300k	0	35.3	31.6	26.3	24.5
0	300k	36.1	32.0	27.8	25.0
150k	150k	**36.6**	**32.7**	**28.5**	**26.5**

Table 2. Comparison of using different instances sources in X-Paste. # Gen and # Real denotes the number of generated images from StableDiffusion and real images filtered by the CLIP model.

4.2. Analysis of X-Paste

We perform ablation studies to analyze our proposed X-Paste and find several intriguing properties.

More instances, better results. In Table 1, we compare the results without using and using different numbers of instances in X-Paste. We find that using only 100k generated instances or filtered real instances from X-Paste can boost performance significantly. Concretely, it achieves 36.3 box AP and 32.3 mask AP, surpassing baseline by +1.9 mAP and +1.5 mAP. It achieves a more significant gain on rare categories with a +4.0 box AP and +3.6 mask AP. More importantly, if we keep increasing the scale of instances from 100k to 300k, we can obtain better results. This also demonstrates the high value of using text-to-image generated images, as retrieving a large quantity of real object-centric images from search engine is very challenging. To further show this point, we use all the one million retrieved images on LVIS dataset, and compare the results with the baseline of using one million retrieved images + one million generated images in table 3. The results show that, even with a large amount of retrieved images, adding text-to-image generated images can still improve $0.7 \sim 0.9$ on mask AP.

Diverse instance sources matter. We present the results of using different sources of instances for X-Paste in Table 2. Under the same instances scale, Using CLIP filtered real images could achieve better results. Moreover, we observe

Backbone	Data	AP^{box}	AP^{mask}	AP_r^{box}	AP_r^{mask}
ResNet50	Retr	36.5	32.3	28.3	25.2
ResNet50	Retr+Gen	36.7	33.0	29.6	27.8
Swin-L	Retr	48.9	43.5	47.6	41.6
Swin-L	Retr+Gen	50.1	44.4	48.2	43.3

Table 3. Comparison with directly use generated images as training data.

Foreground segmentation	AP^{box}	AP^{mask}	AP_r^{box}	AP_r^{mask}
U2Net (Qin et al., 2020)	35.6	31.7	27.1	24.8
CLIPseg (Lüddecke & Ecker, 2022)	35.4	31.6	25.5	23.1
selfreformer (Yun & Lin, 2022)	35.5	31.7	26.6	24.1
UFO (Su et al., 2022)	35.7	31.7	**27.5**	**25.0**
max CLIP	**35.8**	**31.8**	26.2	23.7

Table 4. Comparison of using different segmentation methods.

that using both instances from the generative model and filtered real images can achieve a better performance than using single instance sources. This validates that the diversity of these instances is of great importance for boosting performance.

Effect of instance mask generation. An accurate instance mask serves an important role in training object detection and instance segmentation. Since an imprecise mask may lead the model to learn wrong knowledge. We study several category-agnostic segmentation methods here. Concretely, we experiment with the instance mask generated by U2Net(Qin et al., 2020) and SelfReformer(Yun & Lin, 2022), UFO(Su et al., 2022), CLIPseg (Lüddecke & Ecker, 2022) our proposed method to select the mask segmentation with the max CLIP score. Table 4 reports the results of using different instance mask generation methods. We find that these segmentation methods can achieve comparable results. Our proposed max CLIP could yield slightly better results in AP_r^{box}, AP_r^{mask}, AP^{box}, and AP^{mask}.

Effect of using different CLIP score thresholds for instance filtering. With different CLIP score thresholds, we choose a fixed number of instances for training the framework. We report the results in Table 5. we find 0.21 is the best threshold, comparing non-filtering (threshold set to 0), it provides a gain of 0.6 mask AP for all classes and 1.4 mask AP for rare classes. We also observe that a higher CLIP score threshold may not produce better performance. We believe the reason is that higher CLIP scores mean the generated samples are more correlated with given texts but they may show limited diversity in terms of appearance, viewpoint, and style. The lack of diversity will cause worse performance. This proves again that both quality and diversity of instances are of great importance for training instance segmentation.

Study different instances placement strategies. We compare our random placement strategy with a reference-based placement strategy. Reference-based placement strategy pastes the generated instances to the location of the original

CLIP score threshold	AP^{box}	AP^{mask}	AP_r^{box}	AP_r^{mask}
0	36.0	32.1	27.6	25.1
0.2	36.4	32.5	**29.0**	26.1
0.21	**36.6**	**32.7**	28.5	26.5
0.22	36.4	32.4	28.3	25.7
0.23	36.3	32.4	28.8	**26.7**
0.24	36.3	32.5	28.1	25.8
0.25	36.3	32.4	28.4	25.8

Table 5. Ablation of CLIP score thresholds for instance filtering.

Placement	N_{max}	AP^{box}	AP^{mask}	AP_r^{box}	AP_r^{mask}
reference	-	33.7	30.1	22.9	21.2
random	10	36.2	32.2	28.3	25.1
random	20	**36.6**	**32.7**	**28.5**	**26.5**
random	30	35.5	31.3	26.8	24,6

Table 6. Ablation of different placement strategies.

bounding box in the background image. We also study how the maximum number of instances N_{max} pasted a single background image affects the performance. Table 6 reports the results. We observe random placement strategy achieves a better performance than reference-based methods. Besides, we find that 20 is a suitable number for max number of instances pasted to a background image.

4.3. Comparison with the previous methods

Comparison with other data-augmentation methods. In table 7, we make a fair comparison with previous data-augmentation related methods. We use the same framework CenterNet2 (Zhou et al., 2021) with Resnet50 as the backbone. We compare our method with Copy-Paste (Ghiasi et al., 2021a) and Detic(Zhou et al., 2022) on LVIS dataset. Table 7 reports the results on LVIS dataset. We train Detic with the same instances as X-Paste to investigate whether X-Paste is more efficient than weakly-supervised learning with image-level annotations. Compared with the strong baseline Copy-Paste (Ghiasi et al., 2021a), our method achieves significantly better results. In all categories, X-Paste outperforms Copy-Paste by +1.2 box AP and +1.2 mask AP. In the rare categories, X-Paste achieves a larger gain with +3.2 box AP and +4.4 mask AP. This validates the effectiveness of a large scale of instances is crucial for achieving better results. Our X-Paste also outperforms the recently proposed Detic (Zhou et al., 2022). Compared with Detic, X-Paste does not involve a complicated mechanism and not requires a larger batch size. This validates the effectiveness of the proposed X-Paste.

To verify whether X-Paste is suitable for large models. We further conduct experiments with the recently proposed large model Swin-Large (Liu et al., 2022)(Swin-L for short) as the backbone. We increase the input resolution to 896 and train the model for 4 $\times$ schedule (48 epochs) with exponen-

Method	AP^{box}	AP^{mask}	AP_r^{box}	AP_r^{mask}
baseline	34.5	30.8	24.0	21.6
Copy-Paste(Ghiasi et al., 2021a)	35.4	31.5	25.3	22.1
Detic(Zhou et al., 2022)	35.3	31.7	27.5	25.4
X-Paste	**36.6**	**32.7**	**28.5**	**26.5**

Table 7. Comparison with related methods on LIVS dataset. We use the same network and training settings here. We use 150k generated images and 150k retrieved images to boost object detection and instance segmentation with X-Paste and Detic.

tial moving average to update the model. We use about 1 million generated and retrieved instances filter with the CLIP score threshold set to 0.21 for X-Paste. Table 8 lists the results. We observe that X-Paste outperforms Copy-Paste by +0.6 box AP and +0.7 mask AP in all categories, and +4.8 box AP and +4.8 mask AP in rare categories. X-Paste can provide more improvements on rare categories since X-Paste offers many instances for rare categories while Copy-Paste can not. This validates that X-Paste could achieve consistent gain over Copy-Paste on top of solid baselines. More importantly, we can combine X-Paste with Copy-Paste to have more instances for pasting to background images for training. We observe a more strong performance. This further validates that the diversity of instances is important.

Open-Vocabulary Object Detection. We follow the setting of previous work and remove the images containing rare categories from the training set. Then we train the framework on the images which contain common and frequent categories. We adopt CLIP classifier(Gu et al., 2021) and initialize our model with the Box-Supervised baseline from Detic (Zhou et al., 2022). In Table 9, we evaluate the open vocabulary detection performance of different methods on LVIS dataset. Compared with CLIP based methods like ViLD (Gu et al., 2021), RegionCLIP (Zhong et al., 2022), PromptDet (Feng et al., 2022), and DetPro (Du et al., 2022). X-Paste achieves a significantly better performance. This shows that, without any architecture or algorithm change, the simple Copy-Paste by using generated or filtered novel instances is already an effective and strong baseline for open-vocabulary object detection. We also notice that previous methods like Detic (Zhou et al., 2022), MEDet (Chen et al., 2022) and Centirc-OVD (Rasheed et al., 2022) achieve better performance, they use external human-annotated data like ImageNet-22k and need to make big architecture or algorithm change accordingly to consume such data. It is worthy to note that, as an simple data augmentation plugin, our X-Paste is orthogonal and complementary with these dedicated designed methods.

Performance on COCO dataset. We also conduct experiments on the COCO dataset with our proposed X-Paste. The COCO dataset is more challenging since it has a large number of instances for each category, which does suffer from the long-tailed distribution issue. We apply the StableDiffusion model with the same setting as on LVIS to generate 5k images for each category. For Instance Filtering, we set the CLIP threshold as 0.21 to filter all the obtained instances. We keep 300k generated instances from StableDiffusion and 80k real instances from the CLIP model. During the Instance Composition module, we paste 20 instances to each background image for training as well. We conduct experiments on the CenterNet2 framework with two backbones: ResNet50 and Swin-L. We report the results Table 10. With ResNet50 backbone, X-Paste outperforms the baseline by 2.5 box AP and 1.8 mask AP. With Swin-L backbone, X-Paste achieves 57.0 box AP and 48.5 mask AP, surpassing baseline by +1.7 box AP and 0.9 mask AP. We can find that our methods could consistently improve the baseline in various settings on the COCO dataset. By combining X-Paste with Copy-Paste, we observe a more strong performance as well. Even for the COCO dataset with multiple instances in each category, X-Paste is still effective.

5. Conclusion

This paper revisits the previous Copy-Paste data augmentation methods and finds it is effective while showing limited scale-up capability. To this end, we propose a scalable version of Copy-Paste called X-Paste. X-Paste leverages the zero-shot recognition models(e.g.,CLIP) and text2image models(e.g.,StableDiffusion) to obtain large-scale images with accurate categories. Then these images are further transformed into instances through a series of carefully designed modules. Finally, these instances can be used in data augmentation for training instance segmentation. X-Paste provides significant improvements on top of strong baselines on LVIS and COCO datasets. We hope that X-Paste will foster further research on utilizing zero-shot recognition or generative models for various vision tasks.

6. Acknowledgements

This work was partially supported in part by the Fundamental Research Funds for the Central Universities under Grant WK5290000003, the Natural Science Foundation of China under Grant U20B2047, 62121002, 62072421, 62002334 and Key Research and Development program of Anhui Province under Grant 2022k07020008.

References

Chen, P., Sheng, K., Zhang, M., Shen, Y., Li, K., and Shen, C. Open vocabulary object detection with proposal mining and prediction equalization. *arXiv preprint arXiv:2206.11134*, 2022. 8, 9

Dai, J., He, K., and Sun, J. Instance-aware semantic segmentation via multi-task network cascades. In *Proceedings of the IEEE conference on computer vision and pattern*

Method	Backbone	AP^{box}	AP^{mask}	AP_r^{box}	AP_r^{mask}
Copy-Paste(Ghiasi et al., 2021a)	EfficientNet-B7	41.6	38.1	-	32.1
Tan et al.(Tan et al., 2020)	ResNeSt-269	-	41.5	-	30.0
Detic(Zhou et al., 2022)	Swin-B	46.9	41.7	45.9	41.7
CenterNet2 (Zhou et al., 2021)	Swin-L	47.5	42.3	41.4	36.8
w/ Copy-Paste(Ghiasi et al., 2021a)	Swin-L	49.5	43.7	43.4	38.5
w/ X-Paste	Swin-L	(+2.6) **50.1**	(+2.1) **44.4**	(+6.8) **48.2**	(+6.5) **43.3**
w/ X-Paste + Copy-Paste	Swin-L	(+3.4) **50.9**	(+3.1) **45.4**	(+7.3) **48.7**	(+7.0) **43.8**

Table 8. Comparision with previous methods on LVIS 1.0 validation set.

Method	AP^{box}	AP^{mask}	AP_{novel}^{box}	AP_{novel}^{mask}
ViLD(Gu et al., 2021)	27.5	25.2	17.4	16.8
RegionCLIP(Zhong et al., 2022)	-	28.2	-	17.1
PromptDet(Feng et al., 2022)	-	25.5	-	21.7
DetPro(Du et al., 2022)	28.4	25.9	20.8	19.8
Box-Supervised(Zhou et al., 2022)	33.8	30.2	17.6	16.4
X-Paste	35.7	31.8	22.8	21.4
Detic*(Zhou et al., 2022)	36.3	32.4	26.7	24.9
MEDet*(Chen et al., 2022)	-	34.4	-	22.4
Centric-OVD*(Rasheed et al., 2022)	-	32.9	-	25.2

Table 9. Comparison of open-vocabulary detection performance on LVIS, * means they use external supervised data for training.

Method	Backbone	AP^{box}	AP^{mask}
CenterNet2 (Zhou et al., 2021)	ResNet50	46.0	39.8
w/ Copy-Paste(Dwibedi et al., 2017)	ResNet50	46.4	39.8
w/ X-Paste	ResNet50	(+0.6) 46.6	(+0.1) 39.9
w/ X-Paste + Copy-Paste	ResNet50	(+0.8) 46.8	(+0.2) 40.0
CenterNet2 (Zhou et al., 2021)	Swin-L	55.3	47.7
w/ Copy-Paste(Dwibedi et al., 2017)	Swin-L	56.1	48.3
w/ X-Paste	Swin-L	(+1.3) 56.6	(+0.9) 48.6
w/ X-Paste + retrieval	Swin-L	(+1.5) 56.8	(+1.0) 48.7
w/ X-Paste + Copy-Paste	Swin-L	(+1.6) 56.9	(+1.1) 48.8

Table 10. X-paste works well across a variety of different model architectures on the COCO dataset.

recognition, pp. 3150–3158, 2016. 1

Dai, X., Chen, Y., Xiao, B., Chen, D., Liu, M., Yuan, L., and Zhang, L. Dynamic head: Unifying object detection heads with attentions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7373–7382, 2021. 2

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009. 3

Ding, M., Yang, Z., Hong, W., Zheng, W., Zhou, C., Yin, D., Lin, J., Zou, X., Shao, Z., Yang, H., et al. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34: 19822–19835, 2021. 3

Dong, X., Bao, J., Chen, D., Zhang, W., Yu, N., Yuan, L., Chen, D., and Guo, B. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12124–12134, 2022a. 1

Dong, X., Zheng, Y., Bao, J., Zhang, T., Chen, D., Yang, H., Zeng, M., Zhang, W., Yuan, L., Chen, D., et al. Maskclip: Masked self-distillation advances contrastive language-image pretraining. *arXiv preprint arXiv:2208.12262*, 2022b. 3

Du, Y., Wei, F., Zhang, Z., Shi, M., Gao, Y., and Li, G. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14084–14093, 2022. 3, 8, 9

Dvornik, N., Mairal, J., and Schmid, C. Modeling visual context is key to augmenting object detection datasets. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 364–380, 2018. 1, 2, 3, 5

Dwibedi, D., Misra, I., and Hebert, M. Cut, paste and learn: Surprisingly easy synthesis for instance detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 1301–1310, 2017. 1, 2, 3, 9

Fan, W.-C., Chen, Y.-C., Chen, D., Cheng, Y., Yuan, L., and Wang, Y.-C. F. Frido: Feature pyramid diffusion for complex scene image synthesis. In *Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI)*, 2023. 3

Fang, H.-S., Sun, J., Wang, R., Gou, M., Li, Y.-L., and Lu, C. Instaboost: Boosting instance segmentation via probability map guided copy-pasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 682–691, 2019. 3, 5

Feng, C., Zhong, Y., Jie, Z., Chu, X., Ren, H., Wei, X., Xie, W., and Ma, L. Promptdet: Expand your detector vocabulary with uncurated images. *arXiv preprint arXiv:2203.16513*, 2022. 3, 8, 9

Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T.-Y., Cubuk, E. D., Le, Q. V., and Zoph, B. 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, pp. 2918–2928, 2021a. 1, 2, 3, 7, 8, 9

Ghiasi, G., Gu, X., Cui, Y., and Lin, T.-Y. Open-vocabulary image segmentation. *arXiv preprint arXiv:2112.12143*, 2021b. 3

Gu, S., Chen, D., Bao, J., Wen, F., Zhang, B., Chen, D., Yuan, L., and Guo, B. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10696–10706, 2022. 3

Gu, X., Lin, T.-Y., Kuo, W., and Cui, Y. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 3, 8, 9

Gupta, A., Dollar, P., and Girshick, R. LVIS: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2, 6

Hafiz, A. M. and Bhat, G. M. A survey on instance segmentation: state of the art. *International journal of multimedia information retrieval*, 9(3):171–189, 2020. 1

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016. 6

He, K., Gkioxari, G., Dollár, P., and Girshick, R. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017. 1, 2

Hinterstoisser, S., Lepetit, V., Wohlhart, P., and Konolige, K. On pre-trained image features and synthetic images for deep learning. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pp. 0–0, 2018. 3

Hodaň, T., Vineet, V., Gal, R., Shalev, E., Hanzelka, J., Connell, T., Urbina, P., Sinha, S. N., and Guenter, B. Photorealistic image synthesis for object instance detection. In *2019 IEEE international conference on image processing (ICIP)*, pp. 66–70. IEEE, 2019. 3

Hong, S., Yeo, D., Kwak, S., Lee, H., and Han, B. Weakly supervised semantic segmentation using web-crawled videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7322–7330, 2017. 3

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pp. 4904–4916. PMLR, 2021. 3

Jin, B., Ortiz Segovia, M. V., and Susstrunk, S. Webly supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3626–3635, 2017. 3

Li, F., Zhang, H., Liu, S., Zhang, L., Ni, L. M., Shum, H.-Y., et al. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. *arXiv preprint arXiv:2206.02777*, 2022. 1, 2

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014. 6

Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12009–12019, 2022. 1, 7

Lüddecke, T. and Ecker, A. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7086–7096, 2022. 5, 7

Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., and Van Der Maaten, L. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 181–196, 2018. 2

Meng, L., Dai, X., Chen, Y., Zhang, P., Chen, D., Liu, M., Wang, J., Wu, Z., Yuan, L., and Jiang, Y.-G. Detection hub: Unifying object detection datasets via query adaptation on language embedding. *arXiv preprint arXiv:2206.03484*, 2022. 2

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 3

Pan, T.-Y., Zhang, C., Li, Y., Hu, H., Xuan, D., Changpinyo, S., Gong, B., and Chao, W.-L. On model calibration for long-tailed object detection and instance segmentation. *Advances in Neural Information Processing Systems*, 34: 2529–2542, 2021. 2

Park, G., Son, S., Yoo, J., Kim, S., and Kwak, N. Matteformer: Transformer-based image matting via prior-tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11696–11706, 2022. 5

Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O. R., and Jagersand, M. U2-net: Going deeper with nested u-structure for salient object detection. *Pattern recognition*, 106:107404, 2020. 2, 5, 7

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021. 2, 3, 4, 6

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pp. 8821–8831. PMLR, 2021. 3

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2, 3

Rasheed, H., Maaz, M., Khattak, M. U., Khan, S., and Khan, F. S. Bridging the gap between object and image-level representations for open-vocabulary detection. *arXiv preprint arXiv:2207.03482*, 2022. 8, 9

Richter, S. R., Vineet, V., Roth, S., and Koltun, V. Playing for data: Ground truth from computer games. In *European conference on computer vision*, pp. 102–118. Springer, 2016. 3

Richter, S. R., Hayder, Z., and Koltun, V. Playing for benchmarks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2213–2222, 2017. 3

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022. 2, 3, 4, 6

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 3

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022. 3

Shen, T., Lin, G., Shen, C., and Reid, I. Bootstrapping the performance of webly supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1363–1371, 2018. 3

Su, H., Qi, C. R., Li, Y., and Guibas, L. J. Render for cnn: Viewpoint estimation in images using cnns trained with rendered 3d model views. In *Proceedings of the IEEE international conference on computer vision*, pp. 2686–2694, 2015. 3

Su, Y., Deng, J., Sun, R., Lin, G., and Wu, Q. A unified transformer framework for group-based segmentation: Co-segmentation, co-saliency detection and video salient object detection. *arXiv preprint arXiv:2203.04708*, 2022. 2, 5, 7

Tan, J., Zhang, G., Deng, H., Wang, C., Lu, L., Li, Q., and Dai, J. 1st place solution of lvis challenge 2020: A good box is not a guarantee of a good mask. *ArXiv*, abs/2009.01559, 2020. 9

Tan, J., Lu, X., Zhang, G., Yin, C., and Li, Q. Equalization loss v2: A new gradient balance approach for long-tailed object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1685–1694, 2021. 2

Wang, J., Zhang, W., Zang, Y., Cao, Y., Pang, J., Gong, T., Chen, K., Liu, Z., Loy, C. C., and Lin, D. Seesaw loss for long-tailed instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9695–9704, 2021. 2

Wang, J., Chen, D., Wu, Z., Luo, C., Zhou, L., Zhao, Y., Xie, Y., Liu, C., Jiang, Y.-G., and Yuan, L. Omnivl: One foundation model for image-language and video-language tasks. In *Thirty-sixth Conference on Neural Information Processing Systems (NeurIPS 2022)*, 2022. 3

Wang, T., Li, Y., Kang, B., Li, J., Liew, J., Tang, S., Hoi, S., and Feng, J. The devil is in classification: A simple framework for long-tail instance segmentation. In *European conference on computer vision*, pp. 728–744. Springer, 2020. 2

Wei, Y., Liang, X., Chen, Y., Shen, X., Cheng, M.-M., Feng, J., Zhao, Y., and Yan, S. Stc: A simple to complex framework for weakly-supervised semantic segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2314–2320, 2016. 3

Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., and Girshick, R. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 6

Yu, J., Xu, Y., Koh, J. Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B. K., et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022. 3

Yuan, L., Chen, D., Chen, Y.-L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 2, 3, 4

Yun, Y. K. and Lin, W. Selfreformer: Self-refined network with transformer for salient object detection. *arXiv preprint arXiv:2205.11283*, 2022. 2, 5, 7

Zang, Y., Huang, C., and Loy, C. C. Fasa: Feature augmentation and sampling adaptation for long-tailed instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3457–3466, 2021. 2

Zareian, A., Rosa, K. D., Hu, D. H., and Chang, S.-F. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14393–14402, 2021. 3

Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L. H., Zhou, L., Dai, X., Yuan, L., Li, Y., et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16793–16803, 2022. 3, 8, 9

Zhou, X., Koltun, V., and Krähenbühl, P. Probabilistic two-stage detection. *arXiv preprint arXiv:2103.07461*, 2021. 6, 7, 9

Zhou, X., Girdhar, R., Joulin, A., Krähenbühl, P., and Misra, I. Detecting twenty-thousand classes using image-level supervision. *arXiv preprint arXiv:2201.02605*, 2022. 3, 7, 8, 9