Gen2Det: Generate to Detect

Saksham Suri[1*] Fanyi Xiao[2] Animesh Sinha[2] Sean Chang Culatana[2]
Raghuraman Krishnamoorthi[2] Chenchen Zhu[2†] Abhinav Shrivastava[1†]
University of Maryland, College Park[1] Meta[2]

Abstract

Diffusion models have shown improvement in synthetic image quality as well as better control in generation. We motivate and present Gen2Det, a simple modular pipeline to create synthetic training data for object detection by leveraging state-of-the-art grounded generation methods. Unlike existing works which generate individual object instances, require identifying foreground followed by pasting on other images, we simplify to directly generating scene-centric images. In addition to the synthetic data, Gen2Det also proposes a suite of techniques to best utilize the generated data, including image-level filtering, instance-level filtering, and better training recipe to account for imperfections in the generation. Using Gen2Det, we show healthy improvements on object detection and segmentation tasks on standard benchmarks like COCO and LVIS.

1. Introduction

Recent developments in generative modeling using diffusion models has drastically improved the generation quality. Works like LDM [17], DALL-E [15, 16], Imagen [19], Parti [23] have shown the generation power which diffusion models possess. In addition to these models which can generate high quality images given a text input, there have been multiple developments in the direction of higher control in generation. Along this direction exist works which use conditional control like ControlNet [25] and GLIGEN [9]. There are also works like LoRA [8] and Dreambooth [18] which provide means to adapt these large models in a quick and efficient manner to generate specific kinds of images. With these developments in both quality and control of synthetically generated images, it is only natural to come back to the question of "How to best utilize data generated from these models for improving recognition performance?".

Most previous works [6, 20, 22] have explored using synthetic data from diffusion models for classification and pre-training tasks. The goal of our work is to look at utilizing more realistic scene configurations by generating images conditioned on the existing layout of boxes and labels and use them for object detection and segmentation. A recent work, XPaste [26], explores this problem and utilizes techniques similar to simple copy-paste [4] to paste synthetically generated object-centric instances onto real images for object detector training. However, their method relies on off-the-shelf segmentation methods built over CLIP [12–14, 21, 24] to extract masks, and is thus subject to segmentation errors. In addition to the extra components and com-

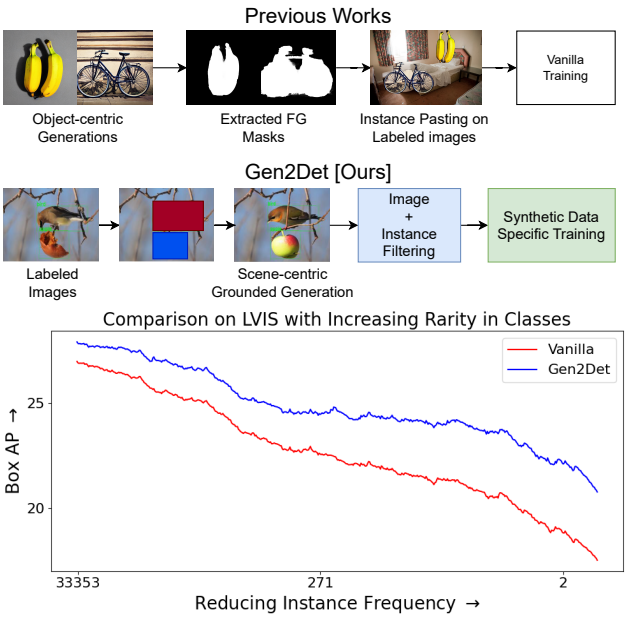

Figure 1. Existing approaches which utilize synthetic data for detection training follow a common methodology of generating object-centric images and pasting instances on real images (top). Gen2Det (middle) instead utilizes state-of-art grounded inpainting diffusion model to directly generate scene-centric images. Further, using filtering and modified training Gen2Det is able to utilize synthetic instances better. As a result, our method consistently improves over vanilla training and the AP improvements increase (bottom) as the class becomes rare (*i.e.*, long-tailed classes).

*Work done during internship at Meta.
†*Equal Advisory Contribution.

pute, the generated images from XPaste are also not realistic as they do not respect natural layouts due to random pasting as shown in Figure 1 (top). In contrast, as shown in Figure 1 (middle), Gen2Det leverages state-of-art diffusion models for grounded inpainting to generate scene-centric images which look more realistic. The goal of our approach is to show that utilizing such generated data from state-of-art diffusion models can lead to improvement in performance of object detection and segmentation models. By using a grounded inpainting diffusion model we are able to generate synthetic versions of common detection datasets like LVIS [5] and COCO [11] in a layout conditioned manner. We then carefully design a set of filtering and training strategies, with which we demonstrate that we can improve detection performance when training on the joint set of real and synthetic images. More specifically, we sample batches of synthetic and real images with a sampling probability. During loss computation we also modify the loss for synthetic data to account for filtered instances. The clear improvement over vanilla training shown in Figure 1 (bottom) for classwise Box AP with increasing rarity of classes makes a strong case for such a pipeline especially in long-tailed or low data regimes.

2. Related Works

Diffusion Models for Controllable Image Generation. Diffusion models have been shown to generate images with unprecedented high quality. Amongst these models a few of the popular models which can generate images from text prompts include LDM [17], DALL-E [15, 16], Imagen [19], and Parti [23]. One of the popular works ControlNet [25] provides multiple ways to enable control including using edge maps, scribbles, segmentation masks amongst other modalities. Another work GLIGEN [9] brings in more grounding control where they show image generation and inpainting conditioned on boxes, keypoints, HED maps, edge maps and semantic maps. We also utilize a similar pretrained state-of-art diffusion model for grounded inpainting.

Leveraging Synthetic Data from Diffusion Models. Due to its high quality generation and the flexibility in handling multimodal data (*e.g.*, vision and language), there has been a lot of recent work in using pretrained diffusion models for different tasks. StableRep [22] uses data generated by diffusion models to train self-supervised models. There have also been works exploring the use of synthetic data from diffusion models to improve image classification [1, 2, 6, 7, 20]. While the classification task has had a lot of exploration with synthetic data due to the object-centric nature of images which these models can generate easily, the detection task is less explored as it's harder to generate data with grounded annotations using these models. A recent study [10] explored the use of such data in highly constrained few shot settings for detection where the gains are expected. Recently, XPaste [26] looked at more general benchmarks and showed consistent improvements using the Centernet2 [27] architecture. Specifically, XPaste [26] uses diffusion models to generate object-centric images and uses multiple CLIP [14] based approaches [12, 13, 21, 24] to extract segmentation maps which makes it slow and error-prone. Following the extraction of instances and their segmentation maps, they use synthetic and real instances (retrieved from an additional real dataset) to perform copy-paste [4] augmentation to generate synthetic data. Gen2Det is able to directly generate diverse instances in a scene-centric manner and with proper filtering and training recipes show improvements in performance. Also in terms of training speed, we are $3.4\times$ faster compared to XPaste with the same configuration.

3. Approach

Through Gen2Det, we provide a modular way to generate and utilize synthetic data for object detection and instance segmentation. As shown in Figure 2, we start by generating data using state-of-art grounded image inpainting diffusion model. As the generations may not always be perfect we perform image level and instance level filtering. The image level filtering is performed using a pre-trained aesthetic classifier [3] while the instance level filtering is performed using a detector trained on the corresponding real data. To train we perform a probability based sampling of a batch to be composed of synthetic or real samples. Further while computing the losses, we modify the negatives corresponding to the synthetic data to be ignored from loss computation to account for the filtering we performed. We also do not apply mask loss for synthetic samples as we do not have the segmentation masks corresponding to them.

Image Generation. We start our pipeline by generating synthetic images by utilizing a state-of-art grounded inpainting diffusion model. This model is trained to support multiple kinds of input conditions. We utilize the model trained for image inpainting with the image, boxes, and corresponding labels as input. As inpainting model requires an image and box level text description, for each box b_i we use the class name $<c_i>$ as the box level prompt and a concatenation of strings $<$a c_1, a c_2, ... and a $c_n>$ as the image level prompt where n is the number of instances in the image. We show examples of images generated using this technique in Figure 3.

Image Level Filtering. The first level of filtering we perform is an image level filtering using a pretrained classifier. The model is trained to predict how pleasing the image looks visually by assigning an aesthetic score to each image. We utilize the publicly available weights from [3] and pass every generated image through the classifier. Any image with an aesthetic score less than τ_a is discarded and its

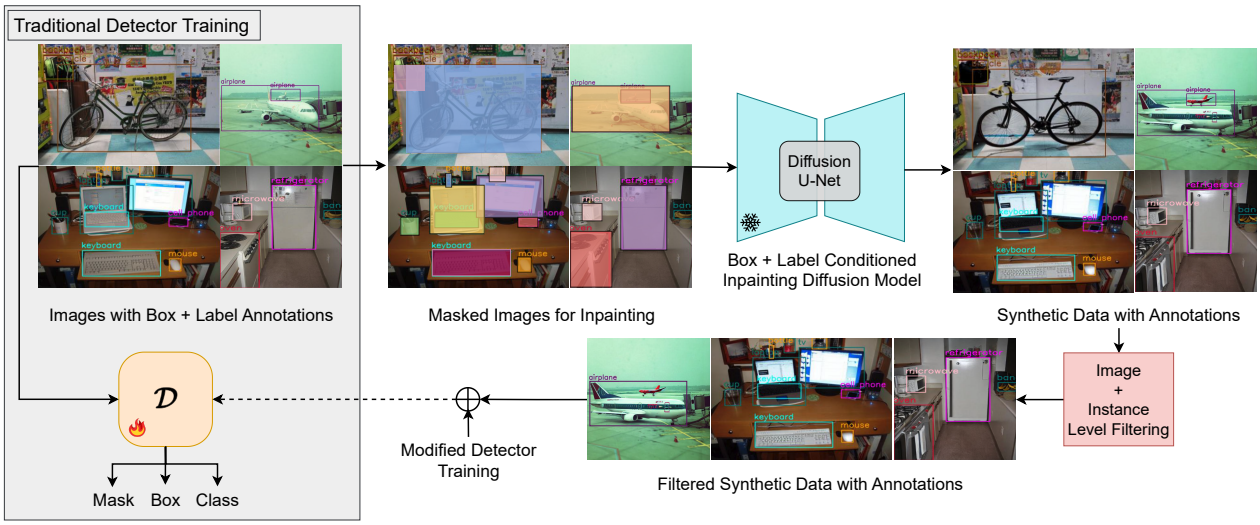

Figure 2. **Gen2Det: our proposed pipeline for generating and utilizing synthetic data for object detection and segmentation.** Gen2Det starts by generating grounded inpainted images using state-of-art diffusion model. The generated images are then filtered at image and instance level. Finally, we train object detection and segmentation models using the filtered data along with our improved training methodology by introducing sampling and background ignore.

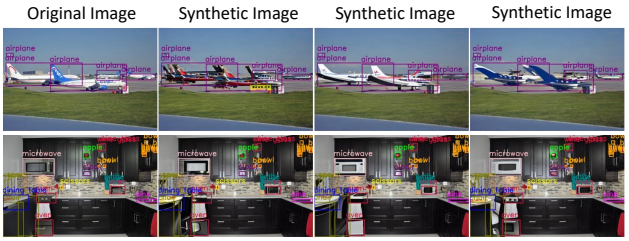

Figure 3. Examples of generations using the inpainting diffusion model on the COCO dataset.

annotations are removed. We show the effect of this filtering by visualizing discarded samples in Figure 4.

Instance Level Filtering. We next perform instance level filtering to remove annotations for specific generated instances which do not have good generation quality. In order to perform this filtering we first train a detector on only the real data. We then pass all the generated images through the trained detector and store its predictions. Based on the detectors predictions we evaluate whether a ground truth annotation corresponding to an inpainted region should be utilized for training or not. To do this, for each generated image we iterate over all ground truth annotations used to generate it. For each annotation we go through all predictions and remove the ground truth annotation if there is no overlapping prediction with a score greater than τ_s and IoU greater than τ_{iou}. We show results of the kind of annotations removed using this filtering in Figure 5.

Training. We utilize both the real and synthetic data for training the final model using a probabilistic sampling of real and synthetic image batches. While the filtering we perform at both image and instance level can deal with bad quality generations, they also introduce some noise especially for classes for which the detector itself has poor quality predictions. Essentially, the instance level filtering removes ground truth annotation corresponding to bad quality/incorrect generations but that does not remove the bad instance itself from the image. Additionally, the generative model could have also hallucinated multiple instance of an object class leading to missing annotations. To counter the effect of both these scenarios we introduce an ignore functionality during training. Essentially, for both the region proposal network (RPN) and detector head we ignore the background regions from the loss computation, if their fg/non-bg class prediction score is higher than a threshold τ_i. This lets us train even in the presence of some bad quality regions without incorrectly penalizing the detector for predicting objects at those locations.

4. Results

4.1. Experimental Setting

We evaluate our approach on LVIS [5] and COCO [11] datasets. LVIS is a long tailed dataset with 1203 classes utilizing the same images as COCO which contains only 80 classes. The long tailed nature of LVIS makes it especially appealing to showcase the use of synthetic data for the rare categories. We report the Average Precision (AP) for both these datasets and compare using the box AP (AP_b) as well as mask AP (AP_m). Additionally for LVIS, we also report APs specific to rare, common, and frequent classes.

For data generation, we utilize a state-of-art diffusion model for grounded inpainting using the images from LVIS

Method	Box				Mask			
	AP	AP_r	AP_c	AP_f	AP	AP_r	AP_c	AP_f
Vanilla (real only)	33.80	20.84	32.84	40.58	29.98	18.36	29.64	35.46
Vanilla (synth only)	11.27	6.54	9.35	15.51	-	-	-	-
Vanilla (synth+real)	31.82	20.68	31.18	37.42	27.49	18.35	27.08	31.96
XPaste	34.34	21.05	33.86	40.71	30.18	18.77	30.11	35.28
Ours	34.70	23.78	33.71	40.61	30.82	21.24	30.32	35.59

Table 1. Comparisons on LVIS with baselines using Centernet2 architecture. We report the overall AP and also AP for rare (AP_r), common (AP_c) and frequent (AP_f) classes.

Method	AP_b	AP_m
Vanilla (real only)	46.00	39.8
Vanilla (synth only)	24.51	-
Vanilla (synth+real)	44.35	37.03
XPaste	46.60	39.90
Ours	47.18	40.44

Table 2. Comparisons with baselines on COCO dataset using Centernet2 architecture using AP_b (box) and AP_m (mask).

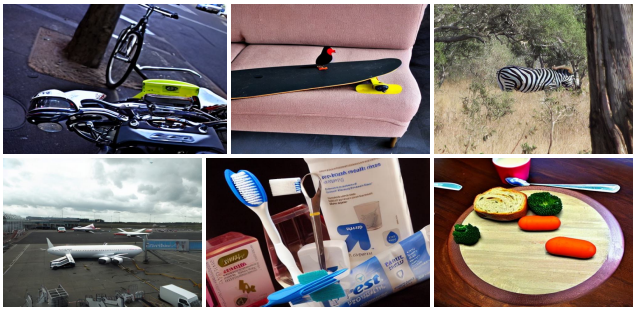

Figure 4. Samples discarded during the image level filtering.

and COCO along with their annotations which contain the boxes and class labels. For aesthetic filtering, we utilize the open source model available in their repository [3]. For aesthetic filtering we set τ_a to 4.5 which is roughly the same as the average aesthetic score on real COCO images. For detector filtering we set τ_s as 0.2 and τ_{iou} as 0.3 for LVIS. We use the same τ_{iou} for COCO and set τ_s as 0.1. During training, we use a sampling probability $p = 0.2$. We set τ_i to 0 for both datasets thus effectively ignore all background regions corresponding to the synthetic data from the loss.

4.2. Quantitative Results

Comparison on LVIS. We start by comparing our approach to the existing works in Table 1 on the LVIS dataset with CenterNet2 backbone. Over vanilla training we show that our method improves the Box and Mask AP over rare categories by 2.94 and 2.88 respectively with a 0.9 and 0.84 Box and Mask AP improvement overall. Despite the fact that XPaste utilizes four different CLIP based models [12, 13, 21, 24] to obtain segmentation masks and is also 3.5× slower to train compared to our approach on LVIS, we are able to outperform XPaste by 2.73 Box AP and 2.47 Mask AP on the rare categories and by 0.36 and 0.64 Box and Mask AP across all categories. The huge gains in rare categories and overall improvements highlight our methods effectiveness in both long tailed as well as general settings.
Comparison on COCO. Similar to LVIS, we compare on the COCO benchmark in Table 2. On COCO too we show an improvement of 1.18 and 0.64 on Box and Mask AP over vanilla training. Compared to XPaste we improve by 0.58 Box AP and 0.54 Mask AP. We note that compared to LVIS

Figure 5. Instances discarded by instance level filtering (red).

the improvements are slightly lower here as LVIS is a more long-tailed dataset where adding synthetic data shines.

It should be noted that the improvement in mask AP for both COCO and LVIS is without additional mask annotations as the synthetic data only contains box annotations.

4.3. Qualitative Results

We show the outputs of different parts of pipeline qualitatively on the COCO dataset. First, in Figure 3 we show a few samples which are synthetic images generated using the pre-trained grounded inpainting diffusion model. In Figure 4 we highlight a few samples which are rejected from the image level filtering. Finally, in Figure 5 we show some examples of instances discarded by detector filtering.

5. Conclusion

With the huge strides in image generation in terms of both quality and control, we try to tackle the problem of training detection and segmentation models with synthetic data. Our proposed pipeline Gen2Det utilizes state-of-art grounded inpainting diffusion model to generate synthetic images which we further filter at both image and instance level before using in training. We also introduce some changes in the detector training to utilize the data better and take care of shortcoming which the data might pose. We show improvement across both LVIS and COCO and show higher improvements on rare classes in the long tailed LVIS setting. Most interestingly, we show improvement in segmentation performance without using any additional segmentation masks like existing works.

Acknowledgements

We thank Justin Johnson for discussion on this work. We also thank Vikas Chandra for helping with the submission.

References

[1] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J Fleet. Synthetic data from diffusion models improves imagenet classification. *arXiv preprint arXiv:2304.08466*, 2023. 2

[2] Hritik Bansal and Aditya Grover. Leaving reality to imagination: Robust classification via generated datasets. *arXiv preprint arXiv:2302.02503*, 2023. 2

[3] Christophschuhmann. Christophschuhmann/improved-aesthetic-predictor: Clip+mlp aesthetic score predictor. 2, 4

[4] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2918–2928, 2021. 1, 2

[5] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 2, 3

[6] Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition? *arXiv preprint arXiv:2210.07574*, 2022. 1, 2

[7] Reyhane Askari Hemmat, Mohammad Pezeshki, Florian Bordes, Michal Drozdzal, and Adriana Romero-Soriano. Feedback-guided data synthesis for imbalanced classification. *arXiv preprint arXiv:2310.00158*, 2023. 2

[8] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 1

[9] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023. 1, 2

[10] Shaobo Lin, Kun Wang, Xingyu Zeng, and Rui Zhao. Explore the power of synthetic data on few-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 638–647, 2023. 2

[11] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2, 3

[12] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7086–7096, 2022. 1, 2, 4

[13] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar Zaiane, and Martin Jagersand. U2-net: Going deeper with nested u-structure for salient object detection. page 107404, 2020. 2, 4

[14] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2

[15] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 1, 2

[16] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 1, 2

[17] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2

[18] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 1

[19] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 1, 2

[20] Mert Bulent Sariyildiz, Karteek Alahari, Diane Larlus, and Yannis Kalantidis. Fake it till you make it: Learning transferable representations from synthetic imagenet clones. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2

[21] Yukun Su, Jingliang Deng, Ruizhou Sun, Guosheng Lin, and Qingyao Wu. A unified transformer framework for group-based segmentation: Co-segmentation, co-saliency detection and video salient object detection, 2022. 1, 2, 4

[22] Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. *arXiv preprint arXiv:2306.00984*, 2023. 1, 2

[23] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022. 1, 2

[24] Yi Ke Yun and Weisi Lin. Selfreformer: Self-refined network with transformer for salient object detection. *arXiv preprint arXiv:2205.11283*, 2022. 1, 2, 4

[25] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 1, 2

[26] Hanqing Zhao, Dianmo Sheng, Jianmin Bao, Dongdong Chen, Dong Chen, Fang Wen, Lu Yuan, Ce Liu, Wenbo Zhou, Qi Chu, et al. X-paste: Revisiting scalable copy-paste for instance segmentation using clip and stablediffusion. 2023. 1, 2

[27] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Probabilistic two-stage detection. *arXiv preprint arXiv:2103.07461*, 2021. 2