Appendix

November 2, 2023

A Related Works

A.1 Diffusion Models for Synthetic Image Generation

Recently, diffusion models have witnessed remarkable successes in various fields [26, 28, 34], revolutionizing the way of generative neural networks. From generation of natural language [18, 10] to computer vision [3, 26, 28] and beyond [23, 12], these models have shown unparalleled capabilities in understanding and generating data. Specifically, in computer vision, diffusion models have excelled in generating high-quality and realistic visual contents for tasks such as text-to-image generation [26, 28], super-resolution [16], inpainting [28, 34], etc. Moreover, their ability to handle uncertainty and generate diverse outputs has made them exceed previous generative models.

A.2 Image Data Augmentation

Thus diffusion models are used to generate synthetic images to help with image classification. [11] demonstrated synthetic data from GLIDE [22] improves zero-shot and few-shot image classification performance. And [1] proves that synthetic data from Imagen [30] improves image classification performance on ImageNet [5]. However, those approaches are not able to generate objects bounded tightly by a given box, and thus are not applicable for object detection tasks.

Recent works try to utilize diffusion models together with the copy-paste method [9] to improve the performance of object detection. [7] use image captioning [27] model to guide the generation for DALLE [26], and used off-the-shelf segmentation model [24] to cut off the object from the synthetic image then paste to real images. While [37] used a fixed prompt and also used off-the-shelf segmentation models [25, 32, 20, 21] to cut the objects off. However, there is a concern that the image generated by copy-paste methods are not realistic [3], and the captioning method [27] is trained on COCO[19] which may cause a data leakage for experiments on COCO [19] and PASCAL VOC [6].

Controllable diffusion models showed the ability to manipulate and guide the generation process while maintaining the advantages of the diffusion process. The novel concept of controllability enables users to have fine-grained control over the generated contents. For example, layout-to-image generation [4, 14] takes the layout description as input and synthesizes an image that corresponds to the given layout. However, such models always require large object detection datasets for pretraining and the effectiveness to help object detectors are not shown. [3] for the first time showed that a layout-to-image model can produce synthetic images to help the training of object detection models, while it requires to use the detection data to train the diffusion model first. [36] emerged as a crucial component of controllable diffusion, making it possible to guide the text-to-image generation with visual priors like edge maps [2, 33], segmentation masks [17], scribbles [33], etc. And it inspires us to build a pipeline generating synthetic images with tight bounding box annotations.

B More Metrics for Important Experiments

We reported only mAP and AP50 for most experiments due to space limits. Here we add results on more metrics, i.e. AP75 and mAP-s/m/l, for more detailed comparisons.

B.1 Few Shot

Table 1 reports mAP, AP50, AP75, mAP-s, mAP-m, mAP-l for the experiments on COCO [19] under few shot settings. Our approach generally improves the detectors' performance by a large margin. An interesting finding is that inpainting methods are sometimes better than our approach on metric mAP-s, which shows that using inpainting may be better when drawing small objects.

B.2 PASCAL VOC and Object Detection in the Wild

Table 2 also add AP75, mAP-s/m/l in addition to previous results for YOLOX-S [8] on PASCAL VOC [6] and downstream datasets [13, 15, 31, 29]. In general our approach boosts the object detection performance significantly. The performance drops in Plantdoc [31] and Deepfruits [29] are largely due to the similar appearance in object edges among different categories. For instance, given a HED edge map as input, the model may get confused to distinguish a healthy leaf over a defected leaf, and is not able to generate the correct data given only the HED map and the prompt as input. This remains a future work for us to further explore.

B.3 Other Visual Priors

Table 3 compare the HED visual prior and other visual priors (i.e., Canny edge map [2], Uniformer segmentation mask [17], Scribble edge map [33, 36]) as in the body part of the paper, but further report AP75, mAP-s, mAP-m, mAP-l in addition to previous mAP and AP50 metrics. From Canny [2], HED [33], Uniformer [17], to Scribble [33, 36], the control of visual prior turns from fine to coarse. The experiment results suggests us to choose a medium control level like HED [33] or Uniformer [17] that does not include too many details of the object or noises, nor too coarse to lose the robustness and cause more distortions.

For the algorithm to generate Scribble edges, see here [33, 36].

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. 1

[2] John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 1986. 1, 2, 4

[3] Kai Chen, Enze Xie, Zhe Chen, Lanqing Hong, Zhenguo Li, and Dit-Yan Yeung. Integrating geometric control into text-to-image diffusion models for high-quality detection data generation via text prompt. *arXiv preprint arXiv:2306.04607*, 2023. 1

[4] Jiaxin Cheng, Xiao Liang, Xingjian Shi, Tong He, Tianjun Xiao, and Mu Li. Layoutdiffuse: Adapting foundational diffusion models for layout-to-image generation. *arXiv preprint arXiv:2302.08908*, 2023. 1

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[6] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010. 1, 2, 3

[7] Yunhao Ge, Jiashu Xu, Brian Nlong Zhao, Laurent Itti, and Vibhav Vineet. Dall-e for detection: Language-driven context image synthesis for object detection. *arXiv preprint arXiv:2206.09592*, 2022. 1

[8] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 2, 3

[9] 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 *Proceed-*

		mAP	mAP50	mAP75	mAP-s	mAP-m	mAP-l
5-shot	YOLOX-S	5.0	10.1	4.1	0.7	4.1	7.7
	+ SDInpaint	5.3	11.1	4.2	0.8	4.0	8.6
	+ PbE	5.5	**11.4**	4.8	**0.9**	4.2	9.1
	+ Ours	**5.9**	**11.4**	**5.3**	0.7	**4.5**	**9.3**
	DINO-SwinL	18.6	26.0	19.4	**4.3**	14.3	29.1
	+ Ours	**20.3**	**28.1**	**21.4**	**4.3**	**15.1**	**32.8**
10-shot	YOLOX-S	9.6	18.1	9.2	2.5	8.4	14.8
	+ SDInpaint	10.6	19.9	10.1	2.4	8.8	16.5
	+ PbE	9.8	18.9	8.9	2.5	7.6	15.5
	+ Ours	**11.1**	**20.6**	**10.5**	**3.2**	**9.8**	**17.3**
	DINO-SwinL	24.3	33.7	25.6	**7.7**	20.8	36.1
	+ Ours	**26.0**	**36.8**	**27.1**	**7.7**	**23.2**	**39.5**
30-shot	YOLOX-S	14.2	26.7	13.6	4.2	13.5	21.4
	+ SDInpaint	14.6	26.4	14.1	**5.1**	13.8	22.6
	+ PbE	14.7	26.5	14.4	4.4	13.7	23.0
	+ Ours	**15.3**	**27.0**	**15.2**	4.2	**14.7**	**24.0**
	DINO-SwinL	**35.8**	**49.5**	**38.6**	**14.3**	**34.8**	50.6
	+ Ours	35.0	48.8	36.9	11.5	33.7	**51.5**

Table 1: We further report AP75, mAP-s, mAP-m, mAP-l in addition to previous mAP and AP50 metrics on COCO [19] at 5/10/30-shot for the light-weight detector YOLOX-S [8] and high-performance state-of-the-art detector DINO-SwinL [35].

		mAP	mAP50	mAP75	mAP-s	mAP-m	mAP-l
PASCAL VOC	YOLOX-S	52.5	77.1	58.1	16.7	38.8	61.0
	+ Ours	**53.7**	**77.9**	**59.5**	**17.0**	**41.0**	**62.1**
Watercolor	YOLOX-S	11.6	26.2	8.4	1.3	3.6	14.1
	+ Ours	**16.5**	**35.7**	**12.0**	**2.4**	**5.5**	**22.2**
Raccoon	YOLOX-S	22.8	70.1	7.2	n/a	20.2	22.8
	+ Ours	**37.5**	**78.8**	**41.1**	n/a	**30.3**	**38.2**
Thermal	YOLOX-S	61.0	89.6	76.3	73.4	78.9	57.9
	+ Ours	**72.2**	**94.0**	**89.1**	**75.5**	**69.0**	**73.2**
Plantdoc	YOLOX-S	**39.8**	**54.0**	**48.5**	n/a	23.0	**40.6**
	+ Ours	38.6	53.4	46.3	n/a	**26.0**	39.5
Deepfruits	YOLOX-S	**57.6**	**87.2**	**64.3**	40.1	**44.5**	**61.8**
	+ Ours	51.5	80.0	58.7	**46.8**	37.8	55.8
Comic	YOLOX-S	10.1	22.2	7.1	**0.8**	5.1	12.1
	+ Ours	**12.0**	**26.2**	**9.3**	0.2	**5.6**	**14.2**

Table 2: We further report AP75, mAP-s, mAP-m, mAP-l in addition to previous mAP and AP50 metrics on PASAL VOC [6] and downstream datasets Watercolor [13], Raccoon [15], Thermal [15], Plantdoc [31], Deepfruits [29], and Comic [13].

ings of the IEEE/CVF conference on computer vision and pattern recognition, pages 2918–2928, 2021. 1

[10] Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and LingPeng Kong. Diffuseq: Sequence to sequence text generation with diffusion models. *arXiv preprint arXiv:2210.08933*, 2022. 1

[11] Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative mod-

		mAP	mAP50	mAP75	mAP-s	mAP-m	mAP-l
5-shot	Canny	5.2	10.6	4.5	1.1	4.4	8.6
	HED	**5.9**	**11.4**	**5.3**	0.7	**4.5**	**9.3**
	Uniformer	5.2	10.8	4.4	**1.2**	3.9	7.9
	Scribble	5.1	10.6	4.4	**1.2**	4.2	7.8
10-shot	Canny	11.2	20.4	10.8	2.8	10.0	17.2
	HED	11.1	20.6	10.5	**3.2**	9.8	17.3
	Uniformer	**11.5**	**20.9**	**11.2**	3.1	**10.5**	**17.7**
	Scribble	9.7	18.5	9.4	3.1	8.5	15.1
30-shot	Canny	14.0	25.4	13.4	3.9	13.5	22.0
	HED	**15.3**	**27.0**	**15.2**	**4.2**	**14.7**	**24.0**
	Uniformer	14.3	25.5	14.0	**4.2**	12.6	22.1
	Scribble	13.9	24.7	14.1	3.8	12.5	21.7

Table 3: We further report AP75, mAP-s, mAP-m, mAP-l in addition to previous mAP and AP50 metrics on our approach with HED edge map [33] comparing with other visual priors, i.e, Canny edge map [2], Uniformer segmentation mask [17], Scribble edge map [33, 36].

els ready for image recognition? *arXiv preprint arXiv:2210.07574*, 2022. 1

[12] Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. *arXiv preprint arXiv:2301.12661*, 2023. 1

[13] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5001–5009, 2018. 2, 3

[14] Naoto Inoue, Kotaro Kikuchi, Edgar Simo-Serra, Mayu Otani, and Kota Yamaguchi. Layoutdm: Discrete diffusion model for controllable layout generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10167–10176, 2023. 1

[15] Chunyuan Li, Haotian Liu, Liunian Harold Li, Pengchuan Zhang, Jyoti Aneja, Jianwei Yang, Ping Jin, Houdong Hu, Zicheng Liu, Yong Jae Lee, and Jianfeng Gao. Elevater: A benchmark and toolkit for evaluating language-augmented visual models. *Neural Information Processing Systems*, 2022. 2, 3

[16] Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2022. 1

[17] Kunchang Li, Yali Wang, Junhao Zhang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unifying convolution and self-attention for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1, 2, 4

[18] Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. *Advances in Neural Information Processing Systems*, 35:4328–4343, 2022. 1

[19] 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. 1, 2, 3

[20] 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*, pages 7086–7096, 2022. 1

[21] 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*, pages 7086–7096, 2022. 1

[22] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided

diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 1

[23] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 1

[24] Lu Qi, Jason Kuen, Yi Wang, Jiuxiang Gu, Hengshuang Zhao, Philip Torr, Zhe Lin, and Jiaya Jia. Open world entity segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1

[25] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R Zaiane, and Martin Jagersand. U2-net: Going deeper with nested u-structure for salient object detection. *Pattern recognition*, 106:107404, 2020. 1

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

[27] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7008–7024, 2017. 1

[28] 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

[29] Inkyu Sa, Zongyuan Ge, Feras Dayoub, Ben Upcroft, Tristan Perez, and Chris McCool. Deepfruits: A fruit detection system using deep neural networks. *sensors*, 16(8):1222, 2016. 2, 3

[30] 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

[31] Davinder Singh, Naman Jain, Pranjali Jain, Pratik Kayal, Sudhakar Kumawat, and Nipun Batra. Plantdoc: A dataset for visual plant disease detection. In *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*, pages 249–253. 2020. 2, 3

[32] Yukun Su, Jingliang Deng, Ruizhou Sun, Guosheng Lin, Hanjing Su, and Qingyao Wu. A unified transformer framework for group-based segmentation: Co-segmentation, co-saliency detection and video salient object detection. *IEEE Transactions on Multimedia*, 2023. 1

[33] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1395–1403, 2015. 1, 2, 4

[34] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18381–18391, 2023. 1

[35] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 3

[36] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 1, 2, 4

[37] 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