V3Det: Vast Vocabulary Visual Detection Dataset
Supplementary Materials

Jiaqi Wang*[1], Pan Zhang*[1], Tao Chu*[1], Yuhang Cao*[2],
Yujie Zhou[1], Tong Wu[2], Bin Wang[1], Conghui He[1], Dahua Lin[1,2,3]
[1]Shanghai AI Laboratory, [2]The Chinese University of Hong Kong,
[3]Centre of Perceptual and Interactive Intelligence
{wangjiaqi,zhangpan}@pjlab.org.cn, dhlin@ie.cuhk.edu.hk

In the supplementary materials, we introduce the license of V3Det dataset in Appendix A, more implementation details in Appendix B, and more experimental results in Appendix C. We show a more detailed visualization of the hierarchy category organization of V3Det in Appendix D, the list of coarse categories which is used during annotation process in Appendix E, some examples of category descriptions written by human experts and a powerful chatbot, *i.e.*, chatgpt[1] in Appendix F, and more visualizations of the V3Det dataset in Appendix G.

A. V3Det Dataset License and Download.

V3Det Images. Around 90% of the images in V3Det were selected from the Bamboo Dataset [20], sourced from the Flickr website. The remaining 10% were directly crawled from the Flickr. We do not own the copyright of the images. Use of the images must abide by the Flickr Terms of Use[2]. We only provide lists of image URLs without redistribution.
V3Det Annotations. The V3Det annotations, the category relationship tree, and related tools are licensed under a Creative Commons Attribution 4.0 License [3].
V3Det Download. The metafile of image URLs, a script to easily download images, annotations, the category relationship tree, and other related tools are available at https://v3det.openxlab.org.cn/.

B. Implementation Details.

B.1. Vast Vocabulary Object Detection.

Two Stage and Cascaded Detectors. We test three representative two-stage or cascaded detectors with ResNet-50 [9] and Swin-B [12] backbones, *i.e.*, Faster R-CNN [15], Cascade R-CNN [2], and CenterNet2 [23]. For Faster R-CNN and Cascade R-CNN, we adopt their standard implementations in mmdetection [4]. For CenterNet2, we adopt its official implementation based on Detectron2 [18]. To have a fair comparison, FPN [11], AdamW optimizer[13], multi-scale augmentation, and repeat factor sampler [8] training are adopted in experiments. Specifically, we use AdamW optimizer [10] with learning rate of 10^{-4}, β_1=0.9, β_2=0.999, and batchsize of 32. Following the best practices on LVIS [8] dataset, we use multi-scale augmentation training by randomly selecting the shorter side of the image from (640, 672, 704, 736, 768, 800) pixels and the longer size of the image less than 1333 pixels. Following LVIS [8], we set the repeat factor to 10^{-3} in the repeat factor sampler [8]. The detectors are trained with 24 epochs, where the learning rate is decreased by 10$\times$ at 16 and 22 epochs.
Single Stage Detectors. We adopt the standard implementation of ATSS[3] and FCOS [16] in mmdetection [4] as representative single stage detectors. Other settings follow the experiments on two stage and cascaded detectors.
DETR Style Detectors. We test two DETR-style detectors, Deformable DETR [24], and DINO [19] with two different backbones, ResNet-50 [9], and Swin-B [12]. We implement Deformable DETR with mmdetection [4] and train it with batch size 32 for 50 epochs with the learning rate of 2×10^{-4}. We implement DINO with detrex [5] and train it with batch size 16 for 24 epochs with the learning rate of 10^{-4}. Both models are trained with AdamW optimizer [10] with β_1=0.9, β_2=0.999 and repeat factor sampler with factor of 10^{-3}. We follow the default data augmentation that is adopted in the corresponding framework.

B.2. Open Vocabulary Object Detection (OVD).

In this section, we provide the implementation details of training Detic [22] and RegionCLIP [21] on V3Det.

For Detic, we follow the original paper that uses CenterNet2 [23] with ResNet-50 [9]. We adopt large-scale jittering [7] with an input resolution of 640x640. The batch size

* equal contribution.
[1]https://openai.com/blog/chatgpt
[2]https://www.flickr.com/creativecommons/
[3]https://creativecommons.org/licenses/by/4.0/

Pretrain (1st)	Pretrain (2st)	Finetune	AP
-	-	LVIS	35.1
V3Det	-	LVIS	36.2
Objects365	-	LVIS	36.9
V3Det	Objects365	LVIS	37.3
Objects365	V3Det	LVIS	37.7

Table A1: Comparisons of different strategies when pretraining a CenterNet2 using R-50 backbone and Norm Linear Layer on V3Det and Objects365, followed by finetuning on LVIS. For each stage, the training schedule is 90k iterations of batch size 64.

is 64, and the learning rate is 2×10^{-4} with the AdamW optimizer. For results in main text Table 6, we first train a model on base classes C_{base} of V3Det for 90k iterations. We then create a subset of ImageNet-21K, named IN-V3Det, which contains 4197 overlapped classes with V3Det. We then do a multi-dataset finetuning on V3Det and IN-V3Det for 90k iterations. For results in main text Table 8, Detic of V3Det* is only trained on V3Det* without ImageNet images for 180k iterations.

For RegionCLIP, following the original paper, we utilize Faster R-CNN [15] with ResNet-50 [9] backbone that is finetuned by CLIP-guided region-text alignment. Initially, the offline Region Proposal Network (RPN) [15] is trained on the base categories for 180k iterations. Subsequently, the finetune stage is applied to transfer the learned knowledge of region-text alignment to open-vocabulary detection for 180k iterations. During training, a batch size of 16 images and an initial learning rate of 2×10^{-3} are utilized.

C. More Experimental Results.

V3Det for pretraining. V3Det aims to be a benchmark for vast-vocabulary and open-vocabulary object detection. Interestingly, our findings indicate that V3Det can additionally be employed effectively as a pretraining dataset. We conducted comparative analyses of diverse strategies for pretraining a CenterNet2 [23] with R-50 on datasets V3Det and Objects365. Subsequent finetuning was performed on the LVIS dataset to assess the effectiveness of the initial pretraining. Table A1 shows that V3Det and Objects365 exhibit comparable performance and can complement each other for pretraining.

Ablation on Norm Linear Layer. In the experiments of Table 2 and Table 4 in the main text, we show the effectiveness of Norm Linear Layer [17], $z = \tau \widetilde{\mathcal{W}}^T \widetilde{x} + b$, where $\widetilde{\mathcal{W}}_i = \frac{\mathcal{W}_i}{\|\mathcal{W}_i\|_2}, i \in C, \widetilde{x} = \frac{x}{\|x\|_2}$. As shown in Table A2, we explore its temperature factor τ, and find that the best performance is achieved when τ is 50. Therefore, we adopt τ of 50 as the default setting in experiments.

Different Stages in Cascaded R-CNN. As in Table A3, we show the performance of Cascade R-CNN [2] ResNet-50 at different stages. The AP increases as the stage grows, demonstrating the effectiveness of cascading refinement cascade. On the other hand, the AP of Cascade R-CNN at stage 1 is 22.7, which is still higher than the AP of Faster R-CNN [15], which is 21.2, indicating the structure of stage cascade is beneficial to the model optimization.

τ	AP	AP_{50}	AP_{75}
30	27.4	32.9	29.7
40	27.8	34.0	30.3
50	**28.3**	**34.5**	**30.8**
60	27.8	33.9	30.3

Table A2: Comparisons of different temperature factors τ in Norm Linear Layer. Cascade R-CNN ResNet-50 trained for 12 epochs with the AdamW optimizer is the framework in this table.

Stage	AP	AP50	AP75
Faster R-CNN [15]	21.2	29.5	24.1
Cascade R-CNN stage 1	22.7	31.1	25.8
Cascade R-CNN stage 2	27.0	33.7	29.7
Cascade R-CNN all stage	**28.3**	**34.5**	**30.8**

Table A3: Performance of Cascade R-CNN at different stages.

Method	Split	AP	AP_{50}	AP_{75}
Cascade R-CNN [2] + Norm Linear Layer [17]	*val*	42.5	49.1	44.9
	test	43.1	49.7	45.6
DINO [19]	*val*	42.0	46.8	43.9
	test	42.4	47.2	44.3

Table A4: Performance differences between *val* and *test* split.

Dataset	AP	CLS ER	LOC ER
LVIS [8]	**62.2**	7.69	**4.78**
V3Det	49.4	**25.32**	0.75

Table A5: Comparison of classification error (CLS ER) and localization error (LOC ER) of EVA [6] on LVIS [8] and V3Det.

Performance Differences Between *Val* and *Test* Split. In Table A4, we evaluate Cascade R-CNN [2] + Norm Linear Layer [17] and DINO [19] with 24 epochs, AdamW optimizer and Swin-B backbone on *val* and *test* split of V3Det, showing the annotation consistency of the two splits.

EVA Error Analysis. Table 5 in the main text shows the AP of EVA [6] on V3Det is 49.4, which is 12.8 lower than the AP on LVIS [8], which is 62.2. In this section, we explore the error source of the AP difference. Table A5 compares the classification and localization errors of EVA on LVIS and V3Det, computed by TIDE [1]. Classification error indicates localized correctly (IoU > 0.5) but classified incorrectly; Localization error indicates classified correctly but localized incorrectly (IoU < 0.5). We can see the localization error of V3Det (0.75) is lower than LVIS (4.78), but the classification error of V3Det (25.32) is much higher than LVIS (7.69). This confirms V3Det exposes a more challenging vast vocabulary classification problem than LVIS, leading to a broader exploration space.

Hierarchy Open Vocabulary Test. To comprehensively

Method	AP^v	AP^h	H-score
Detic [22]	7.25	16.90	42.90
RegionCLIP [21]	5.58	11.70	47.69

Table A6: The results of the hierarchy open vocabulary test.

evaluate the open-vocabulary detectors, we propose a hierarchy open vocabulary test to evaluate the hierarchy capability of the detector, which is built upon our hierarchy category organization. Firstly, we employ two distinct methodologies for assessing the Average Precision metric of non-leaf nodes. One is Vocabulary based Non-leaf Average Precision AP^v, defined as

$$AP^v = \frac{1}{N} \sum_{i \in \{\text{non-leaf nodes}\}} AP(P_i, f_i(Y)), \quad f_i(Y) = \{y | y \in Y \text{ and } y \in \text{descendants of node } i\}, \tag{1}$$

where N is total number of non-leaf nodes in the hierarchy tree. P_i is the predicted boxes of non-leaf node i. Y is the ground truth of all leaf nodes. $f_i(Y)$ is all ground truth of the descendants of node i. $AP(\alpha, \beta)$ is the Average Precision between boxes set α and β. The other is Hierarchy based Non-leaf Average Precision AP^h, defined as

$$AP^h = \frac{1}{N} \sum_{i \in \{\text{non-leaf nodes}\}} AP(f_i(P), f_i(Y)), \quad f_i(P) = \{p | p \in P \text{ and } p \in \text{descendants of node } i\}, \tag{2}$$

where P is the predicted boxes of all leaf nodes. $f_i(P)$ is the merged predictions of the descendants belonging to node i processed by NMS with IoU threshold of 0.5.

Based on AP^v and AP^h, we design Hierarchy Score, dubbed as H-score, which is defined as

$$\text{H-score} = AP^v / (AP^h + \epsilon) \times 100, \tag{3}$$

where ϵ is set to 1e-6. A higher value of the H-score indicates a stronger hierarchical capability of the detector. Table A6 provides the results of the hierarchy open vocabulary test. Although the AP^v and AP^h of RegionCLIP are lower than that of Detic, the elevated H-score of RegionCLIP in comparison to Detic suggests a superior hierarchical capability, owing to the robustness of the leveraging CLIP[14] parameters rather than merely extracted text embeddings.

D. Hierarchy Category Organization.

A more detailed hierarchy category organization of V3Det is shown in Figure A1.

E. Coarse Category List.

Table A7 gives the details of the coarse-grained categories used during the annotation process, including the name, total number of fine-grained categories in each coarse-grained category, and some fine-grained examples for each coarse-grained category.

F. Category Descriptions Examples.

In Table A8, we show examples of category descriptions in V3Det, which is written by human experts and chatgpt.

G. More Dataset Visualizations.

Figure A2 and Figure A3 provide some sampled images with annotations for visualization.

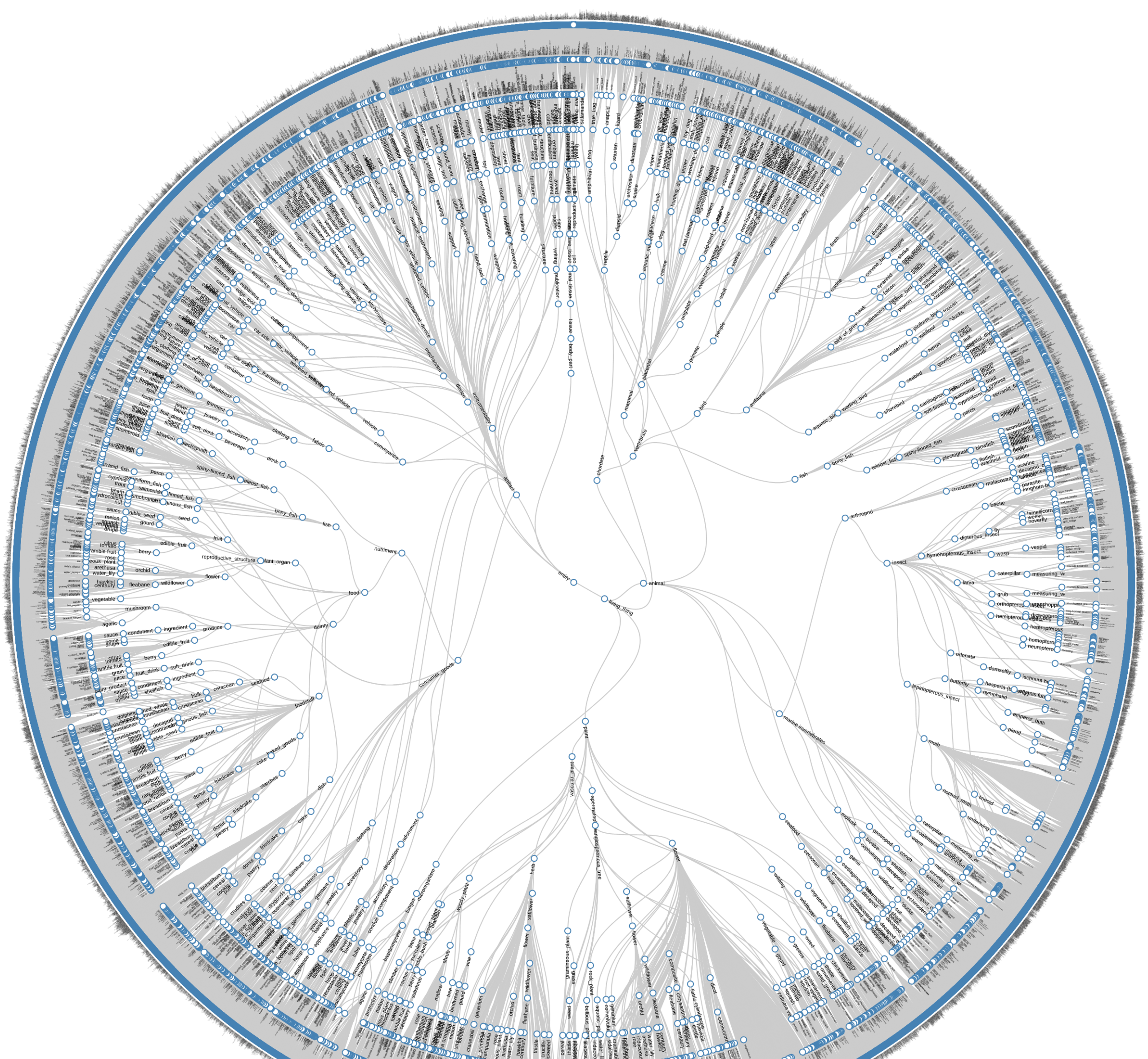
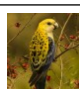
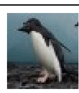
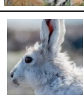
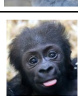


Figure A1: **A Detailed Visualization of Hierarchy Category Organization in V3Det.**

	Coarse category	Numbers	List of fine-grained category
Coarse 1	Animal & Human	7485	siberian tiger; masai lion; net-winged insects; rugby player; man; woman; ...
Coarse 2	Device	241	ceiling fan; tablet computer; display device; mobile device; ipad; game controller; ...
Coarse 3	Table & Chair	30	armchair; conference table; dining table; folding chair; gateleg table; ...
Coarse 4	Flower	1911	hybrid tea rose; floribunda; tagetes; alpine aster; hawaiian hibiscus; orange lily; ...
Coarse 5	Vegetables & Beans & Fruit & Peel	696	squash; kiwifruit; saba banana; calamondin; matoke; eastern prickly pear; ...
Coarse 6	Dishes & Meat & Staple food & Aggs Bean products & Aggs	602	chicken meat; fried noodles; turkey meat; hot dog bun; sliced bread; soba; ...
Coarse 7	Wearable Items	352	baseball uniform; knit cap; martial arts uniform; dog clothes; diving mask; ...
Coarse 8	Fungus	301	medicinal mushroom; lingzhi mushroom; pleurotus eryngii;russula fragilis; ...
Coarse 9	Vehicle	223	microvan; general motors; steam car; solar vehicle; hoverboarding; ambulance; ...
Coarse 10	Sports equipment w/o ball	168	skateboard truck; freebord; training bench; pommelhorse; dart; ...
Coarse 11	Drinks & Seasoning & Oil & Dairy products & Liquid chocolate	113	wiener melange; mocaccino; cream liqueur; ice cream; soy sauce; ...
Coarse 12	Musical instrument	109	drum stick; tabla; banjo; cymbal; drums; guitar; harp; piano; shekere; ...
Coarse 13	Hand & Ignition tools	98	hand fan; alligator wrench; bolt cutter; cap opener; corkscrew; forceps; ...
Coarse 14	Hardware gadgets	66	keychain; plumbing fitting; threading needle; anchor; anvil; awl; bodkin; ...
Coarse 15	Weapon w/o hacking and cutting	62	barbette carriage; battering ram; bomb; brass knucks; bullet; cannon; ...
Coarse 16	Cards & Paper products & Board	62	payment card; academic certificate; blackboard; bookmark; clipboard; doorplate; ...
Coarse 17	Cutting tool & Chop cold weapons	54	axe; bucksaw; crosscut saw; hedge trimmer; knife; cheese cutter; peeler; ...
Coarse 18	Tableware w/o electricity	53	wine glass; stemware; barbecue grill; food steamer; cup; chopstick; fork; ...
Coarse 19	Building & Tent	51	residential; water castle; portable toilet; cabana; bell tent; cenotaph; ...
Coarse 20	Personal care	50	facial cleanser; baby powder; bottlebrush; condom; curler; hairbrush; ...
Coarse 21	Ball	48	basketball; bowling ball; cricket ball; croquet ball; golf ball; handball; ...
Coarse 22	Cactus	39	san pedro cactus; large-flowered cactus; ferocactus cylindraceus; ...
Coarse 23	Window & Door & Delivery hole & Well	38	window covering; water well; artesian well; doorknob; dormer; gusher; ...
Coarse 24	Cloth & Cloth material	36	folding napkins; bath mat; beach towel; cleaning pad; dustcloth; doily; ...
Coarse 25	Office equipment	37	board eraser; abacus; chalk; fountain pen; marker; pencil; inkstone; ...
Coarse 26	Cleaning tools & Nozzle	35	bathroom sink; toilet roll holder; irrigation sprinkler; bathtub; broom; dumpster; ...
Coarse 27	Measuring equipment	28	beaker; compass; detector; divider; plumb bob; protractor; triangular ruler; ...
Coarse 28	Shelf & Storage cabinet	28	wine rack; shoe organizer; pot rack; bookcase; clothes tree; coatrack; ...
Coarse 29	Gem & Fountain	28	fountain; crystal; diamond; ruby; pearl; emerald; chrysoberyl; jadeite; ...
Coarse 30	Faith related objects	27	amulet; christian cross; flag; shoulder board; totem pole; medal; ...
Coarse 31	Medical related objects	26	aspirator; catheter; hypodermic needle; pill; plaster; stethoscope; ...
Coarse 32	Pole & Tube	24	blowgun; chimney; cigar; fire hose; grab bar; meerschaum; test tube; ...
Coarse 33	Lifesaving objects	21	baby float; air cushion; breeches buoy; fire hydrant; life buoy; water wings; ...
Coarse 34	Bottles & Bags & Buckets & Boxes	22	weaving basket; bag; barrel; basin; bottle; briefcase; pot; rain barrel; ...
Coarse 35	Candy & Solid chocolate	21	brittle; chewing gum; candied apple; cocoa powder; lollipop; jello; ...
Coarse 36	Planet & Satellite	18	satellite; meteorite; moon; sun; black hole; Earth; Mars; Mercury; Venus; ...

	Coarse category	Numbers	List of fine-grained categories
Coarse 37	Bedding	18	infant bed; bed; bed pillow; bunk bed; carrycot; crib; futon; hammock; ...
Coarse 38	Chess	16	chessman; chess set; Chinese Chess; Go; International Draughts or Checkers; Shogi; ...
Coarse 39	Bell	12	electric bell; timer; weathervane; wind chime; fire alarm; Windbell; bicycle bell; ...
Coarse 40	Signpost & Roadblock	12	stop sign; crosswalk sign; billboard; pedestrian crossing; yard marker; ...
Coarse 41	Control device & Heating appliances & Hot-water bag w/o electricity	12	brazier; gearshift; handwheel; hot-water bottle; radiator; roaster; ...
Coarse 42	Wheel shaped objects	11	bicycle wheel; bobbin; ferris wheel; gear; inner tube; pulley; automotive tire; ...
Coarse 43	Lens	10	magnifying glass; microscope; telescope; Rifle scopes; Triangular Prism; ...
Coarse 44	Currency & Whistle	8	whistle; gold; money; cash; paper money; coinage; banknote; ...
Coarse 45	Animal nest	9	birdcage; chicken coop; rabbit hutch; nest; wasp nest; cage; ...
Coarse 46	Umbrella & Ladder	9	parachute; cocktail umbrella; Aluminum alloy ladder; Wooden ladder; ...
Coarse 47	Electronic component	6	battery; capacitor; coil; resistor; solar cell; electronic component; ...
Coarse 48	Socket & Plug	6	bung; cork; power outlet; socket; wall socket; ...
Coarse 49	Lure & Aquarium & Post box	6	aquarium; fishbowl; pillar box; Nano aquarium; Spoon lure; Penfold post box; ...
Coarse 50	Industrial machine & Spray paint	5	concrete mixer; crane; generator; spray paint; pumpjack; ...
Coarse 51	Spring & Magnet & Compass	4	compass; rubber band; spring; refrigerator magnet; ...
Coarse 52	Popcorn machine ATM & Ashtray & Incense burner	5	ashtray; automated teller machine; censer; popper; incense burner; ...
Coarse 53	Model	93	toy vehicle; rubik's cube; amphora; armillary sphere; cockhorse; doll; ...

Table A7: **The details of the coarse-grained categories.** All fine-grained categories are divided into 53 coarse-grained categories. 'Numbers' denotes total number of fine-grained categories in each coarse-grained category.

Name	Image	Type	Description
platycercus adscitus		Experts	platycercus adscitus has a brown head, yellow belly and green wings
		chatgpt	green body, yellow head, red markings (pale-headed rosella)
Adelie penguin		Experts	medium-sized penguins occurring in large colonies on the Adelie Coast of Antarctica
		chatgpt	the Adelie penguin has a black head and back, white front, and a distinctive white ring around the eye
polar bear		Experts	compared to other bears, smaller head, round ears, slender neck. Black skin, huge and ferocious.
		chatgpt	large, white-furred bear, with a long neck, a stocky body, powerful legs, and sharp claws
pembroke		Experts	the smaller and straight-legged variety of corgi having pointed ears and a short tail
		chatgpt	small, sturdy dog breed, with erect ears, a foxy face, and a docked tail (or naturally bobtail)
bengal tiger		Experts	yellow-orange coats with dark stripes, white belly and limbs, and an orange tail with black rings
		chatgpt	orange-brown coat with black stripes, white belly, and distinctive white markings above the eyes
polar hare		Experts	a large hare of northern North America; it is almost completely white in winter
		chatgpt	large, white-furred hare with long ears and strong hind legs, found in Arctic regions
gorilla		Experts	the gorilla is strong and hairless on its face and ears. it has a high forehead and protruding jawbone
		chatgpt	large, muscular primate with black or dark brown fur, a broad chest, and a prominent brow ridge

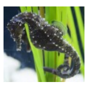
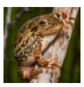
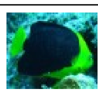
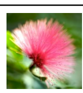
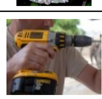
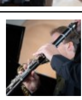
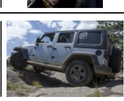
Name	Image	Type	Description
African elephant		Experts	an elephant native to Africa having enormous flapping ears and ivory tusks
		chatgpt	massive, grey mammal with a long trunk, large ears, curved tusks, and wrinkled skin
charolais		Experts	this breed of cattle is heavy, with weighing 700-1650 kg. they have white/cream coats and pink noses
		chatgpt	large, white-colored breed of cattle with a muscular build, broad forehead, and short horns
Grevy's zebra		Experts	it resembles a mule with a big head, elongated nostril openings, large round conical ears, and tall erect mane
		chatgpt	large, wild equid with black and white stripes, a white belly, and a tall, erect mane
box turtle		Experts	it has a domed shell hinged at the bottom, allowing them to close it tightly and escape predators
		chatgpt	small, domed shell with brown or olive-colored skin and a pattern of yellow or orange spots or lines
tuatara		Experts	a reptile with a greenish-grey lizard-like appearance found only on certain small islands near New Zealand
		chatgpt	a reptile with a spiny crest, two rows of teeth, and a third "eye" on the top of its head
northern seahorse		Experts	northern seahorses have long snouts and a curly tail, and can grow up to 8 inches in length
		chatgpt	the northern seahorse has a small body covered in bony plates, a long snout, and a curled tail
lowland burrowing treefrog		Experts	terrestrial burrowing nocturnal frog of grassy terrain and scrub forests having very hard upper surface of head
		chatgpt	small, round-bodied frog with smooth skin, adapted for burrowing with short legs and a pointed snout
rock beauty		Experts	predominately black. The head and front half portion of the body, and the caudal fin are a bright yellow
		chatgpt	bright yellow head, tail, pectoral fins; black body, dorsal, anal fins; abrupt color transitions
clingfish		Experts	very small (to 3 inches) flattened marine fish with a sucking disc on the abdomen for clinging to rocks etc
		chatgpt	the clingfish has a flattened body with a suction cup-like pelvic disc, and ranges in color from brown to green
porpoise		Experts	whales distinguishable from dolphins by their more compact build, smaller size, and curved blunt snout with spatulate teeth
		chatgpt	small, grey marine mammal with a rounded head, a small dorsal fin, and a curved mouth
manta		Experts	extremely large pelagic tropical ray with triangular pectoral fins, horn-shaped cephalic fins and large, forward-facing mouths
		chatgpt	large, flat-bodied, wing-like fins, no tail, black or dark upper side, white or light underside
aglais		Experts	the wings of aglais are rusty red with a unique eyespot in black, blue, and yellow at each wingtip
		chatgpt	eyespots on wings, blue-green-brown hues (European peacock butterfly)
woodland sunflower		Experts	a wildflower native to the United States, which has bright yellow petals that surround a dark brown center
		chatgpt	tall yellow flower with numerous thin petals surrounding a dark center disk, and green leaves
calliandra		Experts	the flowers are produced in cylindrical or globose inflorescences and have numerous long slender stamens
		chatgpt	small shrub with fern-like leaves and vibrant, pink, powder-puff shaped flowers
cistus salviifolius		Experts	a shrub with fragrant, silver-green foliage and white flowers
		chatgpt	a tall, bright yellow or orange flower often grown for its edible seeds and as an ornamental plant
power drill		Experts	a hand tool with a rotating chuck driven by an electric motor for drilling
		chatgpt	handheld tool with a motor and rotating chuck for drilling holes and fastening screws, often with a pistol grip
clarinet		Experts	the clarinet is a single-reed instrument with a nearly cylindrical bore and a flared bell.
		chatgpt	narrow, cylindrical woodwind instrument with a mouthpiece, reed, and numerous keys for playing different notes
off-road vehicle		Experts	the off-road vehicles typically have four-wheel-drive, increased suspension, and large tires
		chatgpt	large, sturdy vehicle with high ground clearance, wide tires, and typically four-wheel drive for use on unpaved terrain

Table A8: Examples of category descriptions of V3Det.

Figure A2: **Visualizations of annotations in complex scenes.** Each box is paired with category name and box size.

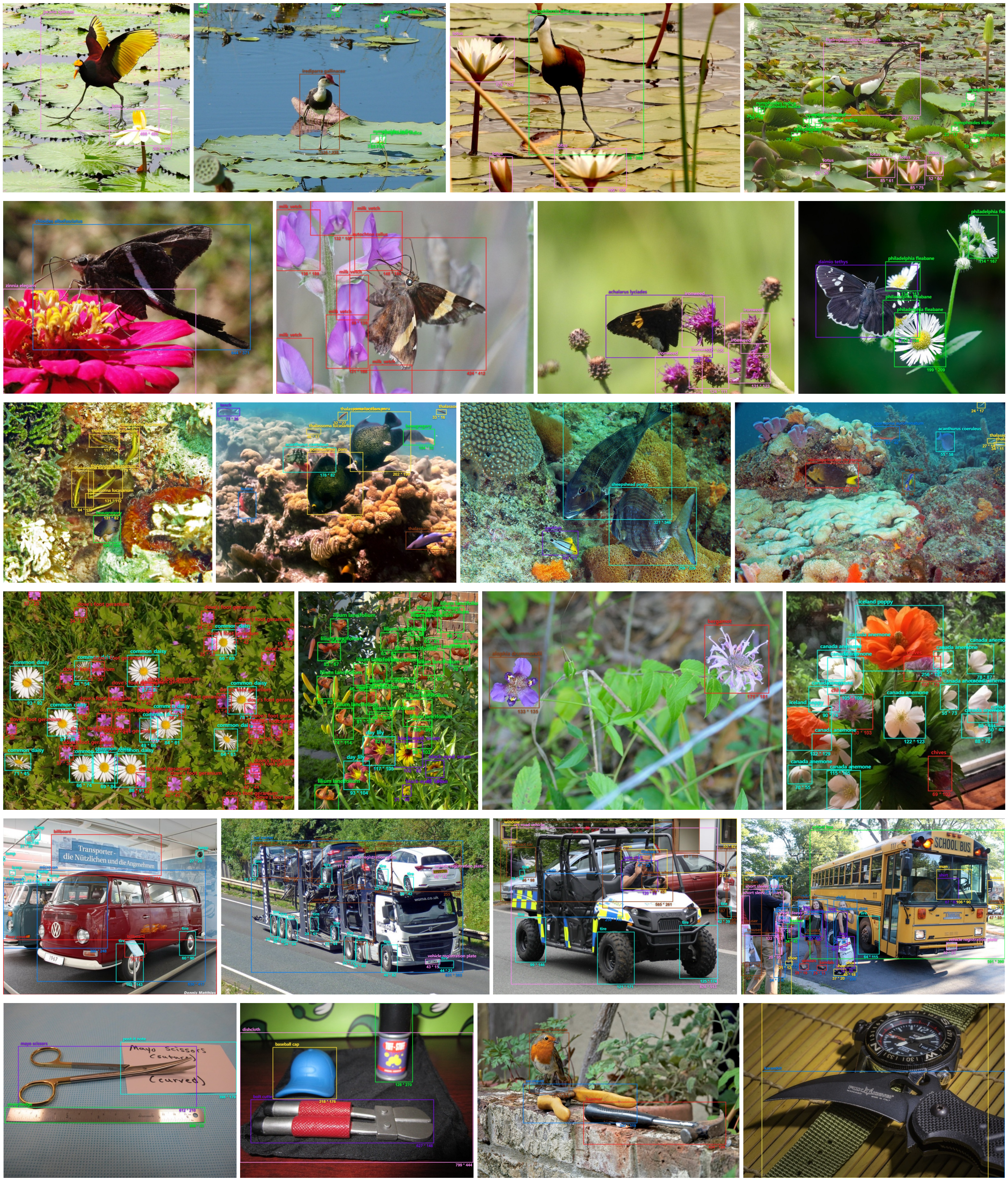


Figure A3: **Visualizations of annotations in fine-grained categories.** Each row shows a set of visually similar category annotations that are easily confused.

References

[1] Daniel Bolya, Sean Foley, James Hays, and Judy Hoffman. Tide: A general toolbox for identifying object detection errors. In *ECCV*, 2020. 2

[2] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: High quality object detection and instance segmentation. *CVPR*, 2018. 1, 2

[3] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. Hybrid task cascade for instance segmentation. In *CVPR*, 2019. 1

[4] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 1

[5] detrex contributors. detrex: An research platform for transformer-based object detection algorithms. `https://github.com/IDEA-Research/detrex`, 2022. 1

[6] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. *arXiv*, 2022. 2

[7] 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 *CVPR*, June 2021. 1

[8] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 1, 2

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1, 2

[10] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv*, 2014. 1

[11] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 1

[12] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 1

[13] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2018. 1

[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 *ICML*. PMLR, 2021. 3

[15] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 1, 2

[16] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: Fully convolutional one-stage object detection. *CoRR*, abs/1904.01355, 2019. 1

[17] Jiaqi Wang, Wenwei Zhang, Yuhang Zang, Yuhang Cao, Jiangmiao Pang, Tao Gong, Kai Chen, Ziwei Liu, Chen Change Loy, and Dahua Lin. Seesaw loss for long-tailed instance segmentation. In *CVPR*, 2021. 2

[18] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019. 1

[19] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Harry Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In *ICLR*, 2022. 1, 2

[20] Yuanhan Zhang, Qinghong Sun, Yichun Zhou, Zexin He, Zhenfei Yin, Kun Wang, Lu Sheng, Yu Qiao, Jing Shao, and Ziwei Liu. Bamboo: Building mega-scale vision dataset continually with human-machine synergy. *arXiv*, 2022. 1

[21] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *CVPR*, 2022. 1, 3

[22] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, 2022. 1, 3

[23] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Probabilistic two-stage detection. *arXiv*, 2021. 1, 2

[24] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv*, 2020. 1