Learning Visual Hierarchies in Hyperbolic Space for Image Retrieval

1. Hierarchy Tree

Our part-based image hierarchy framework is designed to be widely applicable to general image datasets with bounding box annotations. In this section, we outline the key implementation involved in constructing hierarchy trees.

We introduce a general method to generate dataset-specific ground truth hierarchy trees based on data statistics. Following the approach outlined in Sec. 3.1 of the main paper, we begin by identifying bounding box pairs with substantial overlap. In this work, we define a bounding box pair when at least 80% of the smaller bounding box b is contained within either the full image or a larger bounding box. We initially set the overlap threshold at 100% and empirically adjusted it by evaluating the hierarchy's validity using text labels. For the OpenImages dataset, 80% was found to be a suitable threshold. This is a design choice and can be adapted for other datasets.

These pairs are then filtered based on two criteria: a *frequency* threshold and a *proportion* threshold. For each pair, we record the *frequency* of occurrences (*e.g.*, bicycle-to-wheel relationships) and calculate the *proportion* of instances where a child class appears within a parent class (*e.g.*, the percentage of bicycle bounding boxes containing a wheel bounding box). Only pairs meeting both criteria, frequent occurrence and consistent labeling, are preserved. We choose *frequency* $= 50$ and *proportion* $= 10\%$ in our experiments.

Once entailment pairs are established, they are organized into hierarchical trees (see examples in Fig. 1). In the evaluation of hierarchal image retrieval, the order of the hierarchy tree is essential: for parent-to-child retrieval, lower-level concepts below the child in the hierarchy tree are considered correct, while for child-to-parent retrieval, higher-level concepts above the input classes are correct.

2. Experiment Details

2.1. Hyperparameters and training details

We employ the AdamW optimizer with parameters $(\beta_1, \beta_2) = (0.9, 0.999)$ and a learning rate of 2×10^{-5}. Training was conducted using $8 \times$ A10G Nvidia GPUs. For each model, we used the largest batch size that fit in memory: CLIP ViT was trained with a total effective batch size of 320, and MoCo-v2 with a total effective batch size of 800. Each model was fine-tuned for a single epoch on *HierOpenImages* dataset, taking approximately 26 hours for CLIP ViT and 18 hours for MoCo-v2. The embeddings are projected to dimension 128 in the final layer. The hyperbolic model has a learnable curvature parameter.

During training, we filter out bounding boxes that occupy less than 1% of the full image area, as well as pairs involving small bounding boxes labeled as 'IsGroupOf' objects in the bounding box-to-bounding box relationships. For data augmentation, we apply randomly horizontal flip (20%), vertical flip(20%), rotate (degree =15), color jitter (brightness=0.2, contrast=0.2, saturation=0.2, hue=0.1), Gaussian blur (kernel size=5, $\sigma = (0.3, 1.5)$), and then resize each image to 224×224.

2.2. Part-based Image Retrieval

Data. *HierOpenImages* is built from the OpenImages dataset [3], which originally contains approximately 1.9 million images, 14 million bounding boxes and 600 labels. We create image-to-bounding box pairs, including one cross-image bounding box

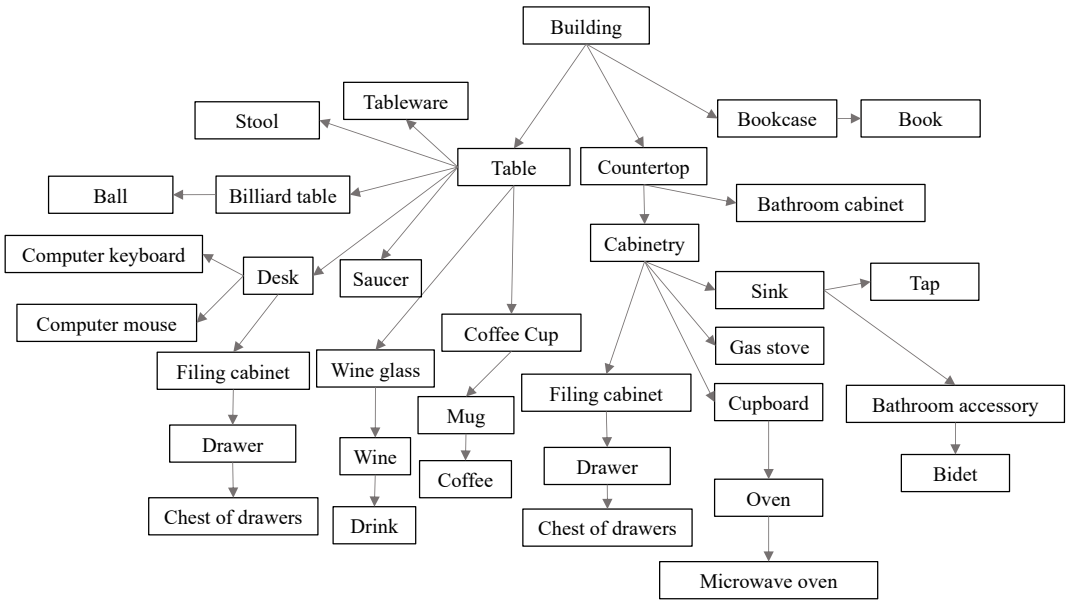

Figure 1. **Example of a subset of hierarchical trees extracted from the OpenImages dataset.**

sample for each bounding box class in the image, and bounding box to bounding box pairs where at least 80% of the smaller bounding box is contained in the larger bounding box.

During part-based image retrieval evaluation, we filter out bounding boxes for very small or large objects, and filter out noisy samples (*e.g.,* wrong labels, highly occluded, *etc*). For part-based image retrieval, we randomly select a subset of 10,000 full images and 10,000 bounding box images as query images, using the entire filtered test set as candidates. The top-50 class frequency distributions for both the query and candidate sets are shown in Fig. 2. Although the query and candidate set distributions are similar, the class distribution is highly imbalanced, highlighting the importance of hierarchical retrieval evaluation using combined precision-recall curves and OT distances (Sec. 3.3; see Fig. 4 and Table 2 in the main paper).

Same-Class Retrieval. For part-to-full retrieval, a retrieval is considered correct if the retrieved full image contains the same object class as the query bounding box image. For full-to-part retrieval, a retrieval is correct if the retrieved bounding box image corresponds to an object class within the query full image.

Hierarchical Retrieval. From a query parent image, correct child classes are all classes located at the lower level on the hierarchy tree of the labeled classes of the parent image (see examples in Fig. 1). For instance, when querying with a high-level full image, such as an image of a car, we expect to retrieve lower-level bounding boxes associated with the car, such as the car mirror, wheel or car plate *etc*. Conversely, when querying with a bounding box image, such as a wheel, we expect to retrieve various types of higher-level full images that include wheels, like cars, bicycles or cyclists *etc*.

Hierarchal Retrieval Evaluation To compute the optimal transport (1-D Wasserstein) distance between the retrieved label distribution and the ground truth (Table 2 and Fig. 4 in the main paper), we construct the ground truth distribution based on the fre-

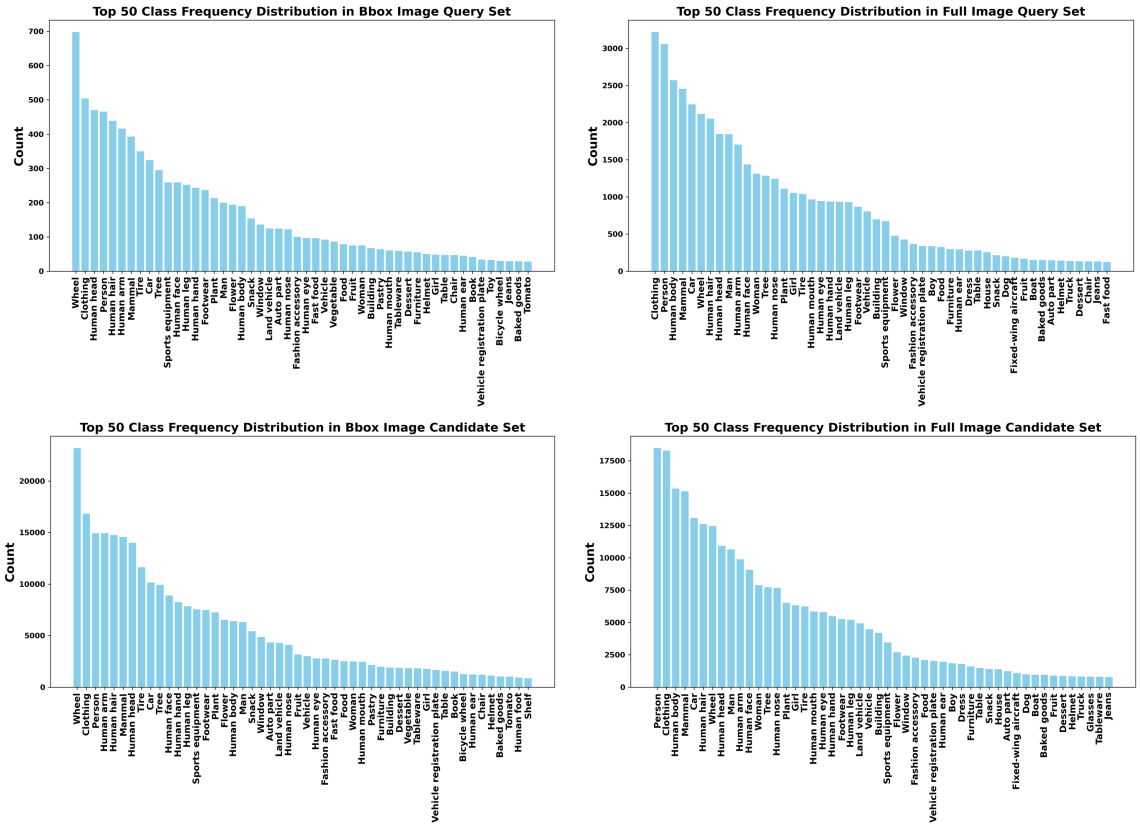

Figure 2. **Bounding box class distribution in the candidate and query sets.**

quency of each class (and its hierarchical classes) in the query parent image. We count the occurrences of each class in the candidate set and build the ground truth distribution by normalizing the frequencies to sum to 1. Similarly, the retrieval distribution is built by counting and normalizing retrieved class occurrences, and assigning any retrieved classes outside the ground truth hierarchy tree to an 'others' class. The two distributions are aligned by class order (ground truth distribution is zero in the 'others' class), and the 1-D Wasserstein distance is computed using the `scipy` library.

Note that the Wasserstein distance has a closed-form formula for 1-D data. If P and Q are represented as discrete empirical distributions (*e.g.*, histograms or sorted samples of size n), let $\{x_1, x_2, \ldots, x_n\}$ to be sorted values P, and $\{y_1, y_2, \ldots, y_n\}$ to be sorted values from Q, then the 1-D Wasserstein distance is:

$$W_p(P, Q) = \left(\frac{1}{n} \sum_{i=1}^{n} \|x_i - y_i\|^p \right)^{1/p},$$

where p refers to the order of the distance in the general p-Wasserstein metric.

In Table 2 of the main paper, we retrieve the Top-K results starting from a large K. This is necessary because each image contains an average of 5.29 distinct classes and 61.5 classes across hierarchical trees, requiring a large number of retrievals to accurately evaluate the distribution.

Image Retrieval Interface via Gradio. We built our image retrieval interface using Gradio [1], as

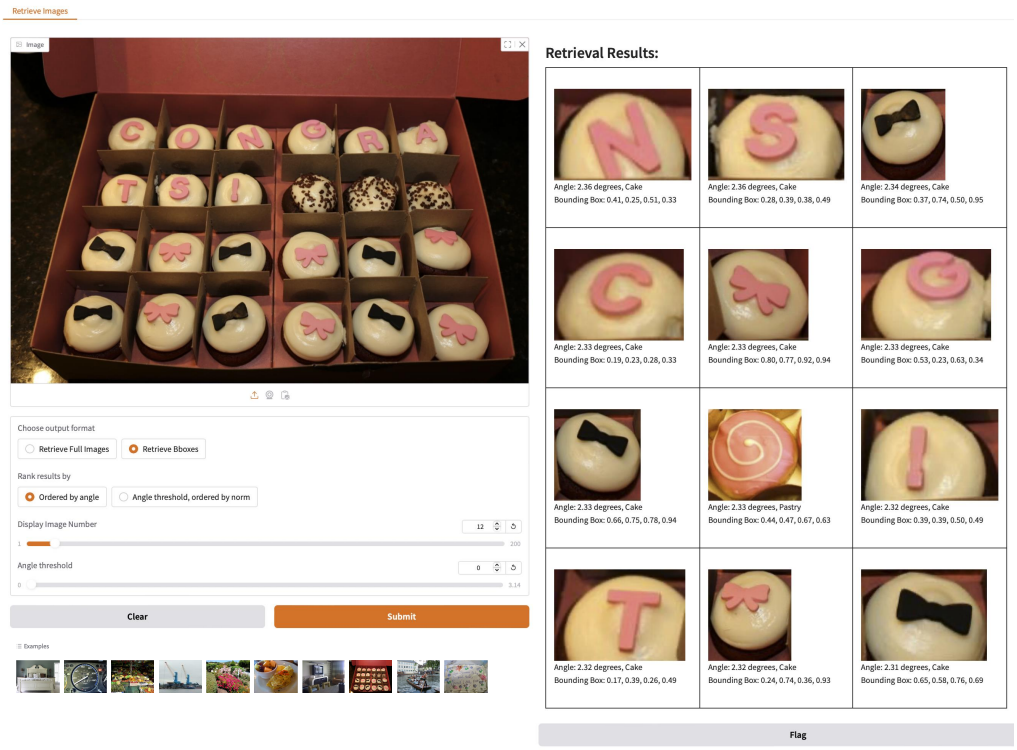

Figure 3. **Example of our image retrieval interface built with Gradio [1].** The interface supports image selection, upload and retrieves results ranked by user-defined modes. Its modular design allows for easy integration of additional functionalities.

shown in Fig. 3. Input images can be selected from a linked image folder, where thumbnail images are displayed in an image gallery, or they can be directly uploaded by users. The retrieval results can be sorted by the hyperbolic angle relative to the input image or filtered using a user-defined threshold value, after which the results are ordered by their embedding norms. Additional functionalities can be easily integrated into the current pipeline.

2.3. Generalization Evaluation

LVIS Dataset. We evaluate the generalization capability of our model on the out-of-domain LVIS dataset [2], which is designed for large-scale ($\sim$1.2M bounding boxes) long-tail instance classification and segmentation. It has a highly imbalanced distribution of 1,203 object categories and contains 897 object categories that are absent from OpenImages [3]. Here are some examples of unseen hierarchies in LVIS [2] but not in OpenImges [3]: {table $\rightarrow$ tablecloth $\rightarrow$ ashtray $\rightarrow$ cigarette}, {backpack $\rightarrow$ strap $\rightarrow$ belt buckle}, {toy $\rightarrow$ teddy bear $\rightarrow$ thread $\rightarrow$ bobbin}, {sofa $\rightarrow$ blanket $\rightarrow$ quilt $\rightarrow$ bedspread}. The full list of these categories can be found in the appendix. Only display the first class of synonyms.

We construct the hierarchical evaluation set from the validation set of the LVIS dataset [2], following the similar process as constructing *HierOpenImages*. To construct the ground truth hierarchy tree, we use the pipeline as described in Sec. 1. We empirically choose parameter $\mathit{frequency} = 5$ and $\mathit{proportion} = 5\%$ in our experiments. This adjustment is necessary because the LVIS dataset [2] has very unbalanced

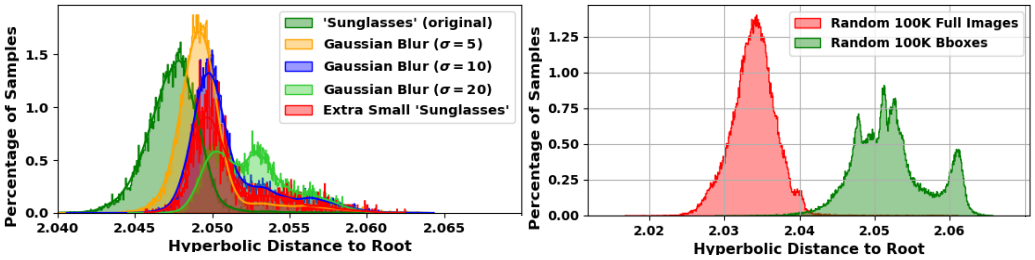

Figure 4. **Example of the model's emergent structure.** In the left plot, the percentage of samples of the original *Sunglasses* crops are shown in dark green. Applying Gaussian blur or selecting extremely small crops shifts the embeddings further from the root. The right figure illustrates a clear separation between full-image and part-image embeddings.

classes.

Emergent structure. We observe that high-frequency, ambiguous image crops are naturally pushed toward the boundary of hyperbolic space, where more angular entailment constraints can be better satisfied. This aligns with the exponentially growing volume near the boundary, and it is an efficient solution for representing diverse, overlapping semantics. In Figure 4, in the left plot, the percentage of samples of the original *Sunglasses* crops are shown in dark green. Adding Gaussian blur or selecting extremely small crops shifts their embeddings further from the root. The small crops (red curve) are not included in other curves. The right plot shows clear separation between full and part image embeddings.

3. Ablation Studies

In this ablation study, we evaluate the impact of cross-image scene-to-object samples. We fine-tuned the CLIP ViT model solely on hierarchical part-based entailment data within individual images (entailment pairs with high visual similarity), excluding any cross-image image-to-bounding-box samples. Fig. 4 in the main paper shows that the hyperbolic model CLIP-hyp† fine-tuned without cross-image samples significantly outperforms the Euclidean model CLIP-euc† and even surpasses the model trained with additional cross-image samples. This indicates that training in hyperbolic space enhances

Model	Cross-Image	Top-5	Top-10	Top-50	Top-100
Child-to-Parent					
CLIP-hyp†	✓	**73.37**	**72.59**	**69.84**	**68.62**
CLIP-euc†	✓	67.83	68.37	67.12	66.04
CLIP-hyp†	✗	70.47	70.29	67.89	66.38
CLIP-euc†	✗	64.98	64.49	63.51	63.10
CLIP	-	26.73	25.95	23.69	22.70
Parent-to-Child					
CLIP-hyp†	✓	**66.02**	**66.63**	**66.50**	**65.91**
CLIP-euc†	✓	65.38	65.70	66.01	65.79
CLIP-hyp†	✗	63.00	63.78	63.65	63.44
CLIP-euc†	✗	60.33	60.73	61.43	61.38
CLIP	-	47.52	46.60	43.50	42.08

Table 1. **Part-based same-class image retrieval evaluation.** Cross-image ✓indicates models fine-tuned on entailment relationships both within and across images at the category level, while ✗ represents models fine-tuned without cross-image sampling. The evaluation setup is the same as Table 1 in the main paper. The best results are marked in bold, and the second-best results are in blue.

Metrics	Model	Cross-Image	Top-150k	Top-200k	Top-250k
Recall % ↑	CLIP-hyp†	✓	**74.06**	**82.34**	**88.79**
	CLIP-euc†	✓	72.96	80.98	87.97
	CLIP-hyp†	✗	73.52	81.58	88.10
	CLIP-euc†	✗	71.34	78.89	86.04
	CLIP	-	51.33	64.28	77.02
OT Distance ↓	CLIP-hyp†	✓	**16.36**	**19.27**	**22.21**
	CLIP-euc†	✓	17.00	20.00	22.45
	CLIP-hyp†	✗	16.95	19.77	22.55
	CLIP-euc†	✗	18.43	21.25	23.53
	CLIP	-	23.81	24.60	25.03

Table 2. **Part-based hierarchical evaluation of parent-to-child image retrieval.** The OT distance is defined in Sec. 3.3 (main paper), and the evaluation setup follows Table 2 (main paper). The best results are marked in bold, and the second-best results are in blue.

the model's ability to recognize visually dissimilar entailment pairs.

In this supplementary material, we further evaluate these models on part-based same-class and hierarchical image retrieval tasks, as shown in Table. 1-2. The best results are highlighted in bold and the second-best results are shown in blue. The results clearly indicate that cross-image sampling improves image retrieval performance. Notably, the hierarchical retrieval results align with Fig. 4 in the main paper. Specifically, the hyperbolic model trained without cross-image samples outperforms the Euclidean model trained with cross-image samples in 5 out of 6 cases, as shown in Table 2.

4. More Qualitative Results

More qualitative parent-to-child retrieval results are shown in Fig. 5 to visualize the latent space. Bounding box images are filtered by angle (CLIP-hyp†) or cosine similarity (CLIP ViT model) thresholds and sorted by increasing embedding norms. Our hyperbolic model retrieves diverse and visually distinct lower-level objects related to the query images, organized according to the predefined *scene-object-part* hierarchy in the embedding space.

References

[1] Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. Gradio: Hassle-free sharing and testing of ml models in the wild. *arXiv preprint arXiv:1906.02569*, 2019. 3, 4

[2] 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. 4

[3] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020. 1, 4

Figure 5. **Example of parent-to-child retrieval using CLIP ViT and our CLIP-hyp† model.** Results are ordered by ascending embedding norms. Our model retrieves images matching the predefined *scene-object-part* hierarchy, placing high-level objects near the origin (*e.g.*, group of fruits → single fruits), and grouping semantically related but visually distinct objects (*e.g.*, chairs & TVs). All retrieved bounding box images are scaled to the same ratio.

Appendix

Object classes in LVIS dataset but not in OpenImage dataset

aerosol can	air conditioner	alcohol	alligator	almond	amplifier
anklet	antenna	applesauce	apricot	apron	aquarium
arctic	armband	armchair	armoire	armor	trash can
ashtray	asparagus	atomizer	avocado	award	awning
baboon	baby buggy	basketball back-board	bagpipe	baguet	bait
ballet skirt	bamboo	Band Aid	bandage	bandanna	banner
barbell	barrette	barrow	baseball base	baseball	baseball cap
basket	basketball	bass horn	bat	bath mat	bath towel
bathrobe	batter	battery	beachball	bead	bean curd
beanbag	beanie	bedpan	bedspread	cow	beef
beeper	beer bottle	beer can	bell	belt buckle	beret
bib	Bible	visor	binder	birdfeeder	birdbath
birdcage	birdhouse	birthday cake	birthday card	pirate flag	black sheep
blackberry	blackboard	blanket	blazer	blimp	blinker
blouse	blueberry	gameboard	bob	bobbin	bobby pin
boiled egg	bolo tie	deadbolt	bolt	bonnet	booklet
bookmark	boom microphone	bouquet	bow	bow	bow-tie
pipe bowl	bowler hat	bowling ball	boxing glove	suspenders	bracelet
brass plaque	bread-bin	breechcloth	bridal gown	broach	broom
brownie	brussels sprouts	bubble gum	bucket	horse buggy	horned cow
bulldog	bulldozer	bullet train	bulletin board	bulletproof vest	bullhorn
bun	bunk bed	buoy	business card	butter	button
cabana	cabin car	cabinet	locker	calendar	calf
camcorder	camera lens	camper	candle holder	candy bar	candy cane
walking cane	canister	canteen	cap	bottle cap	cape
cappuccino	railcar	elevator car	car battery	identity card	card
cardigan	cargo ship	carnation	horse carriage	tote bag	carton
cash register	casserole	cassette	cast	cauliflower	cayenne
CD player	celery	chain mail	chaise longue	chalice	chandelier
chap	checkbook	checkerboard	cherry	chessboard	chickpea
chili	chinaware	crisp	poker chip	chocolate bar	chocolate cake
chocolate milk	chocolate mousse	choker	chopstick	slide	cider
cigar box	cigarette	cigarette case	cistern	clarinet	clasp
cleansing agent	cleat	clementine	clip	clipboard	clippers
cloak	clock tower	clothes hamper	clothespin	clutch bag	coaster
coat hanger	coatrack	cock	cockroach	cocoa	coffee maker

coffeepot	coil	colander	coleslaw	coloring material	combination lock
pacifier	comic book	compass	condiment	cone	control
convertible	cooker	cooking utensil	cooler	cork	corkboard
corkscrew	edible corn	cornbread	cornice	cornmeal	corset
costume	cougar	coverall	cowbell	crabmeat	cracker
crape	crate	crayon	cream pitcher	crib	crock pot
crossbar	crouton	crow	crowbar	crucifix	cruise ship
police cruiser	crumb	cub	cube	cufflink	cup
trophy cup	cupcake	hair curler	curling iron	cushion	cylinder
cymbal	dalmatian	dartboard	date	deck chair	dental floss
detergent	diary	dinghy	dining table	tux	dish
dish antenna	dishrag	dishtowel	dishwasher detergent	dispenser	diving board
Dixie cup	dog collar	dollar	dollhouse	domestic ass	doorknob
doormat	dove	underdrawers	dress hat	dress suit	dresser
drill	drone	dropper	drumstick	duckling	duct tape
duffel bag	dumpster	dustpan	earphone	earplug	earring
easel	eclair	eel	egg	egg roll	egg yolk
eggbeater	eggplant	electric chair	elk	escargot	eyepatch
fan	ferret	Ferris wheel	ferry	fig	fighter jet
figurine	file	fire alarm	fire engine	fire extinguisher	fire hose
first-aid kit	fishbowl	fishing rod	flagpole	flamingo	flannel
flap	flash	fleece	flip-flop	flipper	flower arrangement
flute glass	foal	folding chair	footstool	forklift	freight car
French toast	freshener	frisbee	fruit juice	fudge	funnel
futon	gag	garbage	garbage truck	garden hose	gargle
gargoyle	garlic	gasmask	gazelle	gelatin	gemstone
generator	gift wrap	ginger	cincture	glass	globe
golf club	golfcart	gorilla	gourd	grater	gravestone
gravy boat	green bean	green onion	griddle	grill	grits
grizzly	grocery bag	gull	gun	hairbrush	hairnet
hairpin	halter top	ham	hammock	hamper	hand glass
hand towel	handcart	handcuff	handkerchief	handle	handsaw
hardback book	harmonium	hatbox	veil	headband	headboard
headlight	headscarf	headset	headstall	heart	heron
highchair	hinge	hockey stick	home plate	honey	fume hood
hook	hookah	hornet	hose	hot-air balloon	hotplate
hot sauce	hourglass	houseboat	hummingbird	hummus	icecream
popsicle	ice maker	ice pack	ice skate	igniter	inhaler
iron	ironing board	jam	jar	jean	jeep
jelly bean	jersey	jet plane	jewel	jewelry	joystick

jumpsuit	kayak	keg	kennel	key	keycard
kilt	kimono	kitchen sink	kitchen table	kitten	kiwi fruit
knee pad	knitting needle	knob	knocker	lab coat	lamb
lamb-chop	lamppost	lampshade	lanyard	laptop computer	lasagna
latch	lawn mower	leather	legging	Lego	legume
lemonade	lettuce	license plate	life buoy	life jacket	lightbulb
lightning rod	lime	lip balm	liquor	log	lollipop
speaker	machine gun	magazine	magnet	mail slot	mailbox
mallard	mallet	mammoth	manatee	mandarin orange	manager
manhole	map	marker	martini	mascot	mashed potato
masher	mask	mast	mat	matchbox	mattress
meatball	medicine	melon	microscope	milestone	milk can
milkshake	minivan	mint candy	mitten	money	monitor
motor	motor scooter	motor vehicle	mound	mousepad	music stool
nailfile	napkin	neckerchief	needle	nest	newspaper
newsstand	nightshirt	nosebag	noseband	notebook	notepad
nut	nutcracker	oar	octopus	octopus	oil lamp
olive oil	omelet	onion	orange juice	ottoman	overalls
packet	inkpad	pad	padlock	paintbrush	painting
pajamas	palette	pan	pan	pantyhose	papaya
paper plate	paperback book	paperweight	parakeet	parasail	parasol
parchment	parka	passenger car	passenger ship	passport	patty
pea	peanut butter	peeler	wooden leg	pegboard	pelican
pencil	pendulum	pennant	penny	pepper	pepper mill
persimmon	pet	pew	phonebook	phonograph record	pickle
pickup truck	pie	pigeon	piggy bank	pin	pinecone
ping-pong ball	pinwheel	tobacco pipe	pipe	pita	pitcher
pitchfork	place mat	playpen	pliers	plow	plume
pocket watch	pocketknife	poker	pole	polo shirt	poncho
pony	pop	postbox	postcard	pot	potholder
pottery	pouch	power shovel	projector	propeller	prune
pudding	puffer	puffin	pug-dog	puncher	puppet
puppy	quesadilla	quiche	quilt	race car	radar
radiator	radio receiver	raft	rag doll	raincoat	ram
raspberry	rat	razorblade	reamer	rearview mirror	receipt
recliner	record player	reflector	rib	ring	river boat
road map	robe	rocking chair	rodent	roller skate	Rollerblade
rolling pin	root beer	router	rubber band	runner	saddle
saddle blanket	saddlebag	safety pin	sail	salad plate	salami
salmon	salmon	salsa	saltshaker	satchel	saucepan
sausage	sawhorse	scarecrow	school bus	scraper	scrubbing brush

seabird	seaplane	seashell	shaker	shampoo	sharpener
Sharpie	shaver	shaving cream	shawl	shears	shepherd dog
sherbert	shield	shoe	shopping bag	shopping cart	shot glass
shoulder bag	shovel	shower head	shower cap	shower curtain	shredder
signboard	silo	skewer	ski boot	ski parka	ski pole
skullcap	sled	sleeping bag	sling	slipper	smoothie
soap	soccer ball	softball	solar array	soup	soup bowl
soupspoon	sour cream	soya milk	space shuttle	sparkler	spear
crawfish	sponge	sportswear	spotlight	stagecoach	statue
steak	steak knife	steering wheel	stepladder	step stool	stereo
stew	stirrer	stirrup	brake light	stove	strainer
strap	street sign	streetlight	string cheese	stylus	subwoofer
sugar bowl	sugarcane	sunflower	sunhat	mop	sweat pants
sweatband	sweater	sweatshirt	sweet potato	Tabasco sauce	table-tennis table
table lamp	tablecloth	tachometer	tag	taillight	tambourine
army tank	tank top	tape	tape measure	tapestry	tarp
tartan	tassel	tea bag	teacup	teakettle	telephone booth
telephone pole	telephoto lens	television camera	television set	tequila	thermometer
thermos bottle	thermostat	thimble	thread	thumbtack	tights
timer	tinfoil	tinsel	tissue paper	toast	toaster oven
tongs	toolbox	toothpaste	toothpick	cover	tortilla
tow truck	towel rack	tractor	dirt bike	trailer truck	trampoline
tray	trench coat	triangle	tricycle	truffle	trunk
vat	turban	turnip	turtleneck	typewriter	underwear
urinal	urn	vacuum cleaner	vending machine	vent	vest
videotape	vinegar	vodka	volleyball	vulture	wagon
wagon wheel	walking stick	wall socket	wallet	walrus	washbasin
water bottle	water cooler	water heater	water jug	water gun	water ski
water tower	watering can	weathervane	webcam	wedding cake	wedding ring
wet suit	whipped cream	whistle	wig	wind chime	windmill
window box	windshield wiper	windsock	wine bottle	wine bucket	wineglass
blinder	wolf	wooden spoon	wreath	wristband	wristlet
yacht	yogurt	yoke			