Appendix: Cross-domain Multi-modal Few-shot Object Detection via Rich Text

Zeyu Shangguan, Daniel Seita, and Mohammad Rostami
{zshanggu,seita,rostamim}@usc.edu
Department of Computer Science
University of Southern California

1. Experiments on Additional Benchmarks

In this section, we present our results on other CD-FSOD benchmarks.

Fu *et al*. [3] introduce one such benchmark built on three datasets: Clipart1k [6], DeepFish [13], and NEU-DET [15]. The few-shot sampling strategy is balanced (*i.e*. k shot refers to k instances). The results are presented in Tab 1, demonstrating that our approach outperforms previous methods, particularly on the Clipart1k dataset. Specifically, in the 10-shot scenario, we achieve a score of 47.7 and 48.8, significantly surpassing the best prior method's score of 25.6. However, we observe that in the NEU-DET dataset, the text generated with the assistance of LLM does not yield superior results compared to manually crafted text. This is because the LLM failed to produce texts with sufficient distinguishing features as seen in other datasets (see Section 5.4). Specifically, the LLM-aided text descriptions for the six defect categories tended to exhibit homogeneity, limiting their effectiveness.

Lee *et al*. [8] introduce another CD-FSOD benchmark which includes 10 datasets: VisDrone [23] for aerial images, DeepFruits [12] for agriculture, iWildCam [1] for animals in the wild, Clipart [6] for cartoons, iMaterialist [5] for fashion, Oktoberfest [24] for food, LogoDet-3K [16] for logos, CrowdHuman [14] for people, SIXray [11] for security, and KITTI [4] for traffic/autonomous driving. However, these benchmarks use an unbalanced sampling strategy, *i.e*., k-shot means k images per class. We conduct experiments on two most visual unique datasets: Clipart and SIXray, as illustrated in Tab. 2. Our method surpass the performance of Lee's result.

2. General FSOD Benchmark

The general FSOD framework enables the model to undergo fine-tuning on both novel and base data, helping to mitigate catastrophic forgetting. The benchmark evaluation for general FSOD, as proposed by TFA [17], on the PASCAL VOC dataset comprises 20 categories, with 15 designated as base categories and the remaining 5 as novel categories. To eliminate any potential data bias, 3 fixed category splits have been established for training and testing, enabling an accurate assessment of the model's average performance. The benchmark employs two distinct sampling strategies for sampling the base categories during fine-tuning: balanced sampling and unbalanced sampling. In the case of balanced sampling, the k-shot refers to utilizing k instances for each category, including both novel and base categories. This sampling approach is employed in methods like TFA. On the other hand, the unbalanced sampling strategy involves the k-shot specifically referring to strictly selecting k instances for the novel categories alone, excluding the base categories. This sampling approach is employed in methods like Meta-DETR [20]. Both strategies are acceptable for general FSOD, and for our results reported in Tab.2, we follow the unbalanced sampling strategy.

3. Visualization of the Detection Result

We present additional visualizations of our detection results as an extension of Sec.4.6. The detection results on the ArTaxOr, DIOR and UODD dataset are illustrated in Fig. 1, Fig. 2 and Fig. 3 respectively. We demonstrate consistent improvements compared to the multi-modal method (Next-chat) and the single-modal method (Meta-DETR). Specifically, in Fig. 1, we observe more precise classification of Lepidoptera, improved bounding box regression for Odonate, and enhanced classification confidence. Similarly, notable advancements are also evident in Fig. 2 and Fig. 3. Furthermore, in Fig. 3, we observe additional foreground detection results, such as the sea cucumber in the second row and the scallop in the third row.

4. Computational Overhead

We list our computational overhead as a reference in Tab. 3. The values are reported based on 10-shot (68 images each epoch) experiments on the ArTaxOr dataset. See Sec.4.1 for additional hyperparameters we use. Our method has more than 10X training parameters compared to Meta-DETR, but only has about 2X training time because we cal-

Table 1. Performance Results (mAP) on CD-FSOD benchmarks proposed by Fu *et al*. [3]. The † denotes that the methods are developed or the results are reported by Fu *et al*.; ‡ indicates the results of MM-FSOD are reported by us. Highest scores are in bold font.

Method \ Shot	Backbone	Modality	Clipart1k			DeepFish			NEU-DET		
			1	5	10	1	5	10	1	5	10
ViTDeT-FT† [10]	ViT-B/14	Single	6.1	23.3	25.6	0.9	9.0	13.5	2.4	6.5	15.8
Detic† [22]	ViT-L/14	Multi	11.4	11.4	11.4	0.9	0.9	0.9	0.0	0.0	0.0
Detic-FT† [22]	ViT-L/14	Multi	15.1	20.2	22.3	9.0	14.3	17.9	3.8	14.1	16.8
DE-ViT† [21]	ViT-L/14	Single	0.5	5.5	11.0	0.4	2.5	2.1	0.4	1.5	1.8
Meta-DETR† [20]	ViT-L/14	Single	9.6	17.3	20.7	4.1	7.5	13.6	2.0	13.4	15.2
Next-Chat‡ [19]	ViT-L/14	Multi	5.5	6.1	12.1	0.4	1.9	3.0	1.4	1.6	2.0
Our Method w/ self-built text	DETR-R101	Multi	**22.9**	**42.6**	**47.7**	**21.2**	**23.3**	**25.2**	**6.1**	**15.2**	**19.9**
Our Method w/ LLM text	DETR-R101	Multi	**25.6**	**44.3**	**48.8**	**23.5**	**25.6**	**26.4**	5.8	13.1	13.2

Table 2. Performance Results (mAP) on CD-FSOD benchmarks proposed by Lee *et al*. [8]. Highest scores are in bold font.

Method \ Shot	Backbone	Modality	Clipart		SIXray	
			1	5	1	5
Lee *et al*.	ResNet-50	Single	37.2	49.3	6.6	23.9
Our Method w/ self-built text	DETR-VIT-L	Multi	**46.5**	**58.4**	**13.3**	**34.7**
Our Method w/ LLM text	DETR-VIT-L	Multi	**47.2**	**58.8**	**14.5**	**35.3**

culate and cache the feature embedding of the rich text.

5. Detailed List of the Rich Text

5.1. ArTaxOr Dataset

As depicted in Fig. 4, the ArTaxOr Dataset [2] comprises seven distinct categories of insects. The images within this dataset were primarily captured using a macro lens, resulting in high-definition visuals and a distinct foreground-background boundary. Nevertheless, the inter-class differences within the ArTaxOr Dataset are relatively subtle, posing a challenge in accurately distinguishing between the various insect categories [18].

- *Araneae*:
  - (manually built) Araneae have eight limbs and usually perch on the silk, and do not have antennae.
  - (LLM-aided) Araneae, commonly known as spiders, exhibit a diverse range of appearances, varying from small and delicate to large and imposing, with eight legs, two body segments, and often distinctive markings and patterns on their abdomens.
- *Coleoptera*:
  - (manually built) Beetles can usually be recognized by their two pairs of wings; the front pair is modified into horny covers that hide the rear pair and most of the abdomen and usually meet down the back in a straight line.
  - (LLM-aided) Coleoptera, the beetles, exhibit a remarkable diversity in appearance, ranging from small, shiny, and flightless to large, winged, and brightly colored, often with hardened forewings adapted for protection or flight.
- *Diptera*:
  - (manually built) Diptera use only a single pair of wings to fly, and have a mobile head, with a pair of large compound eyes, and mouthparts designed for piercing and sucking.
  - (LLM-aided) Diptera, or flies, are small insects with two wings, large compound eyes, and long antennae, varying in color from dull to bright, with some species having patterns or stripes on their bodies.
- *Hemiptera*:

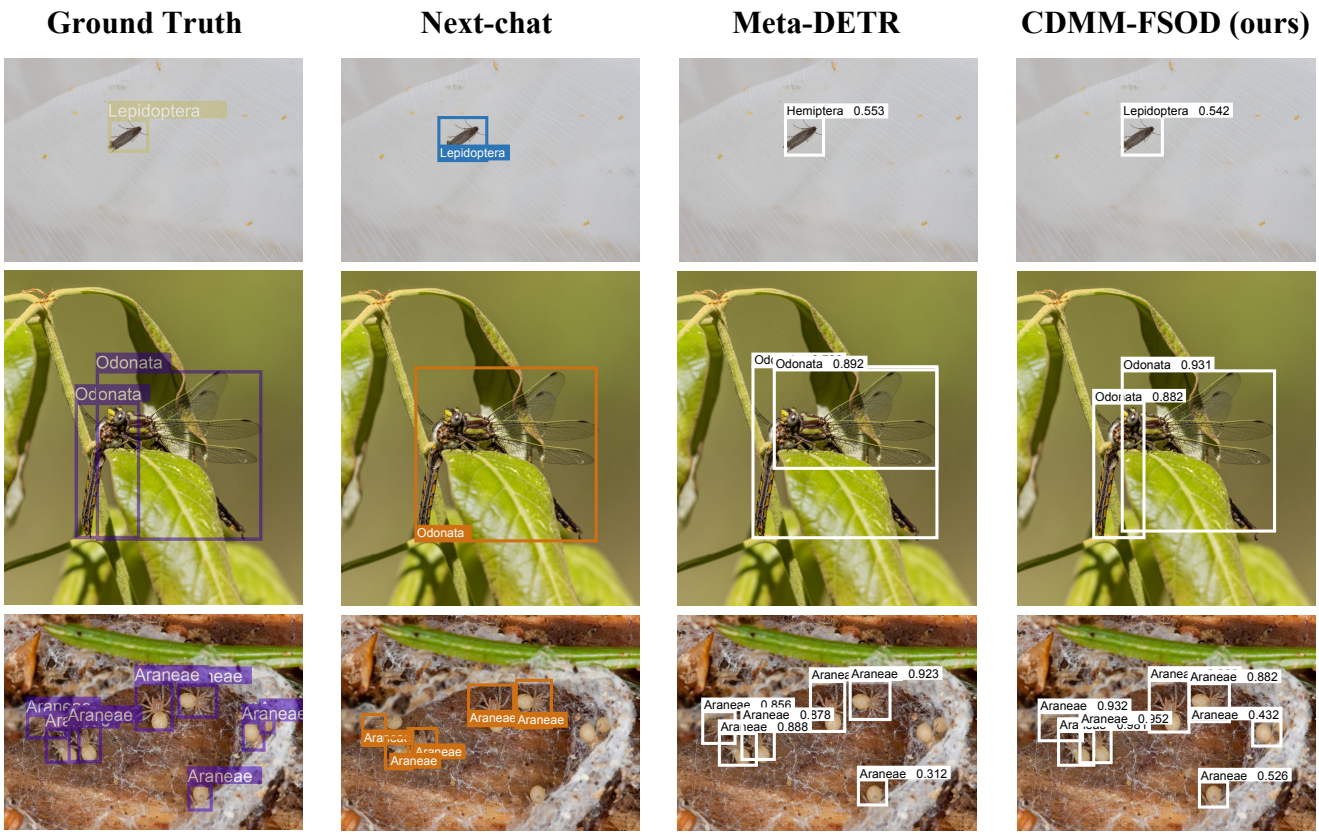

Figure 1. Examples of detection results from ArTaxOr.

Table 3. Computation overhead of our methods

	Trainable parameters	Model size	Training time
Meta-DETR	11,649,794	317MB	36 s/epoch
Our method	162,927,107	671MB	84 s/epoch

  - (manually built) Hemiptera's forewings are hardened near the base and membranous near the ends.
  - (LLM-aided) Hemiptera, or bugs, are small to medium-sized insects typically characterized by a flattened body, strong forewings hardened into protective plates, and prominent eyes. Some species are brightly colored while others are dull-brown or black.

- *Hymenoptera*:
  - (manually built) Hymenoptera usually have two pairs of membranous wings and the forewings are larger than the hind wings.
  - (LLM-aided) Hymenoptera are distinguished by their thin waist connecting the thorax and abdomen, as well as their typically transparent and membranous wings.

- *Lepidoptera*:
  - (manually built) Lepidoptera have scales that cover the bodies, large triangular wings, and a proboscis for siphoning nectars.
  - (LLM-aided) Lepidoptera, including butterflies and moths, are characterized by their wings, which are covered in scales and often display patterns and colors ranging from subtle to vivid. Their bodies are typically slender, and they have antennae and long, thin legs.

- *Odonata*:
  - (manually built) Odonata characteristically have large rounded heads covered mostly by com-

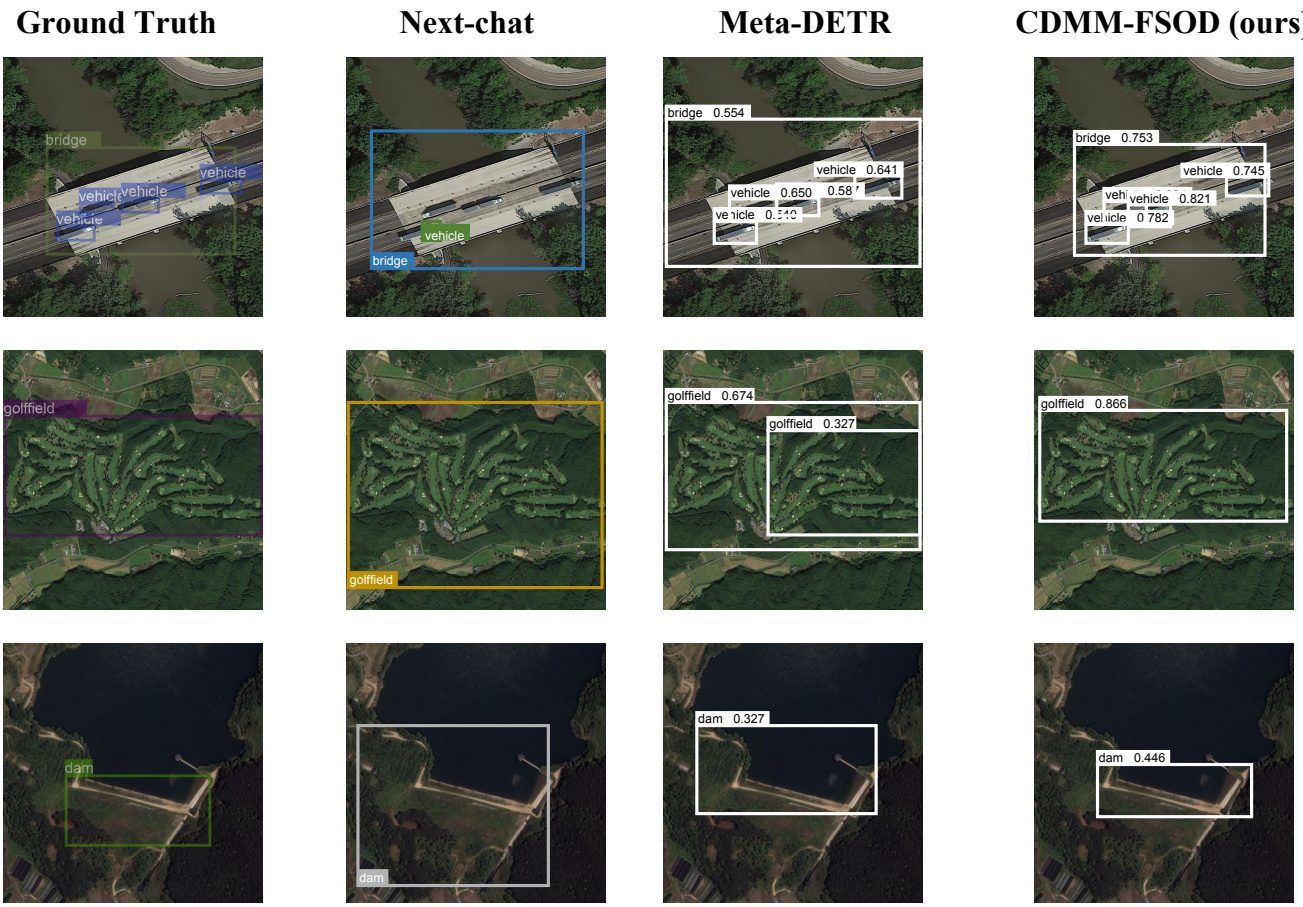

Figure 2. Examples of detection results from DIOR.

pound eyes, two pairs of long, transparent wings that move independently, elongated abdomens, three ocelli and short antennae.

– (LLM-aided) Odonata, commonly known as dragonflies and damselflies, are characterized by their large, often brightly colored eyes, long, thin bodies, and two pairs of wings, the hindwings being broader than the forewings.

5.2. UODD Dataset

The UODD dataset [7], as illustrated in Fig. 5, comprises three distinct categories of underwater creatures. Due to the low visual contrast, it exhibits significant differences from the COCO dataset [18].

- *Sea cucumber*:
  - (manually built) Sea cucumbers have sausage-shape, usually resemble caterpillars; their mouth is surrounded by tentacles.
  - (extended) Sea cucumbers have sausage-shape, usually resemble caterpillars; their mouth is surrounded by tentacles; usually seen together with sea urchins.
  - (LLM-aided) Sea cucumbers are marine invertebrates known for their elongated, leathery bodies, which are typically covered in spines or tentacles and lack a distinct head or tail.
- *Sea urchin*:
  - (manually built) Sea urchins have globular body and a radial arrangement of organs.
  - (extended) Sea urchins have globular body and a radial arrangement of organs; usually seen together with sea cucumbers.
  - (LLM-aided) Sea urchins are marine invertebrates characterized by their globular shape, covered in dense clusters of spines for protection, and having a distinctive oral surface with five radiating rows of teeth-like structures called pedicellariae.
- *Scallop*:

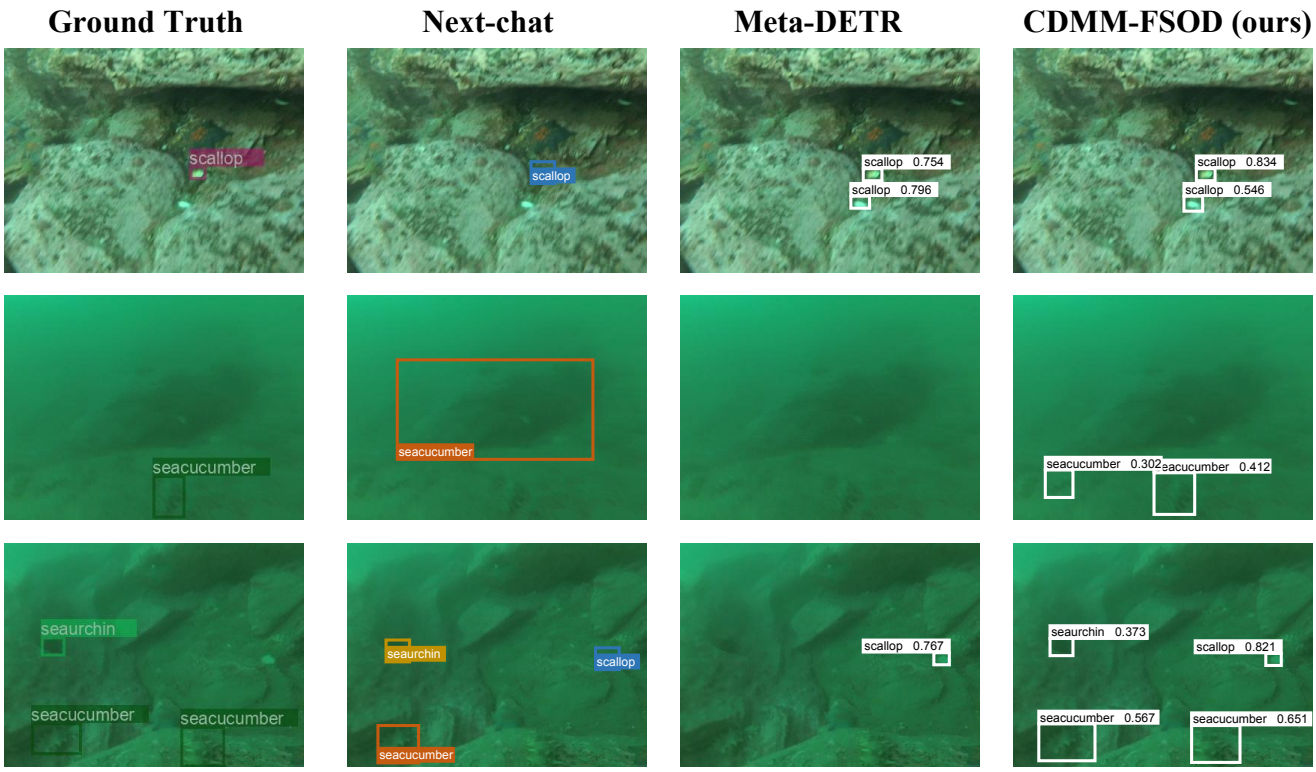

Figure 3. Examples of detection results from UODD.

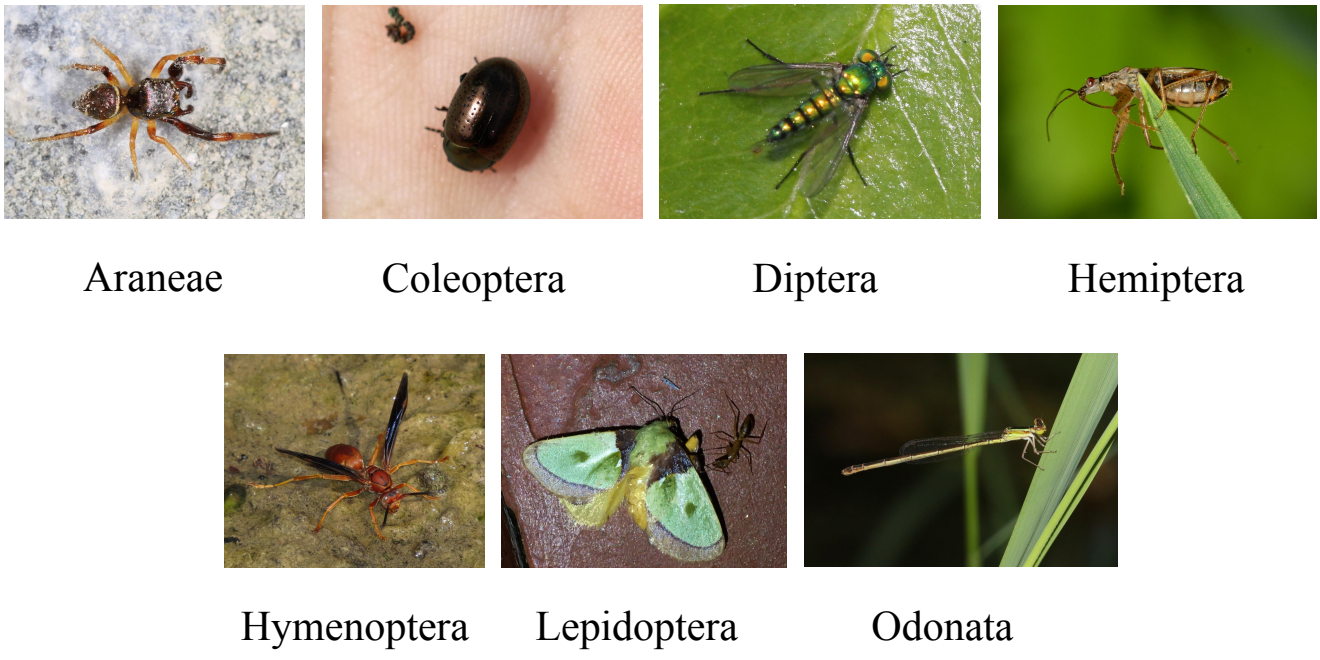

Figure 4. Examples of images and categories from ArTaxOr.

– (manually built) The shell of a scallop has the classic fanned-out shape; one side of the shell is slightly flatter, and the other more concave in shape.

Figure 5. Examples of images and categories from UODD.

  - (extended) Sea cucumbers have sausage-shape, usually resemble caterpillars; their mouth is surrounded by tentacles; usually seen together with sea urchins.
  - (LLM-aided) Scallops are marine bivalve mollusks characterized by their flattened, oval-shaped shells, typically with distinctive radial ribbing and a wavy or scalloped edge.

5.3. DIOR Dataset

As depicted in Fig. 6, the DIOR dataset [9] encompasses 20 categories of Aerial scenes. These scenes differ significantly from those in the COCO dataset due to their unique remote sensing perspective, captured from a top-down shot [18].

- *Expressway service area*:
  - (manually built) A expressway service area is an open space located alongside the highway, typically featuring a main building and multiple regularly arranged parking slots.
  - (LLM-aided) Expressway service areas appear as clusters of buildings and facilities along the highway, typically surrounded by parking lots and green spaces.
- *Expressway toll station*:
  - (manually built) Toll stations on expressways are elongate, thin buildings that span across the highway, and allow vehicles to queue up for entry or exit.
  - (LLM-aided) Expressway toll stations appear as structures located along the highway, typically with multiple lanes entry and exit points, and are distinguished by their regular layout, clear lanes markings, and sometimes by the presence of toll plazas or gantries.
- *Airplane*:
  - (manually built) An airplane is equipped with a pair of symmetric, large wings positioned at the center of its cylindrical fuselage, along with a pair of smaller wings referred to as the horizontal stabilizer and a vertical stabilizer located at the tail end.
  - (LLM-aided) Airplanes are winged vehicles with a fuselage, wings, and engines, typically having a pointed nose and tail.
- *Airports*:
  - (manually built) Airports are designated open areas where airplanes can park, and they typically feature multiple comb-shaped boarding bridges, tall control towers, and lawns, but lack trees.
  - (LLM-aided) Airports are large open spaces with multiple runways, terminals, and parking lots, typically surrounded by fences and gates, and equipped with aircrafts, control towers.
- *Baseball field*:
  - (manually built) The baseball field features a diamond-shaped infield, resembling a quarter of a circle, covered with a checkerboard pattern of natural green grass.
  - (LLM-aided) A baseball field is characterized by a diamond-shaped infield with four bases, a pitcher's mound, and outfield fences, surrounded by stands for spectators and often marked by lines indicating playing areas and distances.
- *Basketball court*:
  - (manually built) The basketball court is the playing surface, consisting of a rectangular floor with tiles at either end, usually made out of a wood, often maple, and highly polished.

Figure 6. Examples of images and categories from DIOR.

  - (LLM-aided) A basketball court is a rectangular playing surface with a raised rim and net at each end, typically surrounded by lines dividing the court into various zones and marked by a center circle and free-throw lines.

- *Bridge*:
  - (manually built) Bridges are elongate, slender structures spanning across water bodies and may accommodate the passage of vehicles.
  - (LLM-aided) Bridges are structures built to span rivers, valleys, or other obstacles, typically with one or more arches, trusses, or cables, connecting two or more points of land.

- *Chimney*:
  - (manually built) A chimney is like a pipe or a tunnel-like channel made by metal or concrete, typically tall and expel gas.
  - (LLM-aided) Chimneys are vertical structures designed to vent smoke and gases from heating appliances, typically made of masonry or metal and having a rectangular or circular cross-section.

- *Dam*:
  - (manually built) A dam is a structure constructed across a stream, river, or estuary to retain water, and from an overhead perspective, it typically appears as a thin, curved white line.
  - (LLM-aided) Dams are large structures built across rivers or streams to control water flow, typically with a wall or gates to hold back water and create a reservoir, and may include spillways to control flooding.

- *Golf course*:
  - (manually built) Golf courses are extensive, irregularly shaped open spaces characterized by meticulously manicured lawns, and the terrain is typically gently undulating.
  - (LLM-aided) Golf courses appear as large, open spaces with regularly spaced tee boxes, greens, and fairways, typically surrounded by fences or hedges and distinguished by their smooth, manicured surfaces.

- *Ground track field*:

  - (manually built) Ground track fields typically consist of red rings forming elliptic paths, with a pair of straight lines facing each other and another pair of half-circles extending in the opposite direction.
  - (LLM-aided) Ground track fields appear as large, open spaces with straight lines and rectangles marking the field boundaries and flight paths, typically surrounded by fences or other markers and distinguished by their regularity and symmetry.

- *Harbor*:
  - (manually built) A harbor refers to a sheltered body of water, either natural or man-made, and it is commonly seen filled with boats or ships parked within its confines.
  - (LLM-aided) Harbors appear as clusters of ships and boats moored along the coast or riverbanks, typically surrounded by docks, piers, and other facilities for loading and unloading cargo, and distinguished by their dense concentration of vessels and associated infrastructure.

- *Overpass*:
  - (manually built) Overpasses are intricate city roadways that resemble bridges and intersect each other, typically elevated above standard road levels to facilitate the passage of vehicles.
  - (LLM-aided) Overpasses appear as elevated structures spanning roads or railroads, typically with multiple arches or girders supporting the roadway above the traffic below, and distinguished by their raised position and distinctive shape.

- *Ship*:
  - (manually built) Ships typically feature a long, thin, bullet-shaped hull that floats on the water, with a sharp prow and a flat stern.
  - (LLM-aided) Ships appear as large, elongated vessels floating on the water surface, typically with a distinctive silhouette and various features such as decks, masts.

- *Stadium*:
  - (manually built) Stadiums resemble bowl-shaped buildings and typically feature stands surrounding a central field.
  - (LLM-aided) Stadiums appear as large, open spaces with distinctive curved or rectangular shapes, typically surrounded by stands and other facilities for spectators.

- *Storage tank*:
  - (manually built) Storage tanks are usually large cylindrical or spherical metal containers with a clean exterior, clustered and well-arranged on the ground.
  - (LLM-aided) Storage tanks appear as large, circular or rectangular structures with flat roofs, typically surrounded by fenced areas and distinguished by their regular shape, smooth surfaces, and sometimes by the presence of multiple tanks clustered together.

- *Tennis court*:
  - (manually built) Tennis courts are rectangular fields marked with straight lines across their surfaces, and they are typically colored green or blue.
  - (LLM-aided) Tennis courts appear as rectangular areas with distinctive lines marking the boundaries and service lines, typically surrounded by fences or hedges and distinguished by their regularity and symmetry.

- *Train station*:
  - (manually built) Train stations are large buildings designed to accommodate trains, and they typically feature a dense network of multiple railways passing through them.
  - (LLM-aided) Train stations appear as clusters of buildings and platforms along the railway tracks, typically with multiple tracks converging and diverging, and distinguished by the presence of trains or other rail vehicles.

- *Vehicle*:
  - (manually built) Vehicles are typically small rectangular objects commonly seen traveling along roads or parked in parking lots.
  - (LLM-aided) Vehicles appear as small to medium-sized objects with distinctive shapes and features, such as wheels, windows, and roofs, typically moving along roads or parking areas.

- *Windmill*:
  - (manually built) Windmills are tall structures topped with multiple fan-like blades that spread out and rotate to harness wind energy.

– (LLM-aided) Windmills appear as tall, thin structures with three or more long blades rotating in the wind, typically located in open spaces such as fields or coastal areas.

5.4. NEU-DET Dataset

As shown in Fig. 7, the NEU-DET dataset [15] features six categories of common steel rolling product defects. These images were captured using an electron microscope, resulting in a distinct black-and-white style. Notably, the industrial image style of the NEU-DET dataset differs significantly from the life scenes presented in the COCO datasets.

- *Rolled-in scale*:
  - (manually built) Rolled-in scale typically consists of multiple rough patches arranged in a fish scale-like pattern.
  - (LLM-aided) In hot-rolled steel strip defect images, rolled-in scale appears as dark, irregular patches or streaks along the surface of the steel strip.
- *Patches*:
  - (manually built) Patch defects typically appear as dark, dense clusters.
  - (LLM-aided) In hot-rolled steel strip defect images, a patch typically appears as a localized area of abnormal texture, color, or brightness
- *Crazing*:
  - (manually built) Crazings typically consist of numerous tiny, thin lines resembling tree branches.
  - (LLM-aided) Crazing in hot-rolled steel strip defect images is characterized by a network of fine cracks or fissures that appear on the surface of the steel strip, typically have a regular, hairline pattern.
- *Pitted surface*:
  - (manually built) Pitted surfaces are characterized by a dense cluster of tiny spots.
  - (LLM-aided) Pitted surfaces are characterized by numerous small, circular indentations or depressions spread across the steel strip's surface.
- *Inclusion*:
  - (manually built) Inclusions typically consist of a series of dark, slender patches.
  - (LLM-aided) Inclusions in hot-rolled steel strip defect images appear as dark, irregularly shaped areas within the steel strip's matrix, typically caused by foreign particles or impurities during the manufacturing.
- *Scratches*:
  - (manually built) Scratches are typically characterized by multiple light, straight, slender lines.
  - (LLM-aided) Scratches appear as long, continuous lines or marks across the steel strip's surface.

5.5. Clipart1k Dataset

The Clipart1k Dataset [6], in Fig. 8, comprises 20 categories of commonly seen objects that overlap with the COCO categories, however, due to their cartoon-like style, models often struggle to accurately recognize them.

- *Aeroplane*:
  - (manually built) The cartoon aeroplane typically features a pair of wings positioned in the center of its cylindrical body, accompanied by a set of small stabilizers at the rear, usually in unified colors and clear contours.
  - (LLM-aided) A cartoon aeroplane is characterized by its sleek fuselage, wings and various mechanical components, all exaggerated in features, enhanced with brighter colors, and adorned in a whimsical or anthropomorphic design.
- *Bicycle*:
  - (manually built) The cartoon bicycle typically comprises of two wheels, one at the front and another at the rear, along with a triangular-shaped frame, usually in unified colors and clear contours.
  - (LLM-aided) A cartoon bicycle is typically defined by its two wheels, a frame connecting them, handlebars for steering, and a seat for the rider, all exaggerated in features, enhanced with brighter colors, and adorned in a whimsical or anthropomorphic design.
- *Bird*:
  - (manually built) The cartoon bird commonly has a furry body, a pair of feathery wings, and a pair of slender feet, usually in unified colors and clear contours.

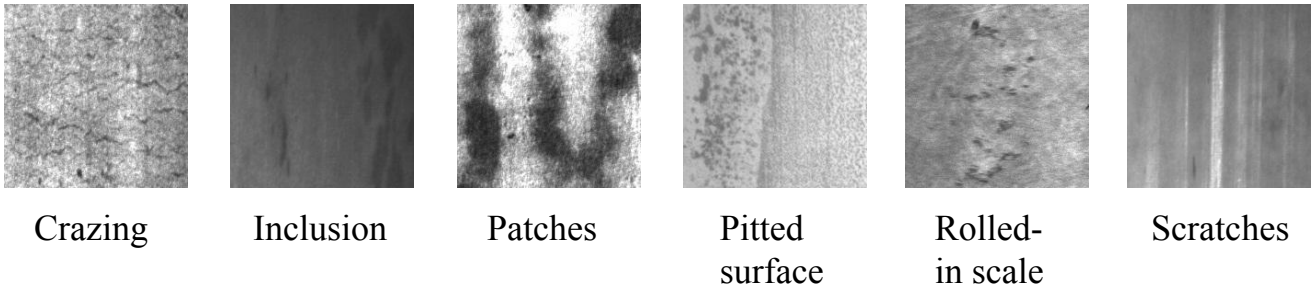

Figure 7. Examples of images and categories from NEU-DET.

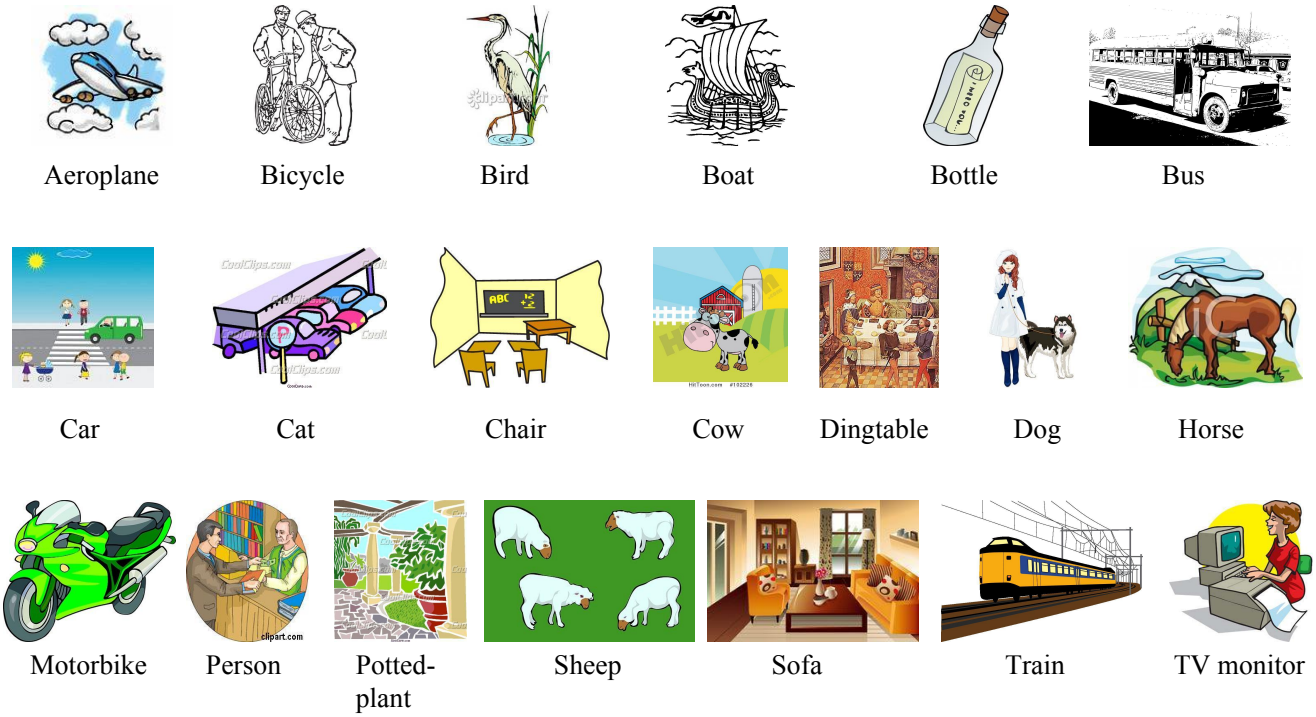

Figure 8. Examples of images and categories from Clipart1k.

  - (LLM-aided) A cartoon bird is typically characterized by its feathered body, beaked head, and wings, all exaggerated in features, enhanced with brighter colors, and adorned in a whimsical or anthropomorphic design.

- *Boat*:
  - (manually built) The cartoon boat typically features a half-moon-shaped hull floating on the water, and sometimes with a set of big sails, usually in unified colors and clear contours.
  - (LLM-aided) A cartoon boat is typically defined by its hull, which floats on water, along with features such as a deck, mast, and sails or engines, all exaggerated in features, enhanced with brighter colors, and adorned in a whimsical or anthropomorphic design.

- *Bottle*:
  - (manually built) The cartoon bottle commonly exhibits a transparent cylindrical body, characterized by a slender neck and a wider abdomen, usually in unified colors and clear contours.
  - (LLM-aided) A cartoon bottle is typically characterized by its distinct shape, often cylindrical or tapered, with a narrow neck and a mouth for pouring or sealing, all exaggerated in features, enhanced with brighter colors, and adorned in a whimsical or anthropomorphic design.

- *Bus*:
  - (manually built) The cartoon bus typically takes the form of a cuboid-shaped vehicle, supported by four wheels at the base. Its body is surrounded by rectangular windows, usually in unified colors and clear contours.
  - (LLM-aided) A cartoon bus is typically characterized by its large size, rectangular shape, and multiple doors for passengers, along with windows for visibility and a roof that often features an advertising panel or other decorative elements, all exaggerated in features, enhanced with brighter colors, and adorned in a whimsical or anthropomorphic design.
- *Car*:
  - (manually built) Cartoon cars are vehicles typically featuring a small, box-shaped driving cab and an flat elongated front engine box, supported by four wheels at the base, usually in unified colors and clear contours.
  - (LLM-aided) A cartoon car is typically defined by its sleek and compact body, with four wheels, a windshield, and various features such as headlights, taillights, and mirrors, all exaggerated in features, enhanced with brighter colors, and adorned in a whimsical or anthropomorphic design.
- *Cat*:
  - (manually built) Cartoon cats possess a furry body with four limbs, a pair of triangular ears on top of their heads, long tails, and slightly pointed mouths, usually in unified colors and clear contours.
  - (LLM-aided) A cartoon cat is typically characterized by its sleek and muscular body, furry coat, pointed ears, and large, expressive eyes, all exaggerated in features, enhanced with brighter colors, and adorned in a whimsical or anthropomorphic design.
- *Chair*:
  - (manually built) The cartoon chair typically features four slender legs for support, along with a small, horizontal, flat area designed for sitting. sometimes includes a vertical backrest and a pair of armrests, usually in unified colors and clear contours.
  - (LLM-aided) A cartoon chair is typically defined by its supportive structure, consisting of a seat, backrest, and legs, all exaggerated in features, enhanced with brighter colors, and adorned in a whimsical or anthropomorphic design.
- *Cow*:
  - (manually built) The cartoon cow commonly possesses four long legs that support its box-shaped body, with a long tail, sometimes, it may also feature a pair of crescent-shaped horns protruding from its head, usually in unified colors and clear contours.
  - (LLM-aided) A cartoon cow is typically characterized by its large, stocky body, brown or black fur, and prominent horns, all exaggerated in features, enhanced with brighter colors, and adorned in a whimsical or anthropomorphic design.
- *Diningtable*:
  - (manually built) The cartoon dining table is typically a large, tall piece of furniture supported by multiple legs; it's surface is often rounded or rectangular, usually in unified colors and clear contours.
  - (LLM-aided) A cartoon dining table is typically defined by its rectangular or circular shape, flat surface for placing food and dishes, and supportive legs, all exaggerated in features, enhanced with brighter colors, and adorned in a whimsical or anthropomorphic design.
- *Dog*:
  - (manually built) The cartoon dog typically exhibits a furry body, characterized by a prominent, elongated nose and mouth, leaf-shaped ears, four legs, and a tail that is often curled, usually in unified colors and clear contours.
  - (LLM-aided) A cartoon dog is typically characterized by its varied coats, distinct ears and muzzle shapes, as well as its four legs and tail, all exaggerated in features, enhanced with brighter colors, and adorned in a whimsical or anthropomorphic design.
- *Horse*:
  - (manually built) The cartoon horse typically stands on four legs, supporting its body, and features a brush-like tail, a long neck and a long face, as well as a long mane along its neck, usually in unified colors and clear contours.

  - (LLM-aided) A cartoon horse is typically characterized by its powerful legs, long mane and tail, and muscular body, all exaggerated in features, enhanced with brighter colors, and adorned in a whimsical or anthropomorphic design.

- *Motorbike*:
  - (manually built) The cartoon motorbike typically features two bold wheels at the front and rear of its body, and a prominent engine positioned in the middle, usually in unified colors and clear contours.
  - (LLM-aided) A cartoon motorbike is typically defined by its compact and aerodynamic frame, two wheels, a handlebar for steering, and an engine for power, all exaggerated in features, enhanced with brighter colors, and adorned in a whimsical or anthropomorphic design.

- *Person*:
  - (manually built) The cartoon person typically possesses two legs to support their body, a pair of arms, and a head with a pair of eyes, ears, and a mouth, usually in unified colors and clear contours.
  - (LLM-aided) A cartoon person is typically characterized by a bipedal stance, upright posture, and a head with facial features such as eyes, nose, and mouth, along with arms, legs, and varying skin tones and hairstyles, all exaggerated in features, enhanced with brighter colors, and adorned in a whimsical or anthropomorphic design.

- *Potted plant*:
  - (manually built) The cartoon potted plants are typically small trees or bushes grown indoors in pots, with green leaves or vibrant flowers, usually in unified colors and clear contours.
  - (LLM-aided) A cartoon potted plant is typically defined by its green foliage or flowers, growing from a soil-filled pot, all exaggerated in features, enhanced with brighter colors, and adorned in a whimsical or anthropomorphic design.

- *Sheep*:
  - (manually built) The cartoon sheep commonly stands on four legs, supporting its body, and is covered in white, curly fur; it features a short, curly tail and a pair of horns that curl downward, usually in unified colors and clear contours.
  - (LLM-aided) A cartoon sheep is typically characterized by its fleecy white coat, rounded body, and curved horns, all exaggerated in features, enhanced with brighter colors, and adorned in a whimsical or anthropomorphic design.

- *Sofa*:
  - (manually built) The cartoon sofa boasts a long, box-shaped body designed for seating, adorned with cushions and featuring a thick backrest as well as a pair of sturdy armrests, usually in unified colors and clear contours.
  - (LLM-aided) A cartoon sofa is typically defined by its cushioned seating, backrest, and arms, all exaggerated in features, enhanced with brighter colors, and adorned in a whimsical or anthropomorphic design.

- *Train*:
  - (manually built) The cartoon train comprises a succession of long, box-shaped compartments that traverse along a rail, usually in unified colors and clear contours.
  - (LLM-aided) A cartoon train is typically characterized by its long and connected series of carriages, each with windows and doors, a powerful locomotive at the front, all exaggerated in features, enhanced with brighter colors, and adorned in a whimsical or anthropomorphic design.

- *TV monitor*:
  - (manually built) The cartoon TV monitor typically comprises a screen for displaying images and a sturdy base for support, usually in unified colors and clear contours.
  - (LLM-aided) A cartoon TV monitor is typically defined by its rectangular screen, displaying images and sounds, with a frame or border, and various buttons or controls for operation, all exaggerated in features, enhanced with brighter colors, and adorned in a whimsical or anthropomorphic design.

5.6. DeepFish Dataset

The DeepFish dataset [13], as illustrated in Fig. 9, exclusively features a single category of high-definition underwater tropical fishes, all sampled in Australia. Its unique image style, characterized by blurred visual boundaries and the fishes' biological camouflage, sets it apart from the COCO dataset.

- *Fish*:

Figure 9. Examples of images and categories from DeepFish.

  - (manually built) Underwater fish typically exhibit a spindle-shaped body covered with scales, along with a pair of pectoral fins and a tail that is crescent-shaped.
  - (LLM-aided) Underwater fish typically exhibit diverse appearance features, ranging from sleek and streamlined bodies to brightly colored scales and fins.

5.7. SIXray Dataset

The SIXray dataset [11], as illustrated in Fig. 9, exclusively features six category of contraband for security check. Its unique X-ray image style, characterized by complex boundaries, sets it apart from the COCO dataset.

- *Knife*:
  - (manually built) Knives appear as long, dark shadows, consisting of a handle and a sharp blade.
  - (LLM-aided) The knife typically exhibits a blade with a sharp edge for cutting
- *Gun*:
  - (manually built) Guns are discernible as dark L-shaped shadows, featuring a long, slender barrel on one side and a shorter, stubby handle on the other.
  - (LLM-aided) The gun typically exhibits a compact and handgun-like shape, featuring a barrel, a grip for holding, a trigger for firing.
- *Wrench*:
  - (manually built) The wrench is discernible as a slim, elongated shape, with one end featuring a horseshoe-shaped head and the other end attached with a circular ring.
  - (LLM-aided) The wrench typically features a handle for gripping and a jaw or socket that can be adjusted to fit nuts, bolts, or other fasteners, enabling it to rotate or tighten them.
- *Pliers*:
  - (manually built) Pliers are characterized by their two slender handles, connected to a pointed, snip-nosed jaw.
  - (LLM-aided) The pliers typically consist of two pivoting levers with jaws at the ends, designed to grasp and manipulate objects, with handles for easy gripping and control.
- *Scissors*:
  - (manually built) Scissors comprise two circular handles connected to a long, sharp blade that forms the cutting jaw.
  - (LLM-aided) The scissors typically consist of two pivoting blades with sharp edges that meet at a point, allowing for cutting action when the handles are squeezed together.

References

[1] Sara Beery, Arushi Agarwal, Elijah Cole, and Vighnesh Birodkar. The iwildcam 2021 competition dataset. *CoRR*, abs/2105.03494, 2021. 1

[2] Geir Drange. Arthropod taxonomy orders object detection dataset, 2019. 2

[3] Yuqian Fu, Yu Wang, Yixuan Pan, Lian Huai, Xingyu Qiu, Zeyu Shangguan, Tong Liu, Yanwei Fu, Luc Van Gool, and Xingqun Jiang. Cross-domain few-shot object detection via enhanced open-set object detector. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pages 247–264, 2025. 1, 2

[4] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012. 1

[5] Sheng Guo, Weilin Huang, Xiao Zhang, Prasanna Srikhanta, Yin Cui, Yuan Li, Hartwig Adam, Matthew R. Scott, and Serge Belongie. The imaterialist fashion attribute dataset. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3113–3116, 2019. 1

[6] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *CVPR*, 2018. 1, 9

[7] Li-Hui Jiang, Yi Wang, Qi Jia, Shengwei Xu, Yu Liu, Xin Fan, Haojie Li, Risheng Liu, Xinwei Xue, and Ruili Wang.

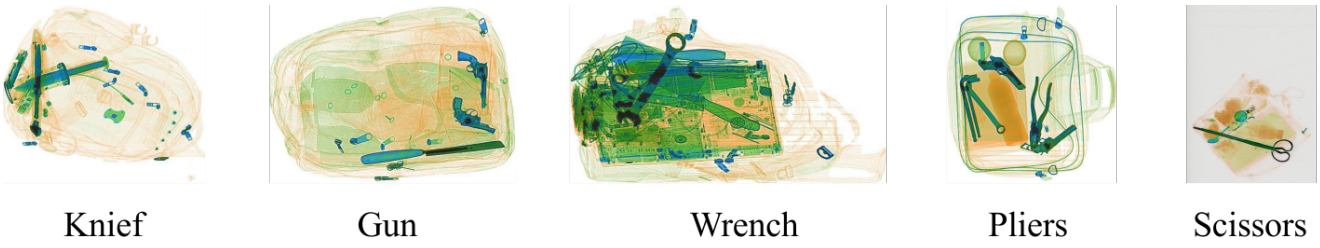

Figure 10. Examples of images and categories from SIXray.

Underwater species detection using channel sharpening attention. *Proceedings of the 29th ACM International Conference on Multimedia*, 2021. 4

[8] Kibok Lee, Hao Yang, Satyaki Chakraborty, Zhaowei Cai, Gurumurthy Swaminathan, Avinash Ravichandran, and Onkar Dabeer. Rethinking few-shot object detection on a multi-domain benchmark. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 366–382, Cham, 2022. Springer Nature Switzerland. 1, 2

[9] Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 159:296–307, 2020. 6

[10] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *ECCV*, 2022. 2

[11] Caijing Miao, Lingxi Xie, Fang Wan, chi Su, Hongye Liu, jianbin Jiao, and Qixiang Ye. Sixray: A large-scale security inspection x-ray benchmark for prohibited item discovery in overlapping images. In *CVPR*, 2019. 1, 13

[12] 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), Aug. 2016. 1

[13] Alzayat Saleh, Issam H Laradji, Dmitry A Konovalov, Michael Bradley, David Vazquez, and Marcus Sheaves. A realistic fish-habitat dataset to evaluate algorithms for underwater visual analysis. *Scientific Reports*, 2020. 1, 12

[14] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018. 1

[15] Kechen Song and Yunhui Yan. A noise robust method based on completed local binary patterns for hot-rolled steel strip surface defects. *Applied Surface Science*, 2013. 1, 9

[16] Jing Wang, Weiqing Min, Sujuan Hou, Shengnan Ma, Yuanjie Zheng, and Shuqiang Jiang. Logodet-3k: A large-scale image dataset for logo detection. *CoRR*, abs/2008.05359, 2020. 1

[17] Xin Wang, Thomas Huang, Joseph Gonzalez, Trevor Darrell, and Fisher Yu. Frustratingly simple few-shot object detection. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9919–9928. PMLR, 13–18 Jul 2020. 1

[18] Wuti Xiong and Li Liu. Cd-fsod: A benchmark for cross-domain few-shot object detection. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. 2, 4, 6

[19] Ao Zhang, Yuan Yao, Wei Ji, Zhiyuan Liu, and Tat-Seng Chua. Next-chat: An lmm for chat, detection and segmentation, 2023. 2

[20] Gongjie Zhang, Zhipeng Luo, Kaiwen Cui, Shijian Lu, and Eric P. Xing. Meta-detr: Image-level few-shot detection with inter-class correlation exploitation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):12832–12843, 2023. 1, 2

[21] Xinyu Zhang, Yuting Wang, and Abdeslam Boularias. Detect every thing with few examples. *arXiv preprint arXiv:2309.12969*, 2023. 2

[22] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, 2022. 2

[23] Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Heng Fan, Qinghua Hu, and Haibin Ling. Detection and tracking meet drones challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7380–7399, 2022. 1

[24] Alexander Ziller, Julius Hansjakob, Vitalii Rusinov, Daniel Zügner, Peter Vogel, and Stephan Günnemann. Oktoberfest food dataset. *CoRR*, abs/1912.05007, 2019. 1