Remedying Target-Domain Astigmatism for Cross-Domain Few-Shot Object Detection

Supplementary Material

1. Cross-Domain Few-Shot Object Detection Problem Formulation

Cross-Domain Few-Shot Object Detection (CD-FSOD) addresses the challenging scenario of adapting object detection models to novel target domains with limited labeled examples. Let $\mathcal{D}_S = \{(I_i^S, \mathcal{B}_i^S, \mathcal{C}_i^S)\}_{i=1}^{N_S}$ represent the source domain dataset with abundant labeled data (typically MS-COCO [6]), where I_i^S is an image, $\mathcal{B}_i^S = \{b_j^S\}_{j=1}^{n_i^S}$ denotes the set of bounding boxes, and $\mathcal{C}_i^S = \{c_j^S\}_{j=1}^{n_i^S}$ represents the corresponding class labels drawn from label space $\mathcal{C}_S$. Similarly, $\mathcal{D}_T = \{(I_i^T, \mathcal{B}_i^T, \mathcal{C}_i^T)\}_{i=1}^{N_T}$ represents the target domain dataset with class labels from $\mathcal{C}_T$.

The fundamental challenges in CD-FSOD stem from two factors: (1) **domain shift**, where the data distributions differ significantly between source and target domains ($P_S \neq P_T$); and (2) **limited supervision**, where only K annotated examples per class (typically $K \in \{1, 5, 10\}$) are available in the target domain. During model development, we first train a detection model on the abundant source domain data $\mathcal{D}_S$, then adapt it to the target domain using the support set $\mathcal{S} = \{(I_i^T, \mathcal{B}_i^T, \mathcal{C}_i^T)\}_{i=1}^{N \times K}$ containing K examples for each of the N novel classes. The model is ultimately evaluated on the query set $\mathcal{Q}$ consisting of unseen target domain images.

Figure 1 illustrates this setting, highlighting the substantial visual differences between the source domain (MS-COCO with diverse everyday objects) and target domains (specialized domains like clipart illustrations and satellite imagery). The cross-domain discrepancy manifests in various forms including changes in visual style, object appearance, background context, and imaging conditions. This setting presents unique opportunities to investigate knowledge transfer across visual domains while operating under extreme data constraints, requiring detection models that can effectively leverage source domain knowledge while adapting to the distinctive characteristics of the target domain with minimal supervision.

Figure 1. Illustration of the Cross-Domain Few-Shot Object Detection setting. The model is first trained on a labeled source domain (MS-COCO) with abundant data, then adapted to target domains (CliPart1k, NEU-DET, etc.) using only K labeled examples per novel class. The figure demonstrates the significant visual domain gaps that models must overcome while learning from limited target domain supervision.

2. Experimental Setup

2.1. Implementation Details

Our model builds on GLIP [5] using PyTorch [7]. Implementation uses a single NVIDIA 3090 GPU with AdamW at base learning rate 5e-5 and weight decay 0.05. Training runs for 200 epochs with early stopping and gradient clipping. We use batch sizes of 2 and 4 for training and evaluation, respectively, and set the prototype similarity threshold τ_{fg} to 0.9 and the enhancement factor γ_{fg} to 0.1.

2.2. Target Domain Datasets

ArTaxOr [1]: This arthropod detection dataset encompasses diverse biological classes including insects, spiders, crustaceans, and millipedes. The dataset presents significant challenges due to the morphological diversity across arthropod species, varying scales from microscopic to macroscopic organisms, and complex natural backgrounds that require fine-grained visual discrimination to distinguish between similar taxonomic groups, as shown in Figure 2.

Clipart1k [2]: This dataset features cartoon-style abstractions that present a significant domain shift from natural photographs. The artistic rendering includes simplified geometric shapes, stylized color schemes, and abstract visual representations that challenge models trained on photorealistic images, as shown in Figure 3.

DIOR [4]: This remote sensing object detection dataset contains diverse annotated satellite imagery featuring objects such as airplanes, ships, and golf fields. The aerial perspective introduces unique challenges including large-scale variations, complex geographic backgrounds, and objects viewed from overhead angles that differ substantially from ground-level detection scenarios, as illustrated in Figure 4.

DeepFish [8]: The underwater fish detection dataset introduces unique challenges including water distortion effects, varying lighting conditions, and complex aquatic environments. The marine setting creates substantial visual differences from terrestrial scenes with specialized underwater optics and marine backgrounds, as illustrated in Figure 5.

NEU-DET [9]: This industrial dataset focuses on surface defect detection with subtle visual patterns and varying scales. The metallic surfaces and microscopic defects require fine-grained visual understanding and precise localization capabilities for quality control applications, as demonstrated in Figure 6.

UODD [3]: The underwater object detection dataset contains marine organisms in complex seafloor environments. The challenging conditions include varying water clarity, diverse lighting scenarios, and intricate backgrounds with coral reefs and marine vegetation that significantly differ from standard object detection scenarios, as shown in Figure 7.

These datasets collectively demonstrate the diverse domain challenges that our center-periphery attention refinement framework addresses, validating the effectiveness of our approach across various types of domain shifts.

Figure 2. Sample images from ArTaxOr dataset showing diverse arthropod species including insects, spiders, and crustaceans with varying scales and natural backgrounds.

3. Computational Efficiency Analysis

Our proposed framework demonstrates exceptional efficiency across multiple computational metrics. As shown in Table 1, our method maintains practical deployment feasibility while delivering significant performance improvements.

Zero-Parameter Design. Our method maintains exactly 232.53M total parameters and 123.33M trainable parameters across all configurations, demonstrating complete parameter efficiency. PPR and NCM operate through feature space manipulations without introducing learnable parameters, while TSA leverages existing BERT encoder and projection layers from the baseline GLIP, ensuring no additional parameter overhead.

Figure 3. Sample images from Clipart1k dataset showing cartoon-style objects with simplified visual features and artistic rendering.

Figure 4. Sample images from DIOR dataset featuring remote sensing objects such as airplanes, ships, and golf fields captured from satellite perspectives.

Memory Efficiency. The framework exhibits excellent memory efficiency, with cached memory increasing modestly from 3.09GB to 3.64GB (+18%) when all modules are activated. Meanwhile, allocated memory remains almost constant (1.95GB) across configurations, indicating stable runtime memory usage.

Computational Efficiency. FLOPs analysis confirms the lightweight nature of the proposed modules. PPR+NCM introduces only 1.3G additional FLOPs (+0.17%), TSA

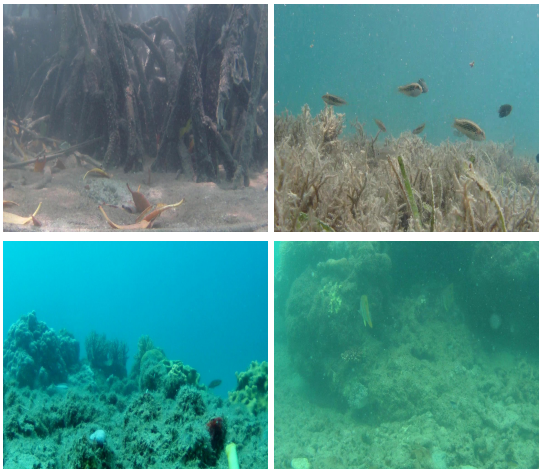

Figure 5. Sample images from DeepFish dataset featuring underwater fish detection with challenging lighting conditions and water distortion effects.

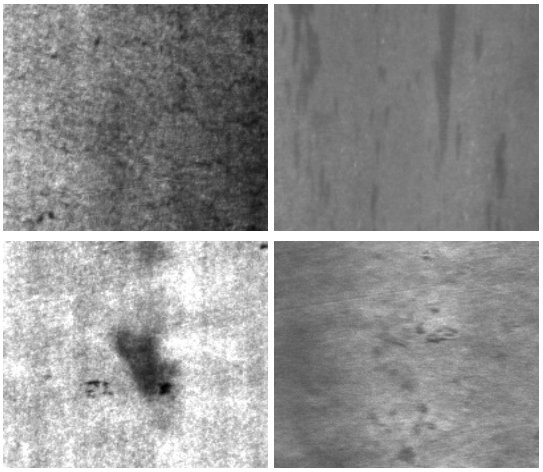

Figure 6. Sample images from NEU-DET dataset showing industrial surface defects with subtle visual patterns and varying scales.

adds 2.9G (+0.39%), and the full framework incurs merely 4.2G additional FLOPs (+0.56%) over the baseline. This reflects the efficiency of our prototype-based attention refinement compared to conventional attention mechanisms.

Practical Implications. The zero-parameter design enables seamless integration into existing architectures, while memory and computational costs are kept minimal. These advantages make our method suitable for real-world deployment scenarios, offering substantial gains in cross-domain detection performance with negligible overhead.

4. Detection Box Quality Analysis

To provide comprehensive evidence of our method's superior detection quality, we present detailed visualizations comparing bounding box predictions across different approaches. Figure 8 demonstrates our method's ability to generate precise, non-redundant detections while effectively suppressing background interference.

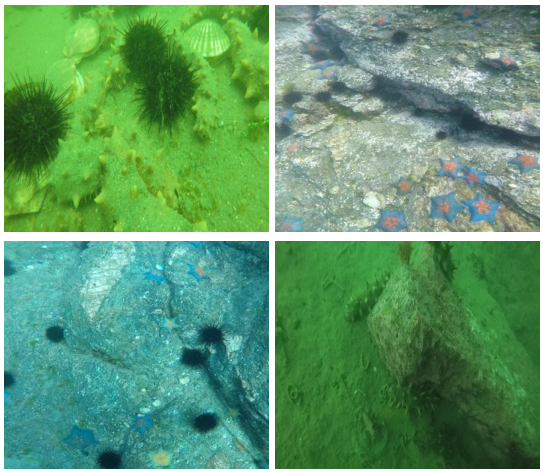

Figure 7. Sample images from UODD dataset containing underwater objects with complex backgrounds and varying visibility conditions.

The visualizations reveal several key improvements:

Precision Enhancement: Our method consistently produces tighter, more accurate bounding boxes that closely align with ground truth annotations. The Positive Pattern Refinement module effectively guides attention toward object boundaries, resulting in more precise localization compared to baseline approaches that often generate loose or imprecise boxes.

Redundancy Reduction: Our approach generates cleaner, non-redundant predictions by suppressing multiple overlapping detections for the same object. This improvement stems from our Negative Context Modulation module, which enhances object-background distinctions and reduces false positive activations.

Background Suppression: The comparisons show superior background suppression capability, where our method effectively reduces false positives in background regions. Our Textual Semantic Alignment module strengthens semantic understanding of object-background relationships, addressing the common issue of misclassifying background regions as objects.

These qualitative results remain consistent across different target domains, supporting our quantitative findings and confirming that our center-periphery framework effectively addresses the Astigmatism problem in cross-domain few-shot detection scenarios.

Method	Cached Memory (GB)	Total Params (M)	Trainable Params (M)	Allocated Memory (GB)	FLOPs (G)
Baseline	3.09	232.53	123.33	1.95	743.9
+PPR+NCM	3.13	232.53	123.33	1.96	745.2
+TSA	3.53	232.53	123.33	1.95	746.8
+PPR+NCM+TSA	3.64	232.53	123.33	1.95	748.1

Table 1. Computational Overhead Analysis

5. Background Text Design and Dataset Categories

Our Negative Context Modulation module leverages carefully crafted background text descriptions to enhance object-background discrimination. This section provides detailed information about the background text design strategy and the specific categories for each evaluated dataset.

5.1. Dataset Categories and Background Text Examples

ArTaxOr Dataset **Target Categories:** 7 arthropod orders including Araneae (spiders), Coleoptera (beetles), Diptera (flies), Hemiptera (true bugs), Hymenoptera (bees, wasps, ants), Lepidoptera (butterflies, moths), and Odonata (dragonflies, damselflies).

Representative Background Text Examples:

- *Natural habitats:* "leaf surface", "tree bark", "plant stem", "forest floor", "meadow grass"
- *Direct negations:* "not Araneae", "not a spider", "not Coleoptera", "not a beetle", "not Diptera"
- *Species negations:* "area without spiders", "region with no beetles", "space lacking flies"
- *Plant elements:* "green leaf", "plant texture", "flower", "twig", "branch"
- *Contextual negations:* "empty leaf without insects", "bare tree bark with no spiders", "clean flower petals without butterflies"
- *Natural environments:* "natural background", "habitat background", "ecological background"

CliPart1k Dataset **Target Categories:** 20 object classes including sheep, chair, boat, bottle, dining table, sofa, cow, motorbike, car, aeroplane, cat, train, person, bicycle, potted plant, bird, dog, bus, TV monitor, and horse.

Representative Background Text Examples:

- *Cartoon-specific elements:* "cartoon blank speech bubble", "clipart empty thought cloud", "illustrated vacant panel"
- *Direct negations:* "not a sheep", "definitely not a chair", "absolutely not a boat"
- *Environmental descriptions:* "empty cartoon theater", "vacant clipart stadium", "cartoon sky without objects"
- *Stylistic backgrounds:* "dotted pattern background", "striped cartoon backdrop", "geometric mosaic background"
- *Contextual negations:* "cartoon room without furniture or people", "clipart outdoor scene without vehicles or animals"

DIOR Dataset **Target Categories:** 20 object classes including Expressway-Service-area, Expressway-toll-station, airplane, airport, baseballfield, basketballcourt, bridge, chimney, dam, golffield, groundtrackfield, harbor, overpass, ship, stadium, storagetank, tenniscourt, trainstation, vehicle, and windmill.

Representative Background Text Examples:

- *Land cover types:* "vegetation", "forest", "grassland", "agricultural field", "bare soil"
- *Direct negations:* "not airplane", "not airport", "not ship", "not harbor", "not bridge"
- *Infrastructure negations:* "area without buildings", "region with no vehicles", "space lacking transportation infrastructure"
- *Satellite perspectives:* "satellite background", "aerial background", "remote sensing background", "earth surface background"
- *Contextual negations:* "forested area without structures", "agricultural land without vehicles", "water body without ships"
- *Natural landscapes:* "natural terrain", "undeveloped land", "pristine landscape"

FISH Dataset **Target Categories:** Multiple fish species in underwater environments.

Representative Background Text Examples:

- *Aquatic environments:* "underwater background", "marine background", "open water", "coral reef", "seagrass bed"
- *Fish negations:* "not a fish", "definitely not a fish", "area without fish", "completely fish-free zone"
- *Water conditions:* "clear water", "murky water", "shallow water", "deep water", "sunlit water"
- *Habitat descriptions:* "coral reef without fish", "seagrass bed with no fish", "underwater cave devoid of fish"
- *Contextual negations:* "underwater habitat that typically

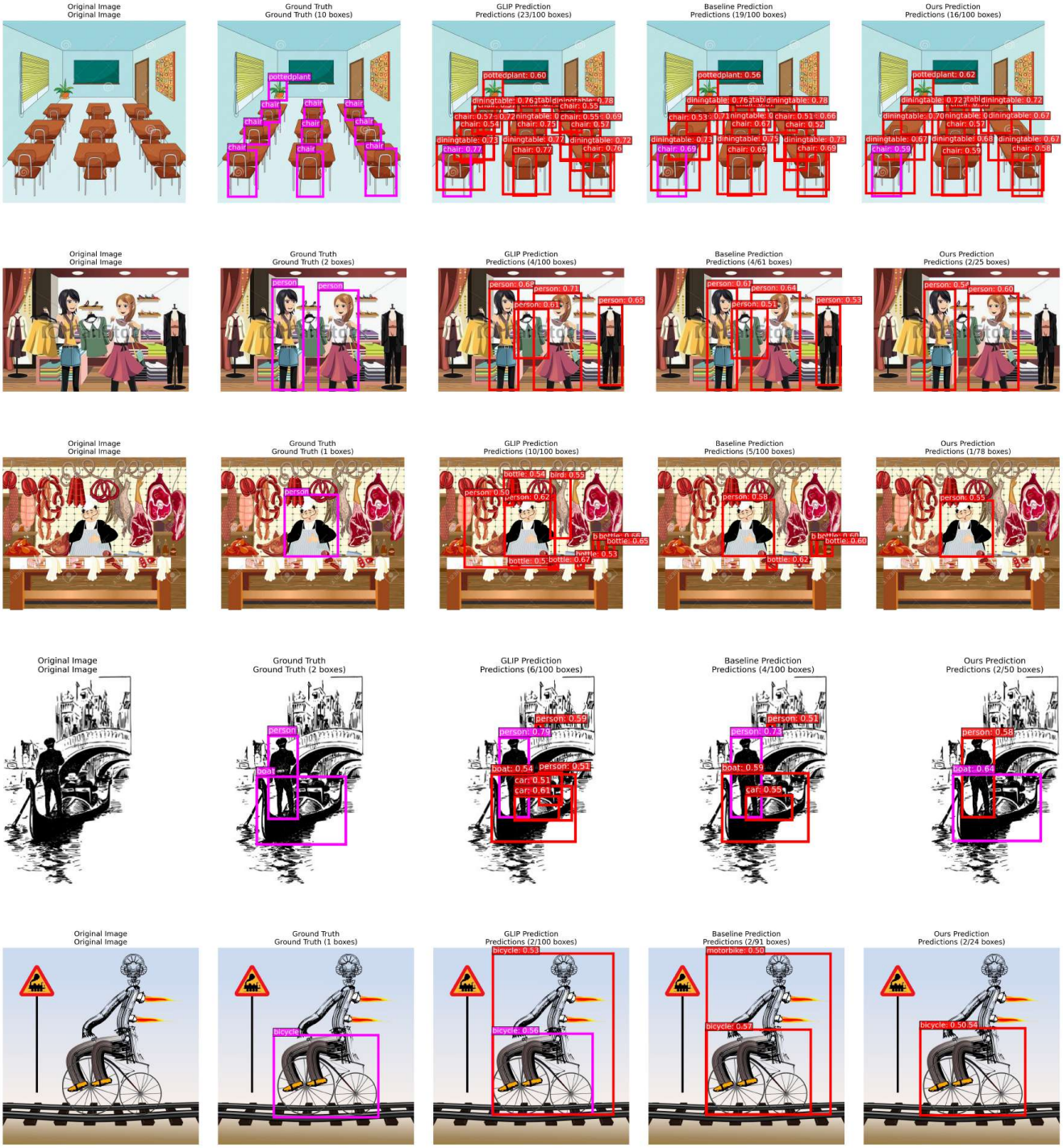

Figure 8. Comprehensive detection quality comparison across multiple test cases. Each row shows (from left to right): Original image, Ground truth annotations, GLIP predictions, Baseline predictions, and Our method predictions. Our approach consistently produces cleaner, more precise bounding boxes with significantly reduced false positives and better object-background separation across diverse scenarios and domains.

contains fish but currently empty", "natural fish shelter without occupants"

NEU-DET Dataset **Target Categories:** 6 steel surface defect types including crazing, inclusion, patches, pitted

surface, rolled-in scale, and scratches.

Representative Background Text Examples:

- *Surface quality:* "defect-free", "clean surface", "unblemished area", "standard surface", "flawless section"
- *Defect negations:* "not crazing", "definitely not inclusion", "absolutely not patches", "certainly not pitted surface"
- *Material descriptions:* "steel surface", "hot-rolled steel", "uniform texture", "metallic texture", "consistent grain pattern"
- *Quality indicators:* "standard production quality", "within-spec finish", "quality control passed surface"
- *Contextual negations:* "uniform steel surface without crazing", "smooth metal with no inclusions"

UODD Dataset **Target Categories:** 3 marine species including sea cucumber, sea urchin, and scallop.

Representative Background Text Examples:

- *Marine environments:* "sandy seafloor", "rocky ocean floor", "underwater terrain", "marine substrate"
- *Species negations:* "not a sea cucumber", "definitely not a sea urchin", "certainly not a scallop"
- *Habitat descriptions:* "area without sea cucumbers", "seafloor devoid of scallops", "marine environment without echinoderms"
- *Water conditions:* "murky water", "clear water background", "turbid water", "suspended particles in water"
- *Contextual negations:* "underwater habitat missing all target species", "seafloor with feeding trails but no sea cucumbers present"

6. Background Text Design Principles

Our background text design follows key principles to maximize the effectiveness of the Negative Context Modulation module:

Domain-Specific Adaptation: Background text is tailored to each dataset's visual domain, incorporating domain-relevant terminology (e.g., artistic elements for cartoons, manufacturing contexts for industrial scenes, aquatic terminology for marine environments, ecological descriptions for biological detection, geographical terms for remote sensing).

Multi-Level Negation Strategy: We employ various negation approaches including direct class negations ("not a sheep"), alternative phrasings ("area without sheep"), and contextual negations ("cartoon room without furniture"), ensuring robust background-foreground discrimination across different linguistic formulations.

Adaptive Text Selection via TSA: Given the extensive background text vocabulary, our Textual Semantic Alignment (TSA) module employs temperature-scaled softmax attention to dynamically select the most semantically relevant background descriptions for each image. This adaptive selection ensures that only the most contextually appropriate negative descriptions are emphasized during training, improving the efficiency and effectiveness of background-foreground discrimination.

The effectiveness of this design is demonstrated through our attention distance analysis and detection quality improvements, confirming that well-crafted negative context descriptions with adaptive selection significantly enhance cross-domain few-shot detection performance.

References

[1] Geir Drange. Arthropod taxonomy orders object detection dataset, 2020. 1

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

[3] Lihao Jiang, Yi Wang, Qi Jia, Shengwei Xu, Yu Liu, Xin Fan, Haojie Li, Risheng Liu, Xinwei Xue, and Ruili Wang. Underwater species detection using channel sharpening attention. In *Proceedings of the 29th ACM International Conference on Multimedia*, page 4259–4267, New York, NY, USA, 2021. Association for Computing Machinery. 2

[4] 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. 1

[5] Liunian Harold Li, Pengchuan Zhang*, Haotian Zhang*, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *CVPR*, 2022. 1

[6] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 1

[7] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 1

[8] 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*, 10(1), 2020. 2

[9] 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*, 285:858–864, 2013. 2