Learn From Human Eyes: Zero-Shot Recurring Pattern Detection on a Multi-Perception Benchmark

Shimian Zhang, Keaton Kraiger, Skanda Bharadwaj, Robert Collins, Yanxi Liu

Department of Computer Science and Engineering
The Pennsylvania State University

May 18, 2025

Overview

1. Introduction: Patterns, Symmetry, and Perception
2. The Challenge: Perceptual Diversity
3. Contributions
4. Evaluation
5. Comparison with MLLMs
6. Qualitative Results & Applications
7. Conclusion

Motivation: The Spectrum of Repetition

- Visual world is rich with repetition.
- From perfect symmetry to abstract recurrence.
- Humans perceive structure intuitively.

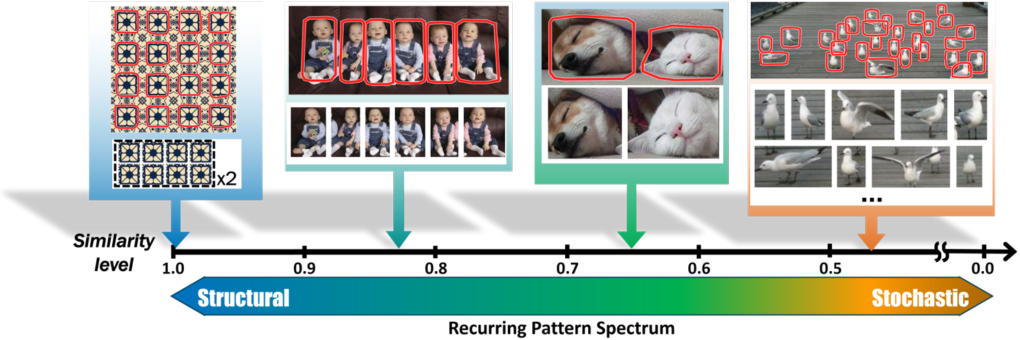


Figure: Recurring Pattern (RP) Spectrum

Defining Recurring Patterns (RPs)

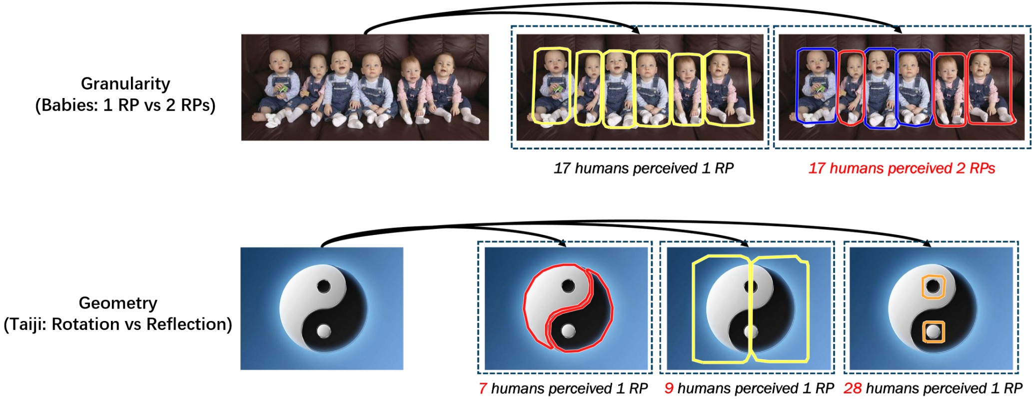
- **Base Definition:** RP $\mathcal{R} = \{I_1, \cdots, I_n\}$: Set of n recurring RP Instances [1, 2].
- **Generalized Symmetry:** RPs encompass symmetry but allow more flexibility.
- **Transformation Flexibility:** Instances $I_j \approx g \circ I_i$, where g can be rigid or non-rigid (e.g., $g \in \mathrm{Diff}(\mathbb{R}^2)$).
- **Perceptual Similarity:** Core criterion based on feature distance d_ϕ after alignment:

$$\min_{g \in \text{AllowedTransforms}} d_\phi(g \circ I_i, I_j) < \tau$$

(τ: similarity threshold, d_ϕ: perceptual distance e.g., DINO [3])

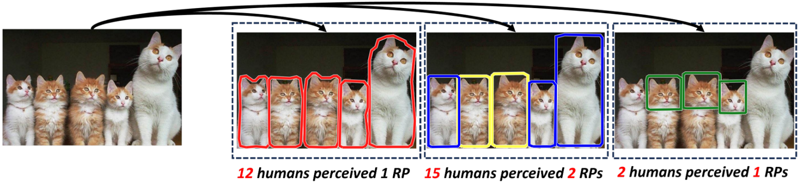
Challenge: Subjective & Diverse Human Perception

- Human perception is subjective [4, 5].
- Same image → Different perceived patterns/symmetries.
- How to define/model consensus?

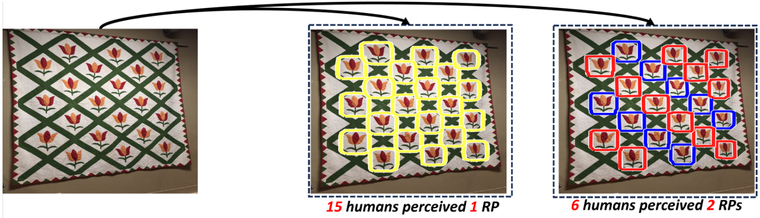


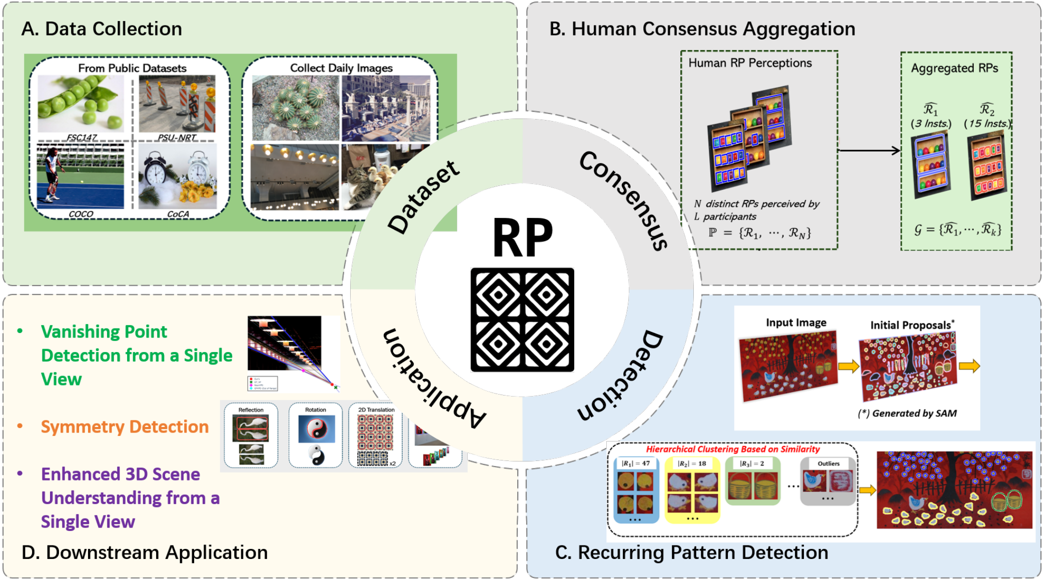
- Human perception is subjective [4, 5].
- Same image $\rightarrow$ Different perceived patterns/symmetries.
- How to define/model consensus?

Semantics
(Cats: Color vs Gaze)



Detail Sensitivity
(Patterns)




A. Data Collection
From Public Datasets
FSC147
PSU-NRT
COCO
CoCA
Collect Daily Images
B. Human Consensus Aggregation
Human RP Perceptions
N distinct RPs perceived by L participants
$\mathbb{P} = \{\mathcal{R}_1, \cdots, \mathcal{R}_N\}$
Aggregated RPs
$\widehat{\mathcal{R}}_1$ (3 Insts.)
$\widehat{\mathcal{R}}_2$ (15 Insts.)
$\mathcal{G} = \{\widehat{\mathcal{R}}_1, \cdots, \widehat{\mathcal{R}}_k\}$
Dataset
Consensus
Detection
Application
RP
Vanishing Point Detection from a Single View
Symmetry Detection
Reflection
Rotation
2D Translation
Enhanced 3D Scene Understanding from a Single View
D. Downstream Application
Input Image
Initial Proposals*
(*) Generated by SAM
Hierarchical Clustering Based on Similarity
$|R_1| = 47$
$|R_2| = 18$
$|R_3| = 2$
Outliers
C. Recurring Pattern Detection

Contribution 1: Multi-Perception RP Dataset

- First large-scale dataset based on diverse human RP perceptions.

- **Key Stats:**
  - 4,625 Images
  - 272 Participants
  - $>$220k Unique Perceptions
  - 12,125 Aggregated RPs
  - $\sim$30 Perceptions/Image

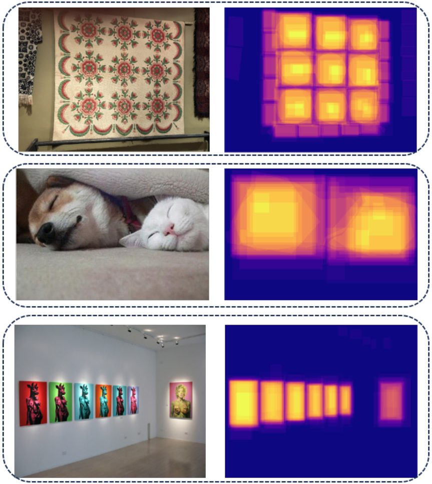


Figure: Perception Examples

Dataset: Collection & Quality Control

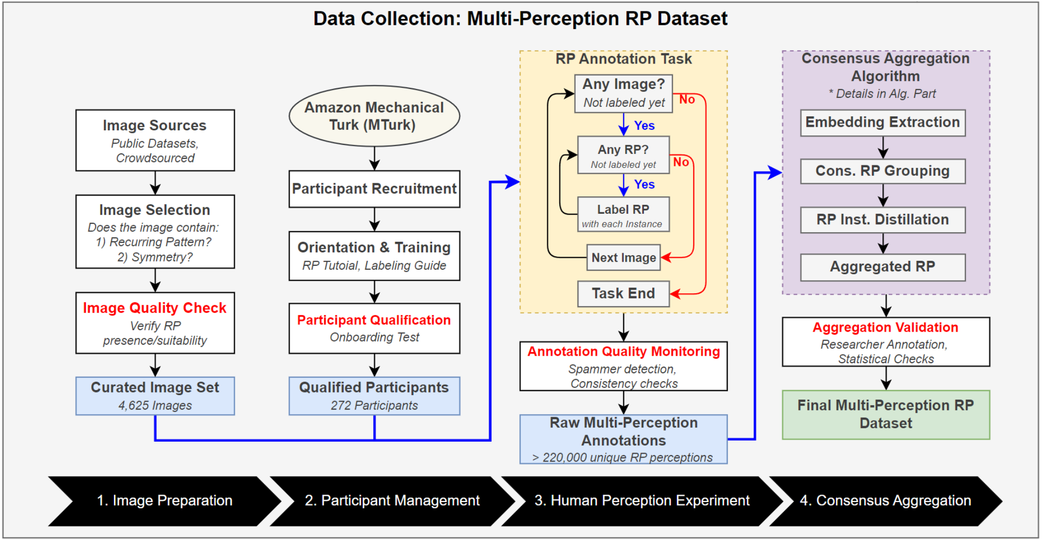


Figure: Data Collection Pipeline

Contribution 2: RP Consensus Aggregation Algorithm

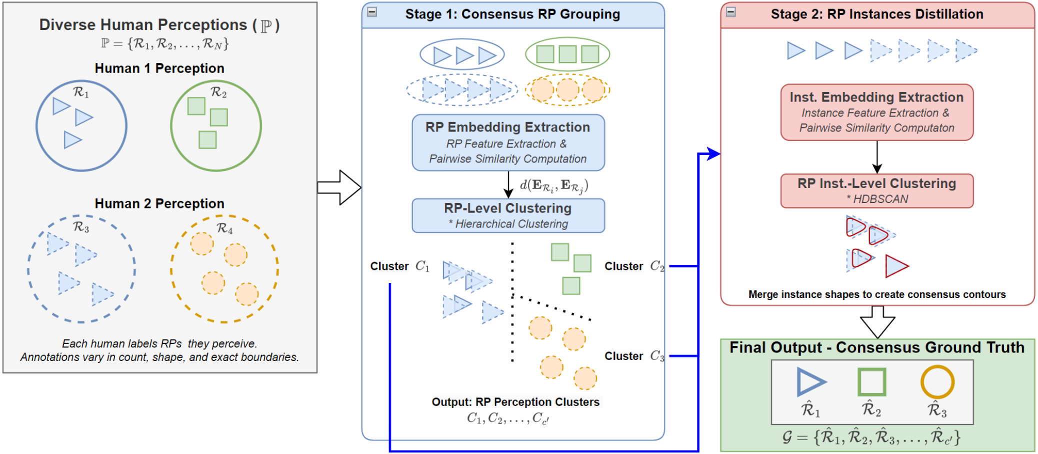


Figure: RP Consensus Aggregation Algorithm

Contribution 2: RP Consensus Aggregation Algorithm (Cont.)

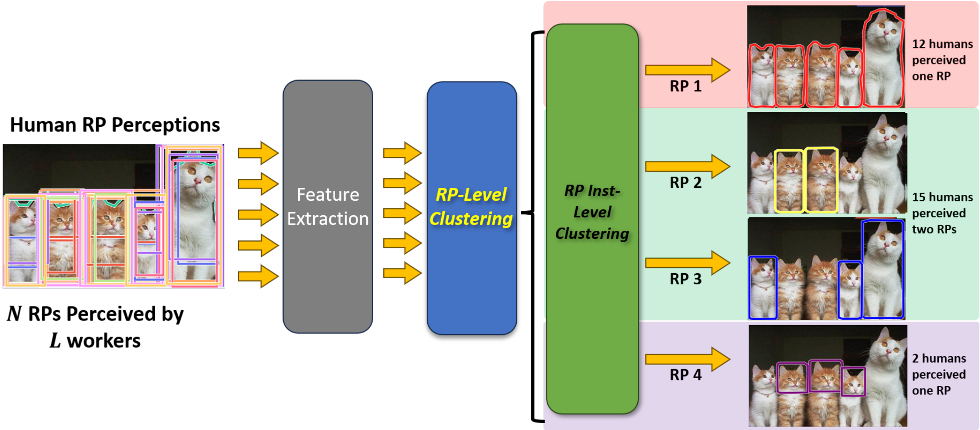


Figure: A Real Example of RP Consensus Aggregation Algorithm

Contribution 3: Zero-Shot RP Detection Approach

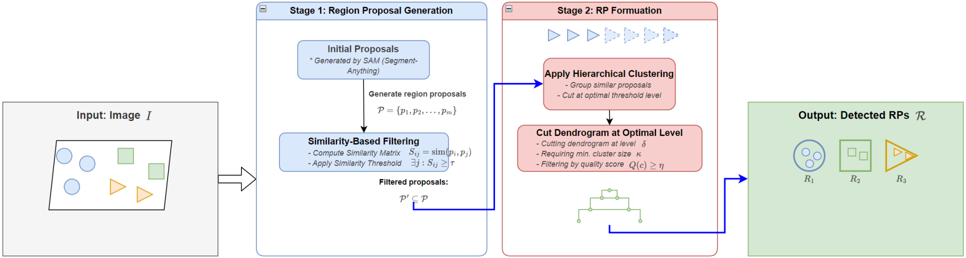


Figure: RP Detection Pipeline

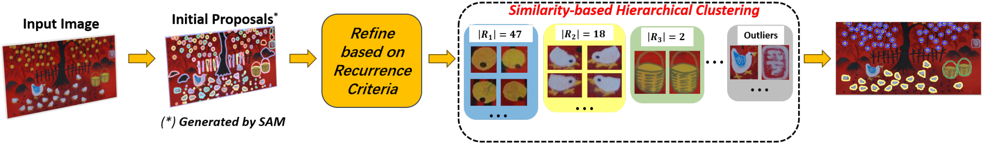


Figure: A Real Example of RP Detection

Evaluation Protocol: A Two-Level Approach

- Standard metrics (e.g., for object detection [6, 7]) insufficient for RPs (class-agnostic, pattern focus).
- Adopt two-level evaluation [2]:

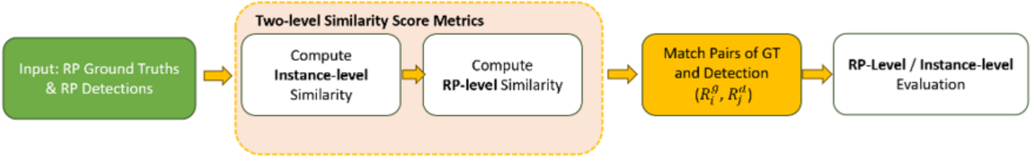


Figure: Two-Level Evaluation

- Alignment using weighted IoU (wIoU) [8].

Quantitative Results

Comparisons:

- Unsupervised Object Detection (CutLER [9])
- Traditional Methods [1, 2]
- Our Zero-Shot Method

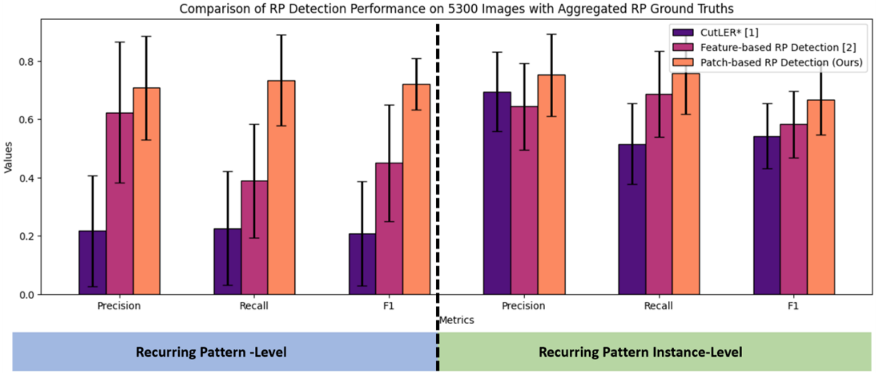


Figure: Performance Evaluation

Comparison with Multi-Modal LLMs (MLLMs)

Goal: Evaluate MLLMs on zero-shot RP detection.

Models Tested:

- GPT-4o / Gemini 2.5 Pro

```
Terminology:
An RP, denoted as $\mathcal{R}$, is a set of $n$ elements, each referred to as an \textit{RP Instance} $I_{i=1}^n$, that that
reoccur in some feature space. Formally, this can be expressed as $\mathcal{R} = \{I_1, \cdots, I_n\}$, where each $I_i$
represents an individual occurrence of the pattern.
Given an image, a human subject, $s$, can identify $c$ distinct RPs within this image. We denote the subject's perceived set of
RPs as $\mathcal{P}_s = \{\mathcal{R}_1, \cdots, \mathcal{R}_c\}$, where each $\mathcal{R}_i$ is an RP perceived by the subject.
RP perceptions of all $L$ subjects for the image are denoted by the set $\mathbb{P}$.

Prompt:
Please identify all recurring patterns you see in this image (image is already segmented into many instance proposals with
number ids), and describe them in a JSON format, like:
{
   "image_description": "A image of two humans standing in front of a building.",
   "RP1": {"rp_keywords: "Humans", "num_instances": 2, "instance_ids": [2, 4]},
   "RP2": {"rp_keywords: "Windows", "num_instances": 15, "instance_ids": [11, 12, 14, 15, ...]},
   ...
}
```

Figure: Consistent Prompting Across Models

Comparison with Multi-Modal LLMs (MLLMs) (Cont.)

Goal: Evaluate MLLMs on zero-shot RP detection.

Models Tested:

- GPT-4o

Figure: A Bad Case of GPT-4o on Detecting RP

Models Compared:

- Our Approach / Gemini 2.5 Pro

Figure: A Comparison of Our Approach vs. Gemini 2.5 Pro

Downstream Application: Symmetry Analysis

- RPs provide primitives for symmetry detection.
- Detects Reflection, Rotation, Translation.
- Naturally handles **near-symmetry**.
- Enables 3D translation symmetry perception from single view (via cross-ratio) [2].

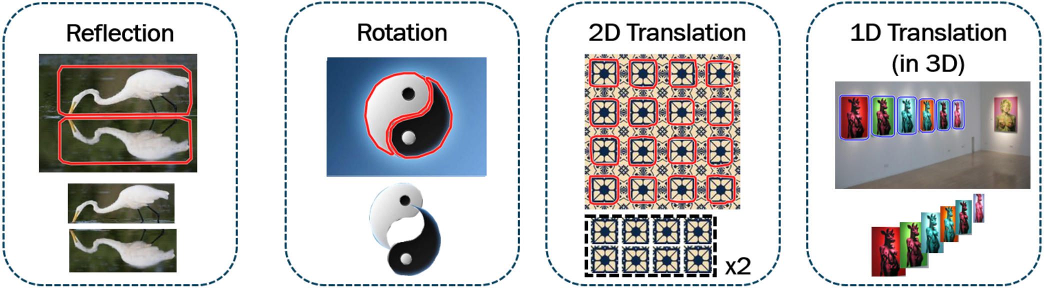


Figure: Symmetry Detection

Downstream Application: Vanishing Point Detection

- RPs provide implicit correspondences for Vanishing Point (VP) detection.
- Alternative to explicit line detection [10].
- More robust in scenes with weak geometric cues or occlusion.

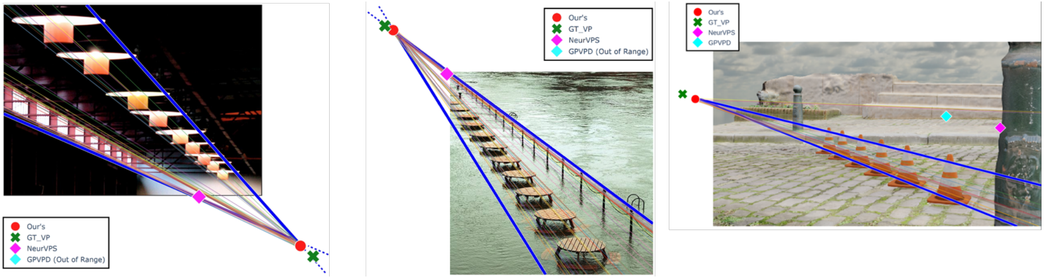


Figure: Vanishing Point Detection

Downstream Application: Counting & Scene Understanding

- **Class-Agnostic Counting:** RP grouping inherently counts similar items without pre-defined classes [11].

- **Enhanced Scene Understanding:** Combine RPs + Symmetry + VPs for richer descriptions.

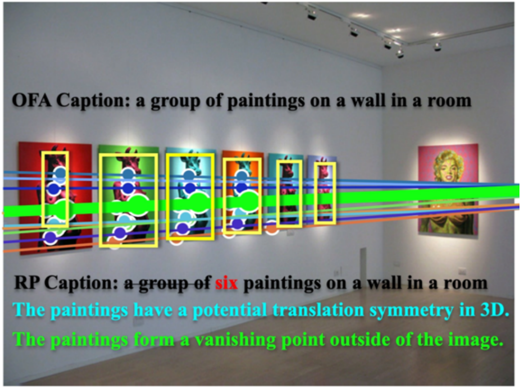


Figure: 3D Scene Understanding from Single View

Contributions Summary:

- Formalized RPs as generalized symmetry, studied perceptual diversity.
- Created first multi-perception RP dataset & benchmark.
- Developed Consensus Aggregation for robust GT generation.
- Proposed a Zero-Shot RP Detection method.
- Demonstrated applications in symmetry analysis & scene understanding.

Main Message: Bridging human visual intelligence and computational perception.

Future Work:

- Deeper integration with reasoning models (LLMs/MLLMs) [12].
- Exploring RP hierarchies explicitly.
- Applications in new domains (robotics, medical).

Thank You!

Contact: svz5303@psu.edu

Acknowledgements: This research is funded by NSF Grants 1909315

References I

[1] Jingchen Liu and Yanxi Liu. "Grasp recurring patterns from a single view". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2013, pp. 2003–2010.

[2] Shimian Zhang et al. "Novel 3D Scene Understanding Applications From Recurrence in a Single Image". In: *arXiv preprint arXiv:2210.07991* (2022).

[3] Maxime Oquab et al. "Dinov2: Learning robust visual features without supervision". In: *arXiv preprint arXiv:2304.07193* (2023).

[4] Evan Heit and Brett K Hayes. "Relations among categorization, induction, recognition, and similarity: Comment on Sloutsky and Fisher (2004).". In: (2005).

[5] Yanxi Liu et al. "Computational symmetry in computer vision and computer graphics". In: *Foundations and Trends® in Computer Graphics and Vision* 5.1–2 (2010), pp. 1–195.

References II

[6] Mark Everingham et al. “The pascal visual object classes (voc) challenge”. In: *International journal of computer vision* 88 (2010), pp. 303–338.

[7] Tsung-Yi Lin et al. “Microsoft coco: Common objects in context”. In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer. 2014, pp. 740–755.

[8] Yeong-Jun Cho. “Weighted intersection over union (wIoU): a new evaluation metric for image segmentation”. In: *arXiv preprint arXiv:2107.09858* (2021).

[9] Xudong Wang et al. “Cut and learn for unsupervised object detection and instance segmentation”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, pp. 3124–3134.

[10] Siyuan Huang et al. “Holistic 3d scene parsing and reconstruction from a single rgb image”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 187–203.

References III

[11] Viresh Ranjan et al. "Learning to count everything". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 3394–3403.

[12] Haotian Liu et al. "Visual instruction tuning". In: *Advances in neural information processing systems* 36 (2024).