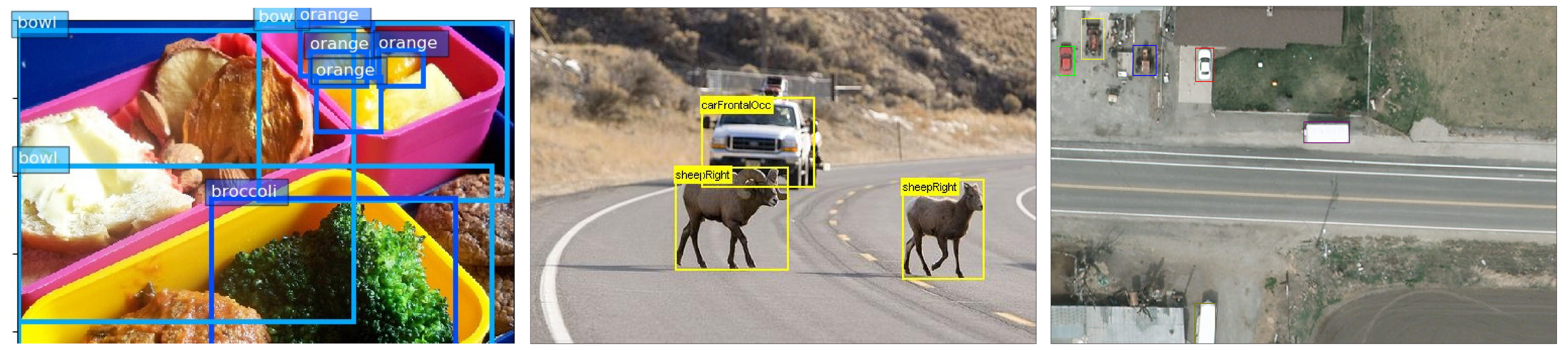
Localizing Grouped Instances for Efficient Detection in Low-Resource Scenarios

Amélie Royer, Christoph H. Lampert
Institute of Science and Technology Austria, Klosterneuburg, Austria

Motivation

Detection in varying scale and sparsity scenarios ?

- Two-stage detectors (e.g. Faster-RCNN [1])
  $\longrightarrow$ exhaustive but slow
- One-pass grid-based models (e.g. YOLO [2])
  $\longrightarrow$ Input resolution must account for small objects
- Trade-off accuracy vs memory/computational efficiency (e.g., applications involving embedded systems)

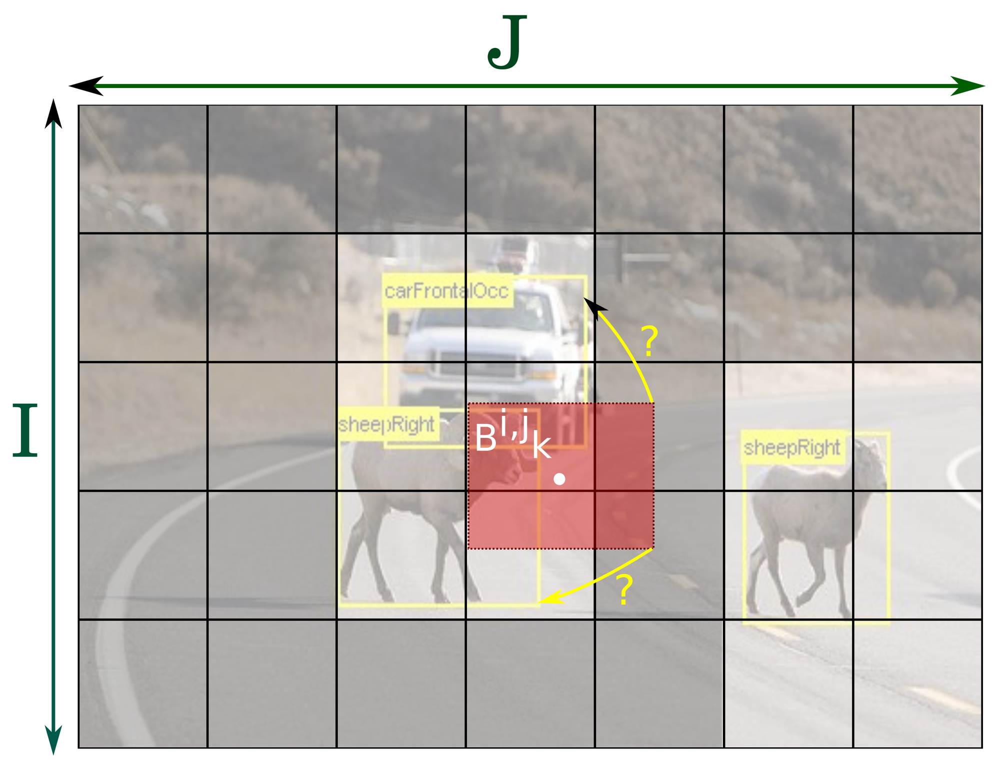


High density and saliency ←- - - - - - - - - - - - - -→ Sparse and small objects

Objective: Design a detection scheme that makes use of group structures in the image to curate few, meaningful, proposals

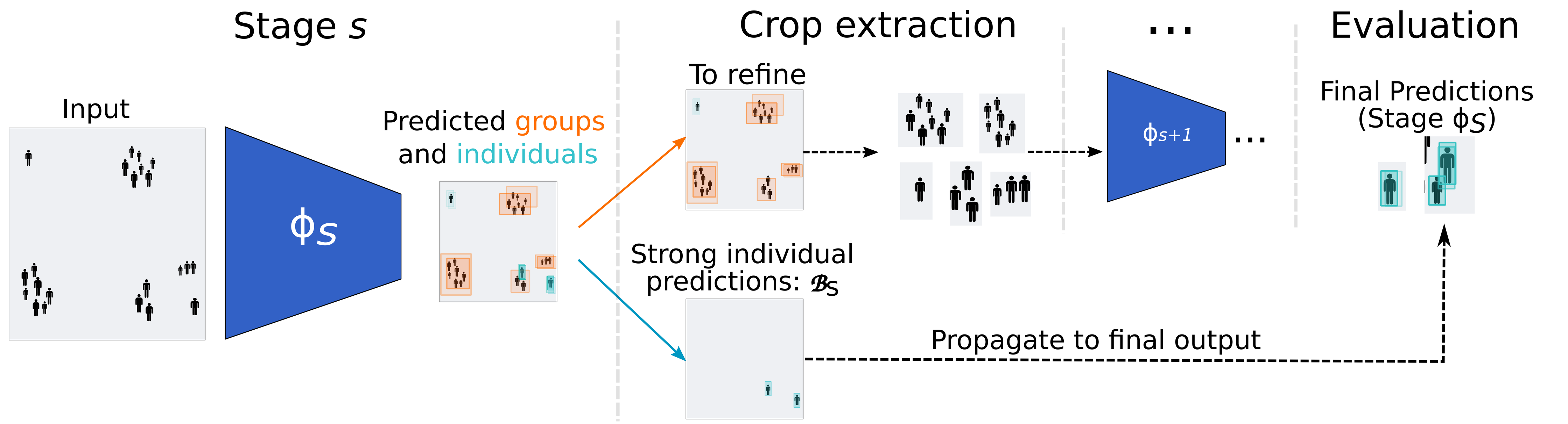
One-stage detectors

Based on **YOLO**: fast (one-pass) yet dense (grid-based)



- In each cell in the $(\mathbf{I}, \mathbf{J})$ grid, output **K** boxes $B_k^{i,j}$
- In **empty cells**, $c_k^{i,j} \to 0$
- In **active cells**, assignment $A(B_k^{i,j}, b^*) \in \{0, 1\}$. $c_k^{i,j} \to \mathtt{iou}(B_k^{i,j}, b^*)$ and match coordinates

Model Overview



Proposed Solution

How is the assignment A defined ?

$$A(B_k^{i,j}, b^*) = [\![\, b^* \text{ in cell } (i,j) \,]\!] \; [\![\, k = \arg\max_{k'} \; \mathtt{iou}(B_{k'}^{i,j}, b^*) \,]\!]$$

The total loss gathers contributions from all **empty cells** (push confidence scores to 0) and **active cells** (match assigned bounding boxes to target coordinates and confidences to iou)

Shortcomings:

- Grid parameters $(\mathbf{I}, \mathbf{J}, \mathbf{K})$ need to match the data resolution
- Choosing **K** is non-trivial. Too **high** (imbalance between empty and active cells) or too **low** (assignment collisions)

Proposed solution, ODGI:

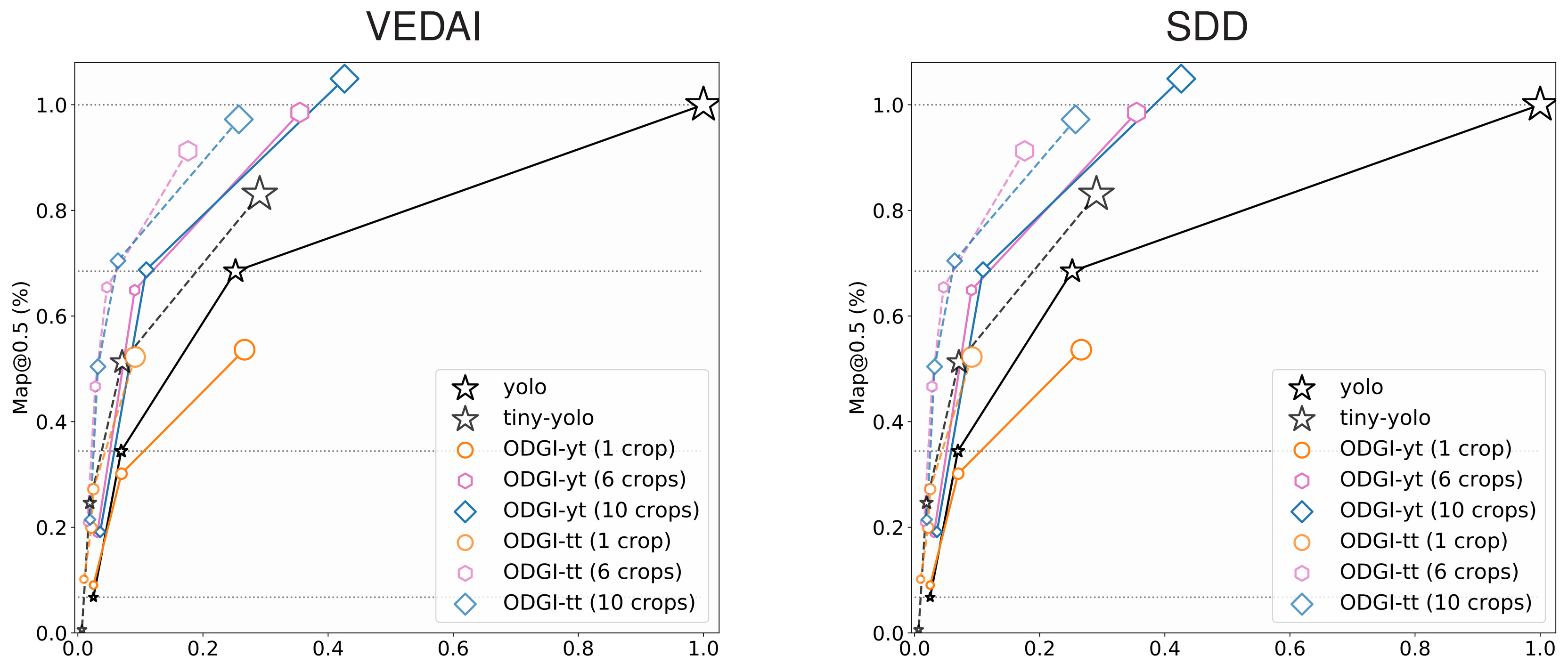
Define **groups of objects** as intermediary structures that are easier to detect at low resolution and can be refined if needed. New group assignment

$$A = \arg\max_{A'} \sum_{i,j,k} \sum_{q \in \{1 \dots |B|\}} \sum_{b^* \in \mathcal{P}^q(B)} A'(B_k^{i,j}, \bigcup b^*) \; \mathtt{iou}(B_k^{i,j}, \bigcup b^*)$$

- Hard to solve (exhaustive search), inefficient training procedure. We choose to fix $\mathbf{K} = 1 \to$ intuitive interpretation
- Ground-truth group flag is defined as $\mathtt{grp}(\bigcup b^*) = [\![\, |b^*| > 1 \,]\!]$

Experiments

- Accuracy (map@0.5) vs efficiency (runtime) trade-off
- Two **aerial views** datasets (VEDAI and SDD), 1024x1024 px
- Three backbones: tiny-YOLOv2, YOLOv2, MobileNet



- **Ablation experiment (1)**: No groups (∼ two stages YOLO). Achieves lower map, though compensated by the learned offsets.
- **Ablation experiment (2)**: No offsets or fixed offsets.

Conclusions

- **[+]** Focusing on meaningful patches defined by groups allows to start at lower resolution and avoid unnecessay regions
- **[+]** Stage transition parameters can be changed without retraining to match the application at hand
- **[-]** Sequential yet not perfectly end-to-end (joint training)

[1] S. Ren et al., *"Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks"*. NIPS, 2015.
[2] J. Redmon et al, *" You Only Look Once: Unified, Real-Time Object Detection"*. CVPR, 2016.