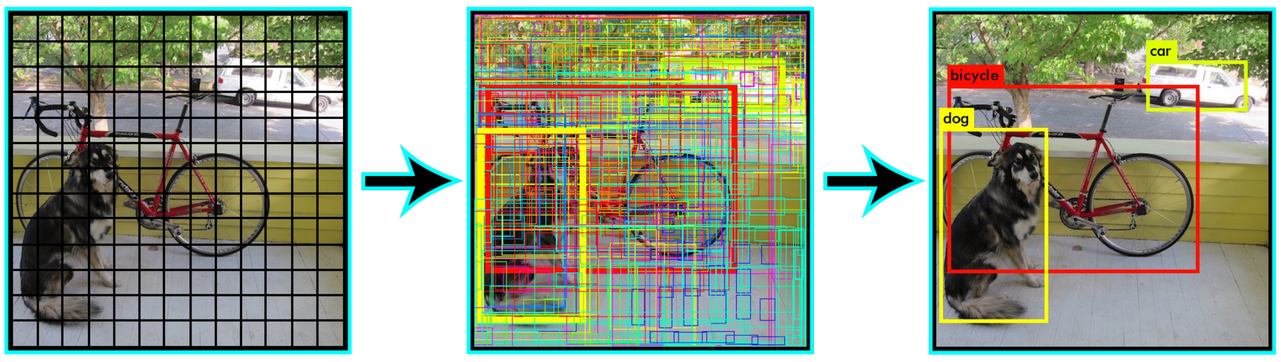
Detective: An Attentive Recurrent Model for Sparse Object Detection

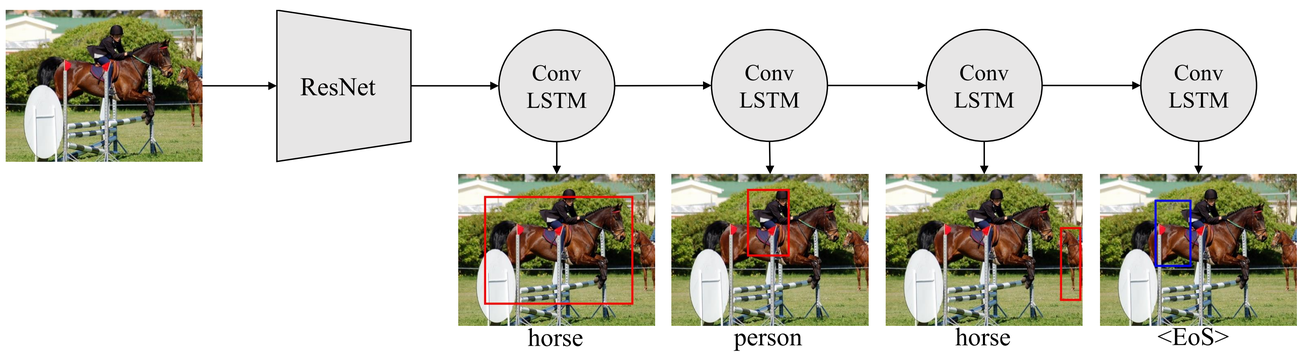
Amine Kechaou, Manuel Martinez, Monica Haurilet, Rainer Stiefelhagen

Computer Vision for Human-Computer Interaction (cv:hci)
Karlsruhe Institute of Technology, Germany
cvhci.anthropomatik.kit.edu

- **Dense Object Detection:** many object predictions are made, and post-processing is needed to fuse and filter them, e.g., YOLO:



- **Sparse Object Detection:** the network generates exactly one detection per object, e.g., **Detective**



Detective is an end-to-end sparse object detector using a ConvLSTM and attention to detect one object each iteration, until the End of Sequence token is predicted.

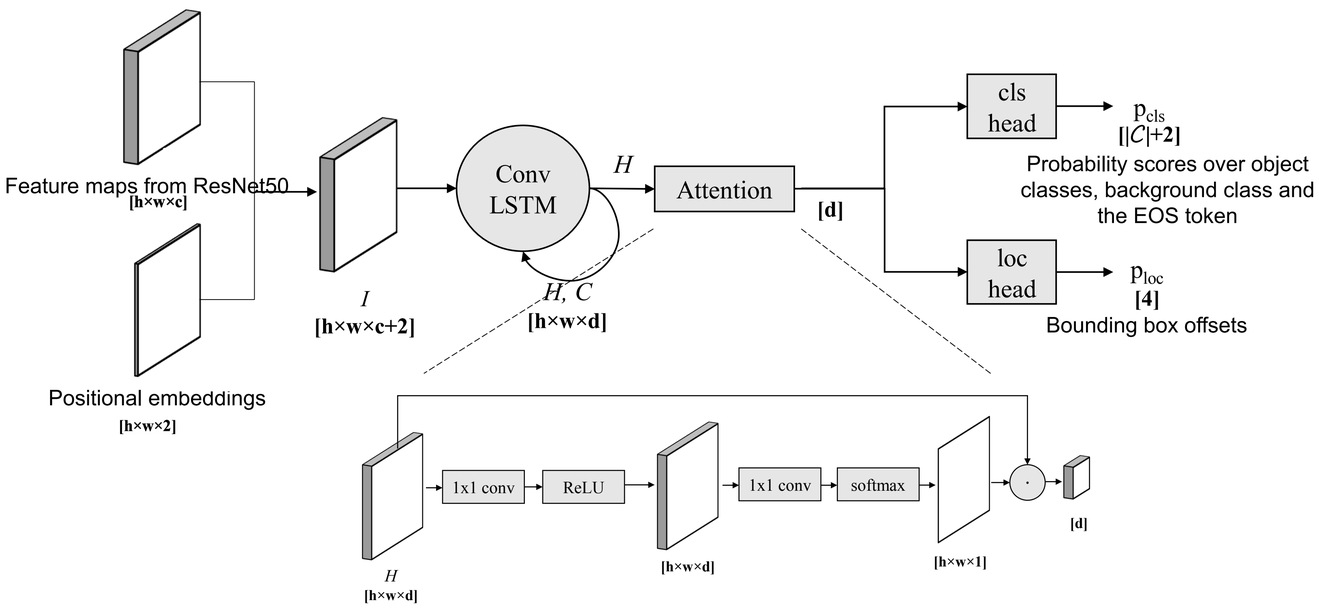
Advantages

- Raw detections are final detections
→ No post-processing required
- Predictions are not made densely over the image
→ No more foreground / background class imbalance
- Objects are predicted one by one
→ Enable better understanding at the instance level
- Possibility for Ordered Object Detection
→ Objects are predicted in the right order in one pass

Challenges

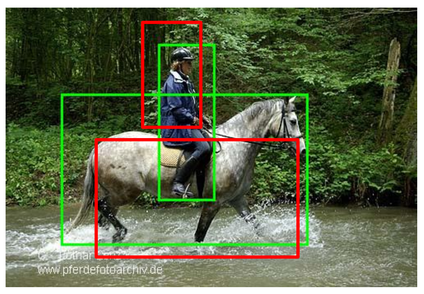
- Assigning target objects to predictions: target objects do not have a predefined order
- Limited related work on sparse object detection
- Combination of classification and localization is challenging for RNNs

Encoder-Decoder Architecture:



Dynamic Matching of Targets and Predictions During Training with the Hungarian Algorithm

- **Challenge:**
How to assign target labels to predictions during training?



- **Solution:**
Dynamic Matching using the Hungarian Algorithm
- **Cost function:**
$$f(t,p) = \mu_{cls}\mathcal{L}_{cls}(t_{cls}, p_{cls}) + \mu_{loc}\mathcal{L}_{loc}(t_{loc}, p_{loc})$$
- The Hungarian Algorithm efficiently finds the matching with the minimal overall cost

Ablation Study on PASCAL VOC

TABLE I
Results of Detective under different decoder settings

Model	Conv-LSTM	Attention	Positional embeddings	Background classification	mAP (%) on VOC07
Baseline	✗	✗	✗	✗	40.7
	✗	✓	✓	✓	49.4
	✓	✗	✓	✓	46.5
	✓	✓	✗	✓	47.9
	✓	✓	✓	✗	49.9
Detective	✓	✓	✓	✓	**52.0**

Examples, and their attention maps each iteration