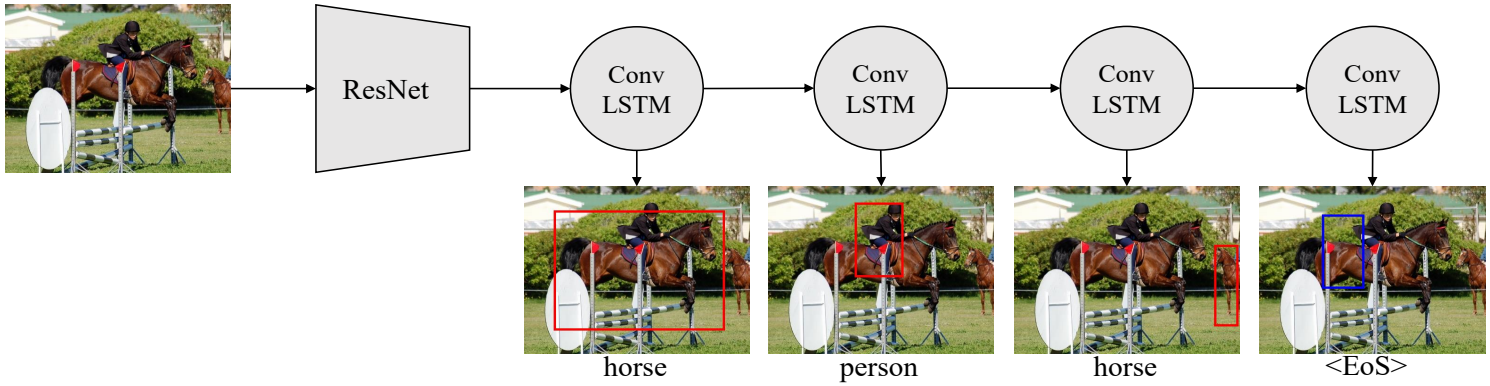
Karlsruhe Institute of Technology

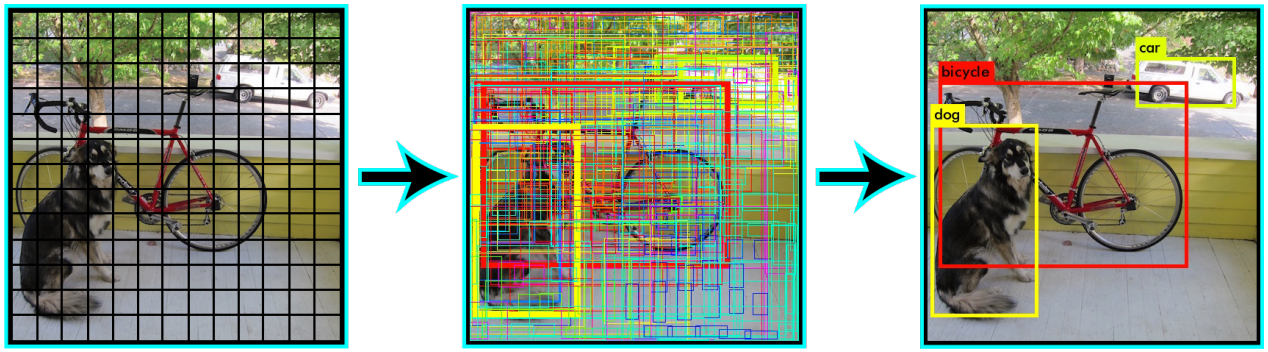
Detective: An Attentive Recurrent Model for Sparse Object Detection

Amine Kechaou, Manuel Martinez, Monica Haurilet, Rainer Stiefelhagen

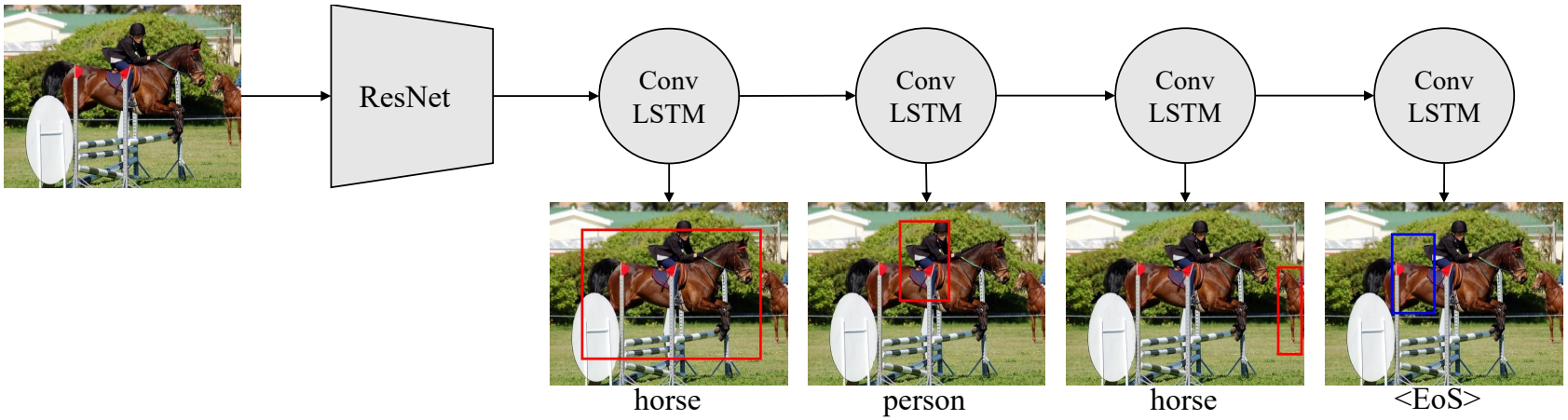
Computer Vision for Human-Computer Interaction Lab
Karlsruhe Institute of Technology



Motivation



Example of a dense object detector: YOLO v1 (**Redmon *et al.* 2016**)



Detective: an attentive sparse object detector

Advantages

- Raw detections are final detections
  → No post-processing required

- Objects are predicted one by one
  → Enable better understanding at the instance level

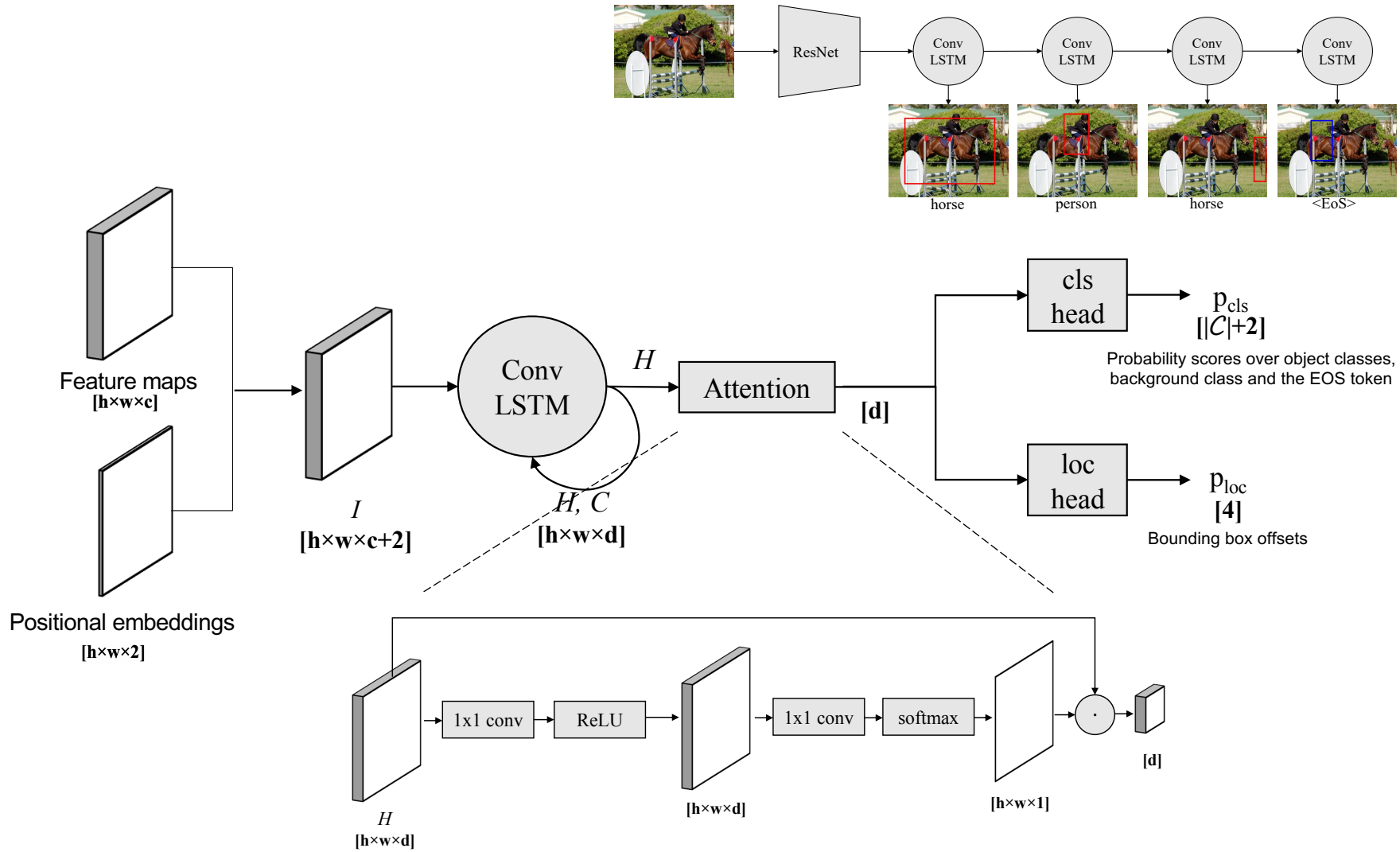
- Predictions are not made densely over the image
  → No more foreground / background class imbalance

- Possibility for Ordered Object Detection
  → Objects are detected in the right order in one pass

Challenges

- Object detection is framed as an unordered problem
  - Target objects' set is unordered
  - How to assign target objects to predictions during training?
- Sparse localization is challenging
  - Limited related work
- RNNs have been rarely used for the combination of localization and classification
  - Balancing localization and classification is challenging

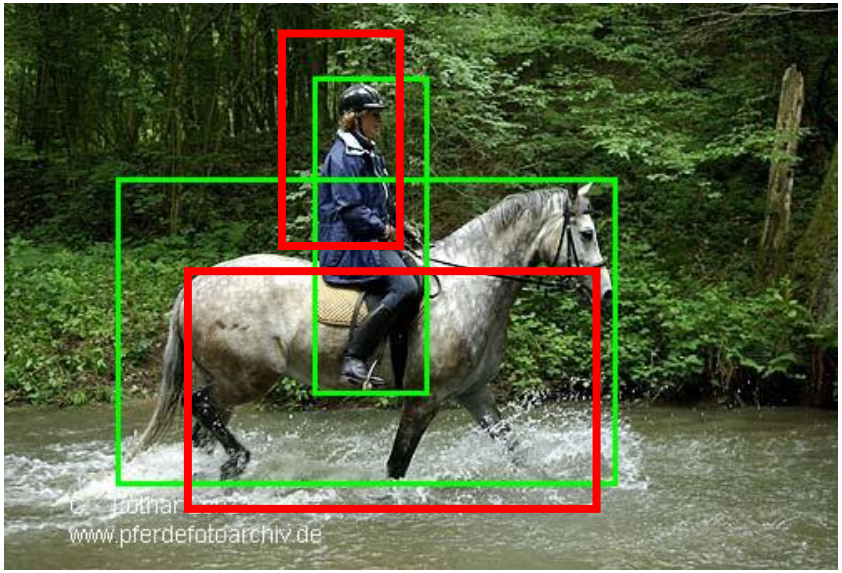
Decoder Architecture



Detective: An Attentive Recurrent Model for Sparse Object Detection
Amine Kechaou, Manuel Martinez, Monica Haurilet, Rainer Stiefelhagen

Dynamic Matching

- How to assign target objects to predictions during training?
- Dynamic matching: match target objects (T) to closest predictions (P)
  - Cost function: $f(t,p) = \mu_{cls}\mathcal{L}_{cls}(t_{cls}, p_{cls}) + \mu_{loc}\mathcal{L}_{loc}(t_{loc}, p_{loc})$
  - Find the matching function g: $T \longrightarrow P$ such that $\sum_{t \in T} f(t, g(t))$ is minimal
  - Hungarian algorithm efficiently finds the optimal solution



Loss Function

- Targets / predictions mapping: $M = \{(t,p) \in T \times P \mid p = g(t) \forall t \in T\}$
- Loss: $L = \lambda_{cls} L_{cls} + \lambda_{loc} L_{loc}$ where:

$$L_{cls} = \sum_{(t,p) \in M_{FG}} \mathcal{L}_{cls}\left(t_{cls}, p_{cls}\right) + \sum_{(t,p) \in M_{BG}} \mathcal{L}_{cls}\left(BG, p_{cls}\right) + \mathcal{L}_{cls}\left(EoS, p_{cls}^{(last)}\right)$$

and

$$L_{loc} = \sum_{(t,p) \in M} \mathcal{L}_{loc}\left(t_{loc}, p_{loc}\right)$$

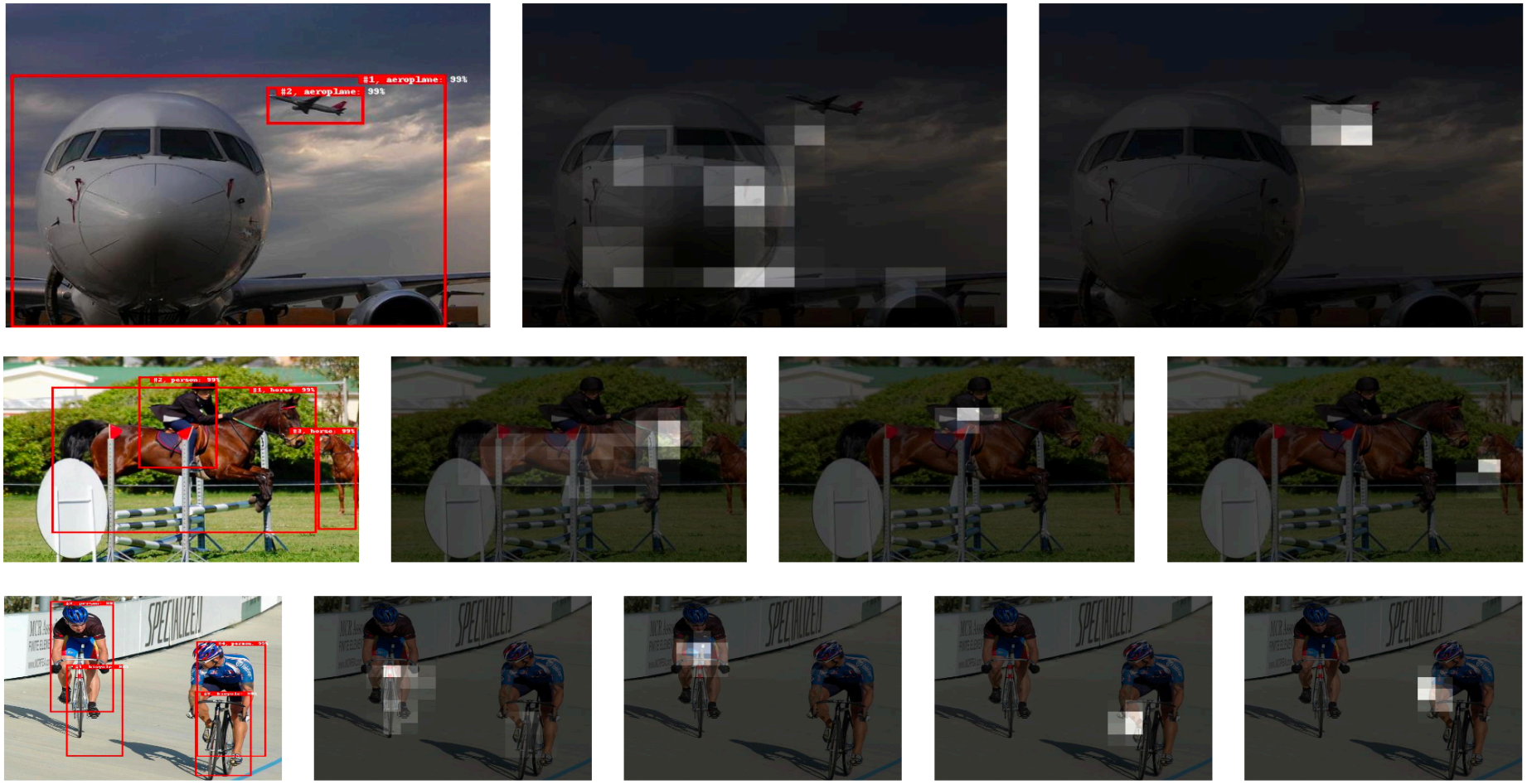
Evaluation
Ablation Study

- Dataset: PASCAL VOC
  - Training set: `trainval07+12`
  - Test set: `test07`
- Metric: mean Average Precision (mAP)

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**

Effects of Attention



Detective: An Attentive Recurrent Model for Sparse Object Detection
Amine Kechaou, Manuel Martinez, Monica Haurilet, Rainer Stiefelhagen

Conclusion

- Detective: sequential detection of object instances in an image
  - Architecture: ConvLSTM + Attention
  - Post-processing (e.g., NMS) **not required**

- Evaluation
  - Promising results on PASCAL VOC
  - ConvLSTM + Attention add spatial awareness
  - Ability to reason at the **instance** level