Multi-view X-ray R-CNN

Jan-Martin O. Steitz, Faraz Saeedan, and Stefan Roth

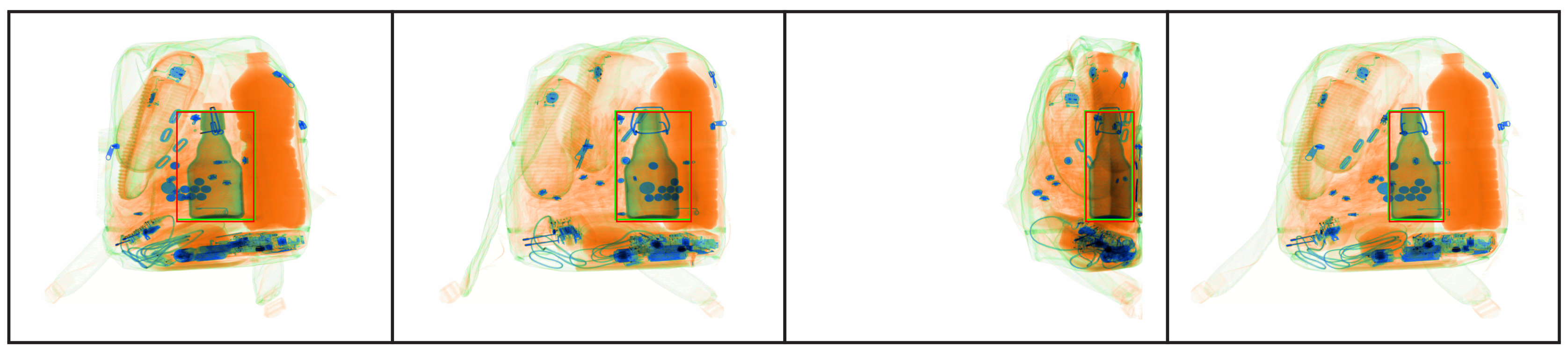
TECHNISCHE UNIVERSITÄT DARMSTADT

Goal

- **Detect prohibited objects** in X-ray recordings of luggage
- Utilize **multi-view information**
- Leverage features from **pre-trained deep CNN-backbones**

Contributions

- **Multi-view pooling** layer for 3D aggregation of 2D features
- End-to-end trainable **multi-view detection pipeline**

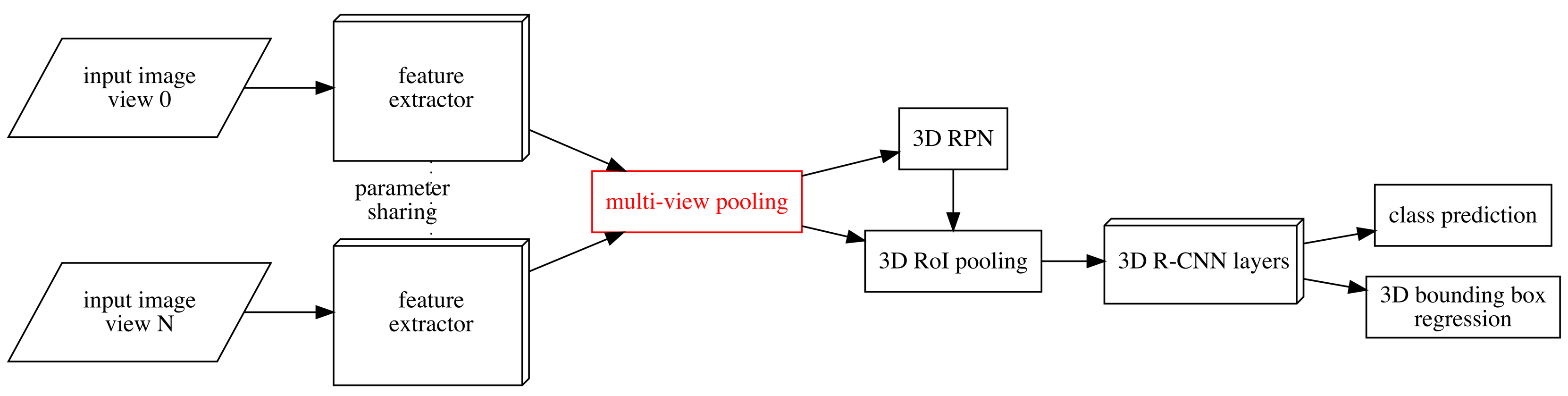


Example of multi-view X-ray images of hand luggage. Ground-truth (*green*) and detected (*red*) bounding boxes (detection confidence of 99.2 %).

MX-RCNN Architecture

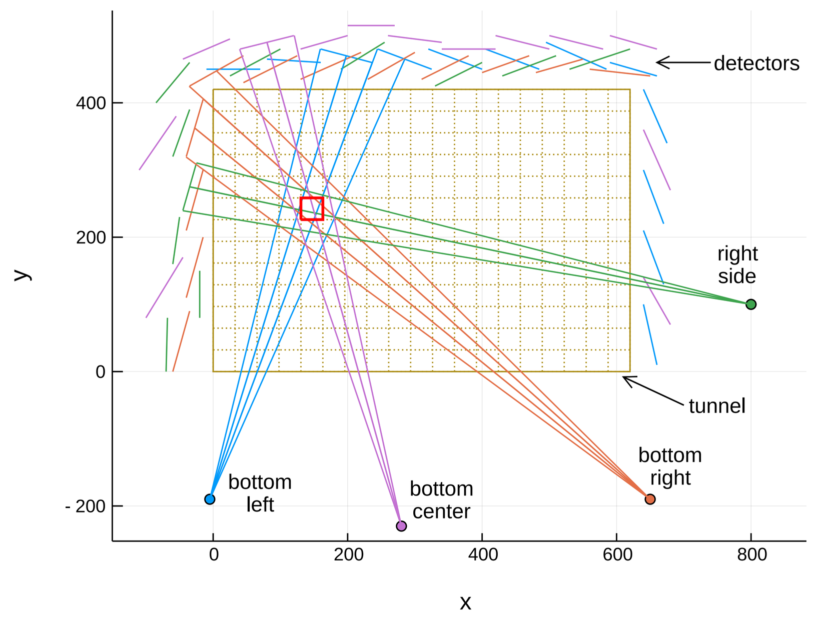
- **Multi-view end-to-end trainable** object detection pipeline
- **Hybrid 2D-3D** architecture
- Based on **Faster R-CNN** [1]
- **ResNet-50** backbone [2]
- Feature extraction of each view independently in 2D
- Combine 2D features in **multi-view pooling layer**
- Propose and evaluate **3D bounding boxes**
- Computationally faster and cheaper than separate processing of views

Multi-view
end-to-end trainable



Multi-view Pooling Layer

- Maps **2D feature maps** of views to common **3D feature volume**
- Uses **known geometry** of recording setup
- Weighted **average** or weighted **maximum** across all X-ray beams
- Normalized volume of intersection as weights



Multi-view pooling
converts features
from 2D to 3D

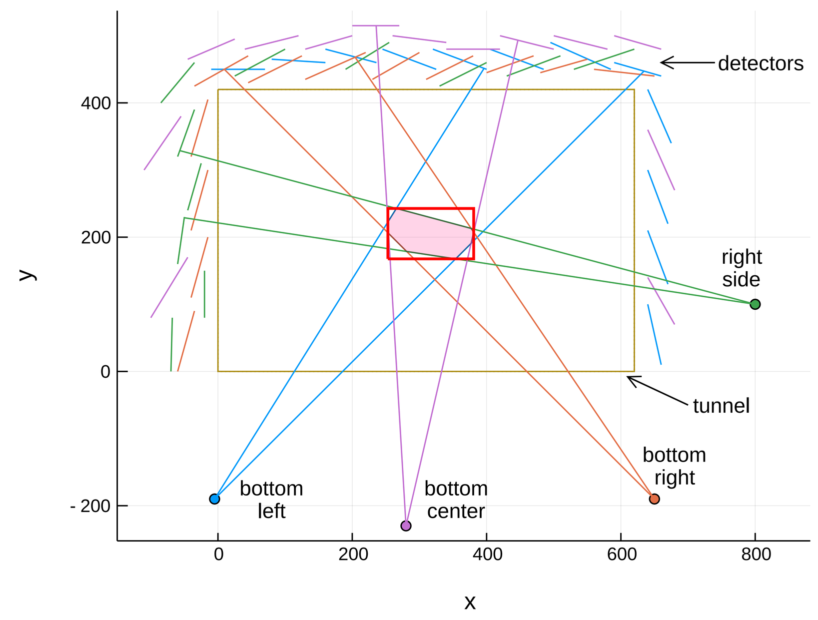
Multi-view X-ray Dataset

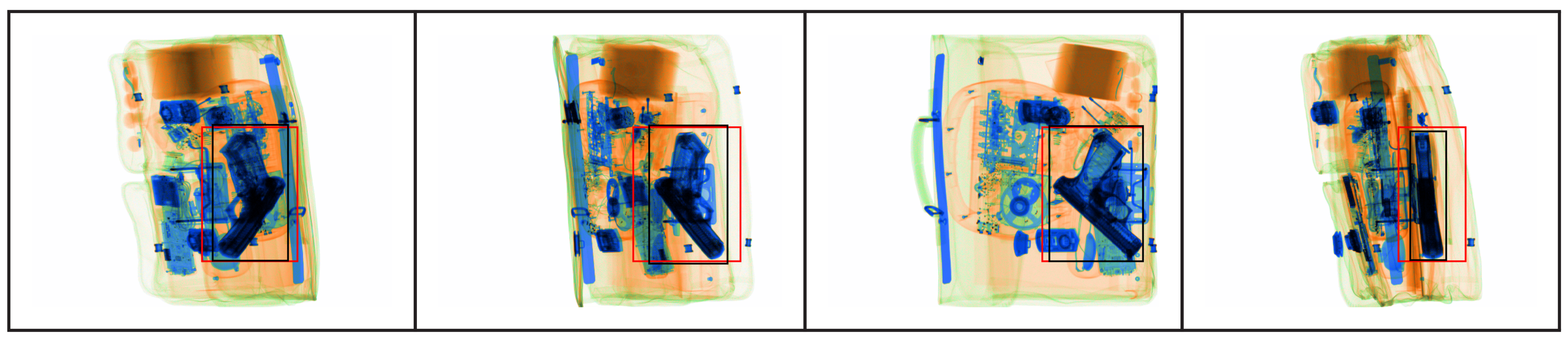
- **Dual-energy X-ray** recordings
- Converted to **false-color RGB**
- **4 different views** per recording
- **2 object classes:** ***weapon*** and ***glassbottle***
- Image resolution of **[704, 832]** × **[101, 1400] px**

Type	Images
Glassbottle	2428
TIP Weapon	8640
Real Weapon	1856
Negative	3800
All	**16724**

3D bounding box annotations

- Generate **axis-aligned 3D bounding boxes** from 2D bounding box annotations
- Intersection of projection lines in the *xy*-plane
- Choose minimal bounding box enclosing the polygon
- Project 3D bounding boxes back onto 2D views for propagation of estimation error



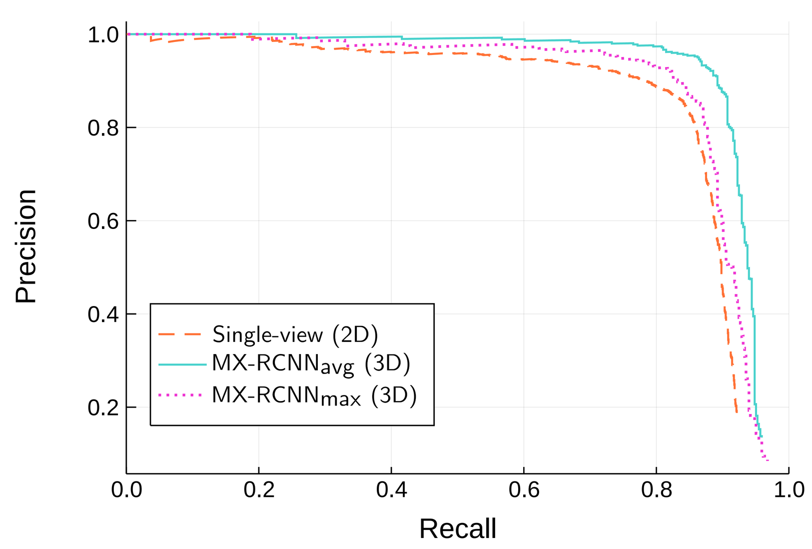


Bounding boxes show the original 2D annotations (*black*) and the reprojected 3D annotations (*red*).

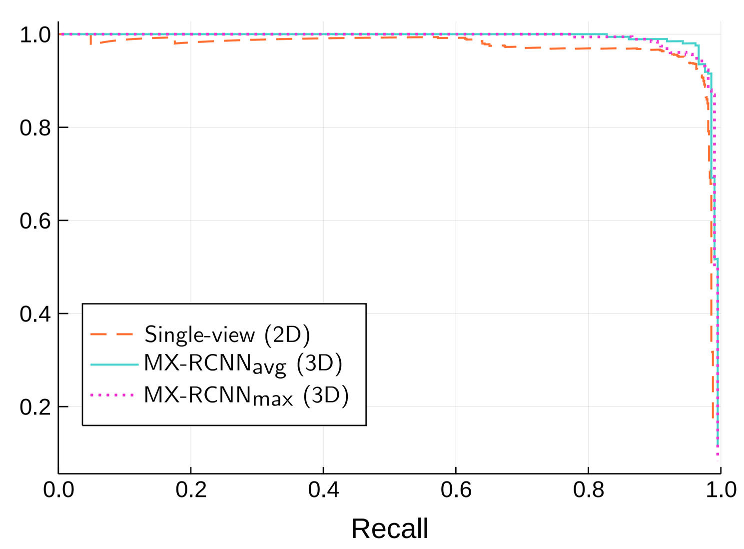
Results

Results (in %) of MX-RCNN networks compared to single-view baseline (standard Faster R-CNN). Evaluation of proposed 3D bounding boxes as well as projections onto 2D views.

Method	Single-view	MX-RCNN_{avg}		MX-RCNN_{max}	
Evaluation	2D	3D	2D	3D	2D
Weapon AP	85.6	**92.3**	90.3	89.0	87.7
Glassbottle AP	96.9	**98.8**	95.4	98.7	95.6
Mean AP	91.2	**95.6**	92.8	93.9	91.7



(a) *Weapon* class.



(b) *Glassbottle* class.

References

[1] Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE T. Pattern Anal. Mach. Intell. **39**(6), 1137–1149 (Jun 2017)

[2] He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016, LNCS, vol. 9908, pp. 630–645. Springer, Cham (2016)

Conclusion

- Multi-view end-to-end trainable MX-RCNN detector
- Novel multi-view pooling layer
- Clear accuracy gains, particularly in the high-recall regime

Acknowledgements

The authors gratefully acknowledge support by Smiths Heimann GmbH.