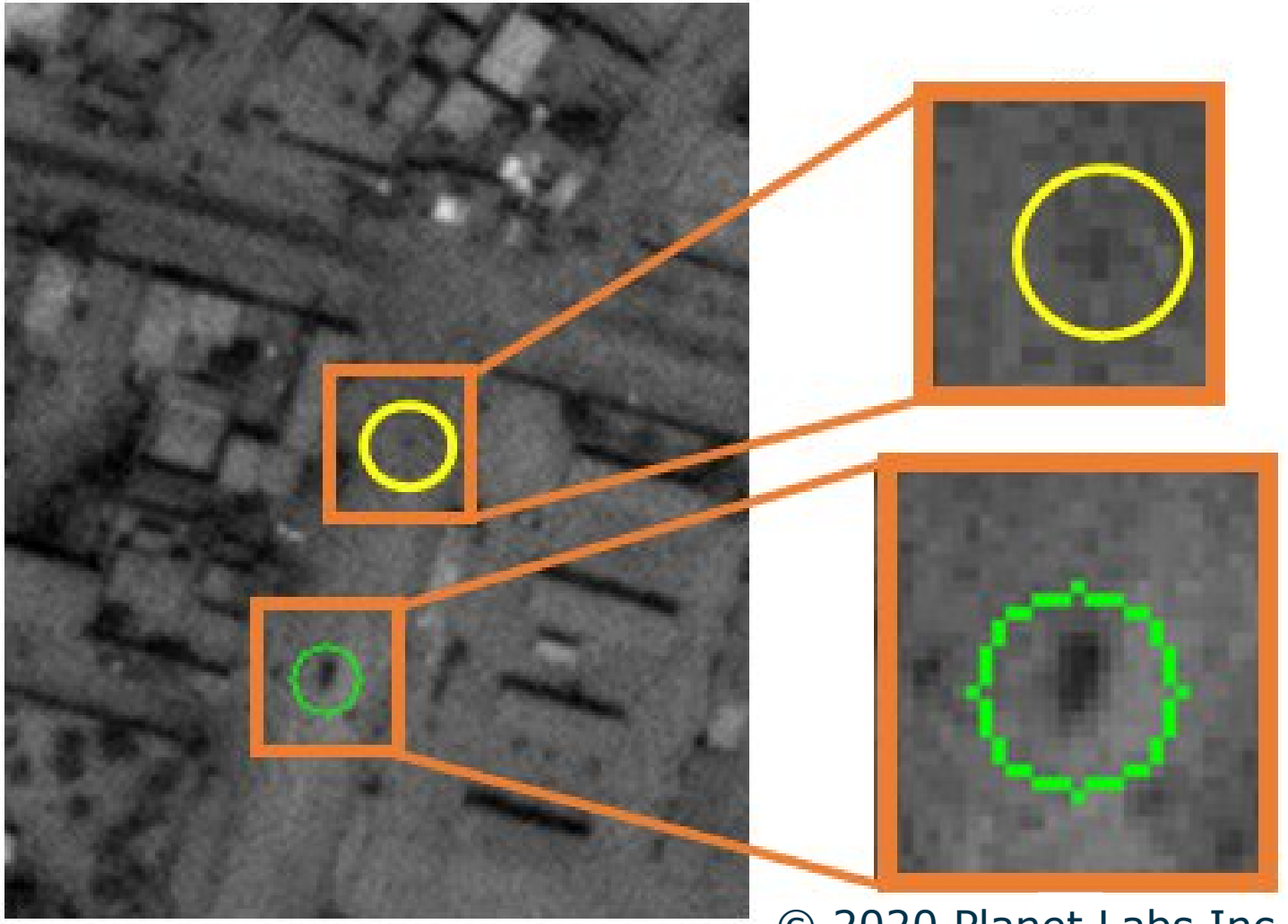
# CNN-based Detection of Tiny Objects in Remote Areas Using Spatiotemporal Earth Observation Data

Roman Pflugfelder (1), Axel Weissenfeld (2)

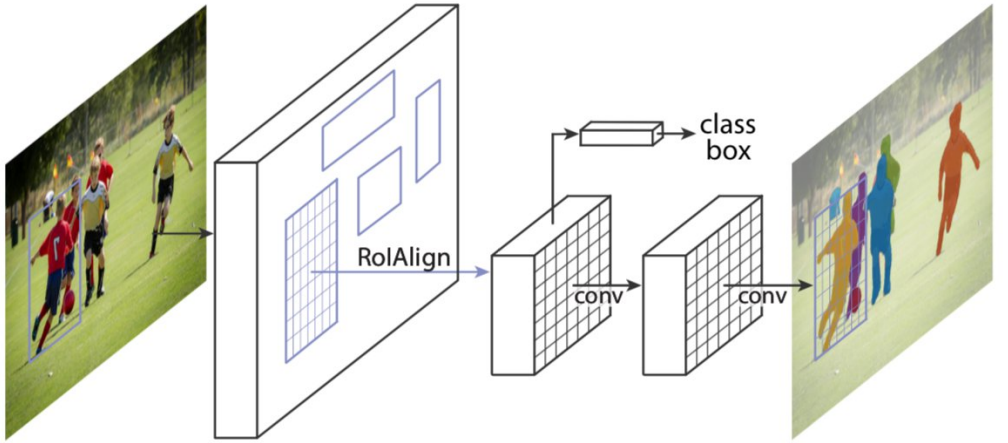
*(1) TU Wien, (2) AIT Austrian Institute of Technology*

# Task challenge: Moving vehicle detection in remote sensing



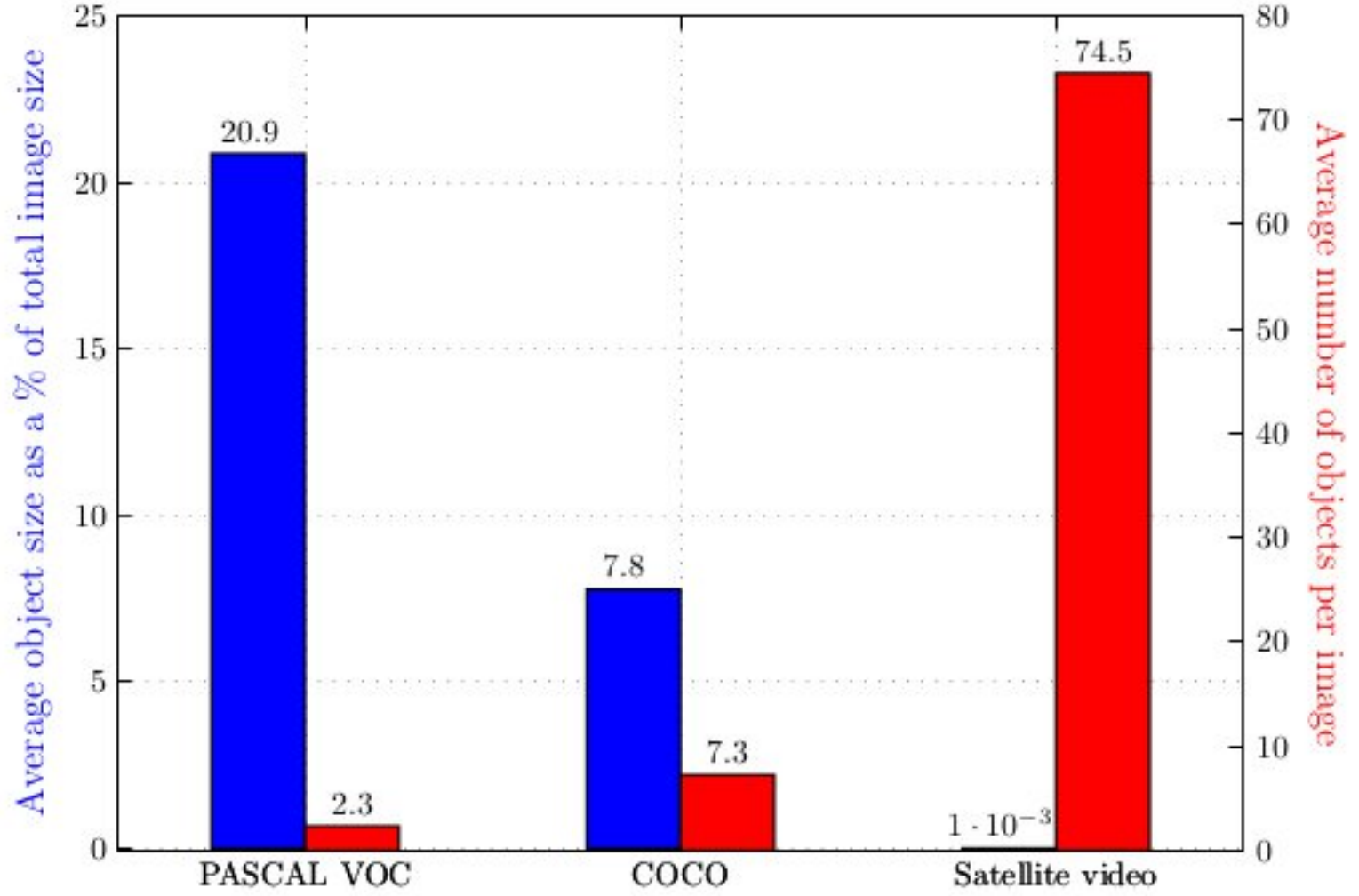
# Prominent approaches for object detection

- Faster R-CNN
  (S. Ren et al., NeurIPS'15)
- Mask R-CNN (see illustration)
  (K. He, ICCV'17)
- SSD - Single Shot Detector
  (W. Liu et al., ECCV'16)
- Yolo V.3
  (J. Redmon, A. Farhadi, CoRR'18)




- Vehicle detection in remote sensing imagery is challenging as vehicles appear **tiny** (four to ten pixels) in **very large images**
- Standard object detections fail…

# Challenge: Typical object detection datasets



# Spatiotemporal Earth Observation Data



multispectral imaging
GSD ~ 0,30m

asynchronous stereo images (WorldView-3)

https://www.satimagingcorp.com/satellite-sensors/worldview-3/

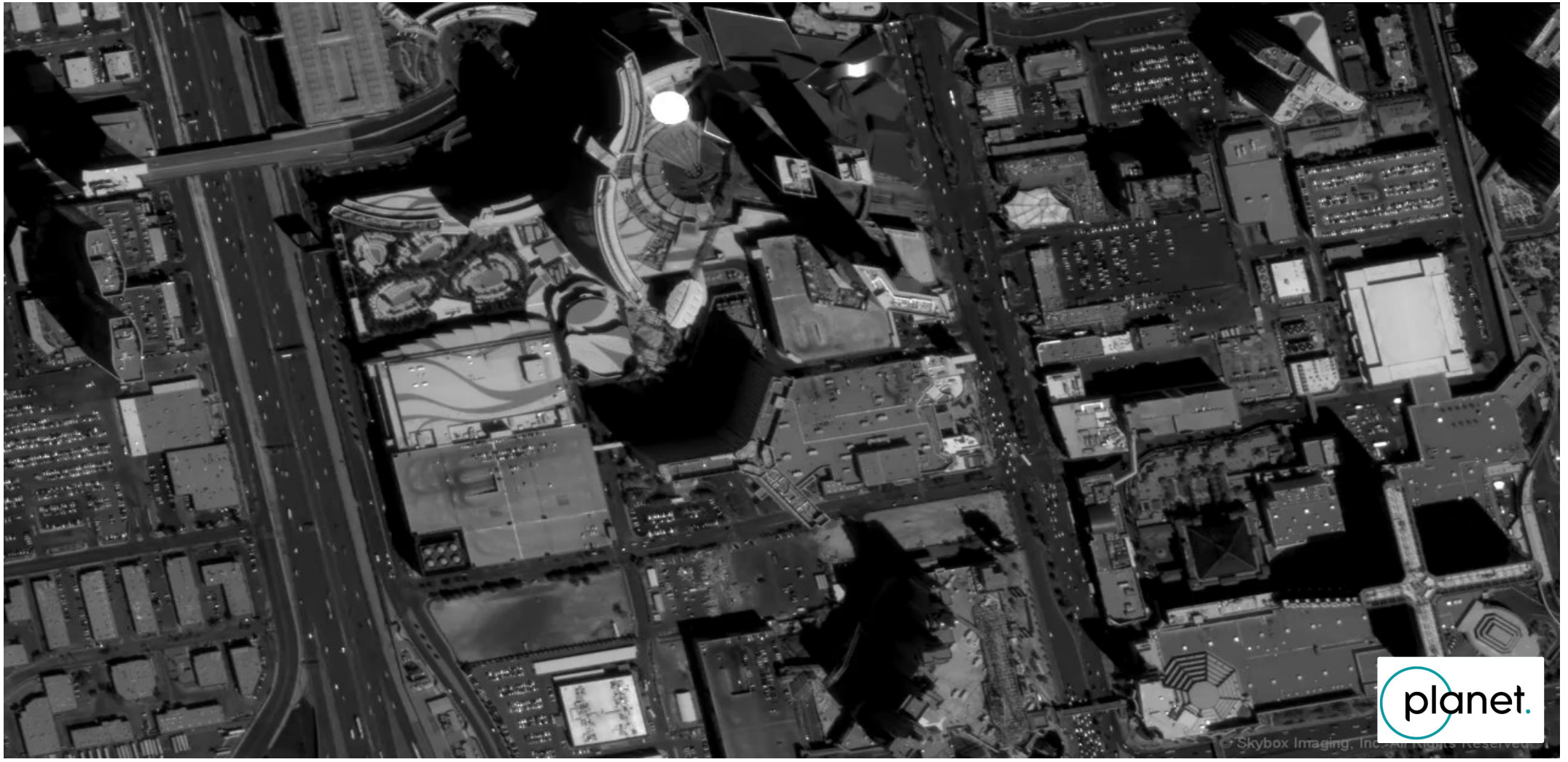
all-frames product

https://assets.planet.com/docs/Planet_Basic_L1A_All-Frames_User_Guide.pdf

satellite video
GSD ~ 0,70m

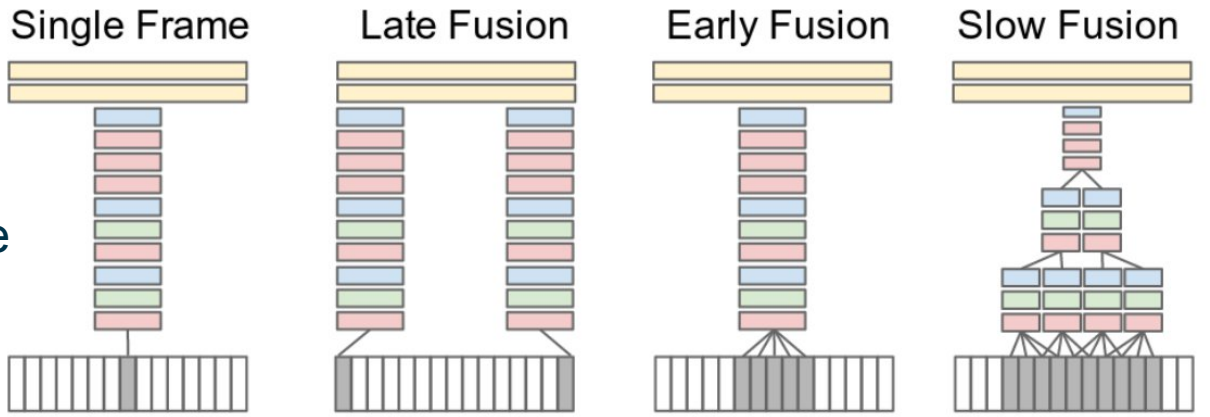
Increasing tempoal sammpling

# Las Vegas Video - courtesy of Planet Inc. (100cm GSD)



# Recurrent Networks

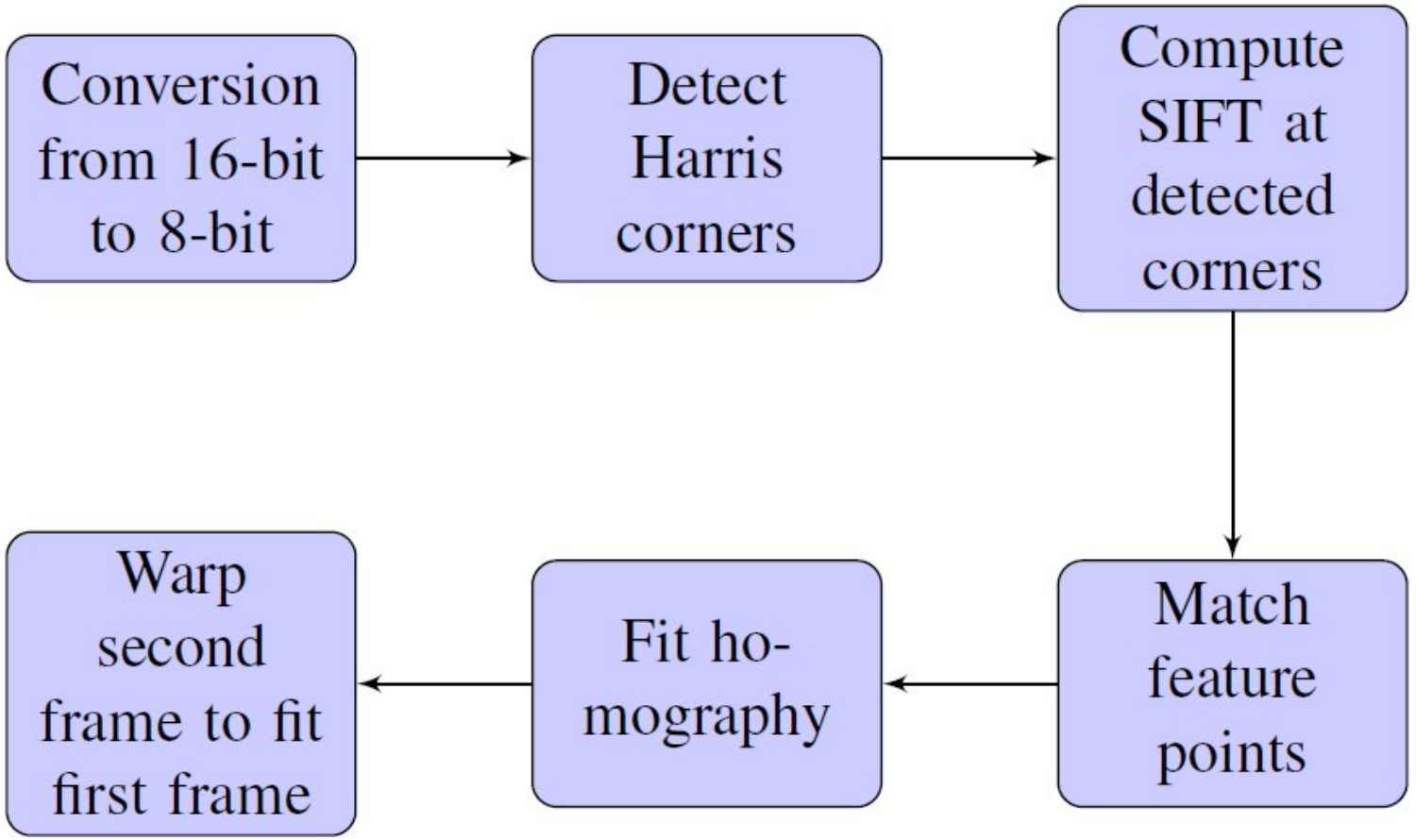
- We propose a fully **spatiotemporal convolutional neural network** (CNN) to detect moving vehicles in satellite videos.
- Inspired by the work of Lalonde et al. [1] – vehicle detection with aerial WAMI data
- Input to spatio-temporal CNNs are a sequence of images. In this way, these CNNs can take advantage of local spatio-temporal information [2]
- Our approach consists of a pre-processing step and the object detection.



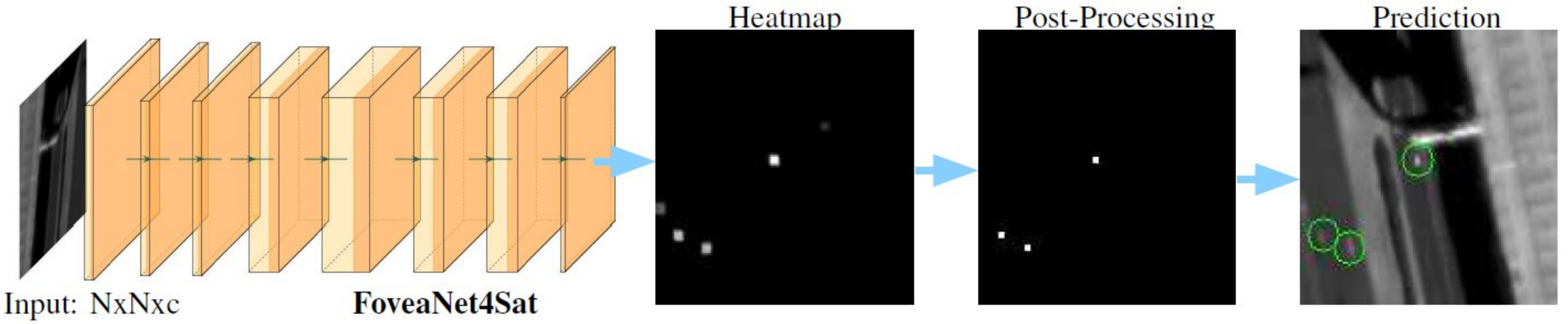

[1] LaLonde, Rodney, Dong Zhang, and Mubarak Shah. "Clusternet: Detecting small objects in large scenes by exploiting spatio-temporal information." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

[2] Karpathy, Andrej, et al. "Large-scale video classification with convolutional neural networks." Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2014.

# Pre-processing: Frame registration pipeline



# Prediction




The object detection process consists of two steps:

- The FoveaNet4Sat predicts a heatmap, which indicates the likelihood that an object is at a given image coordinate.

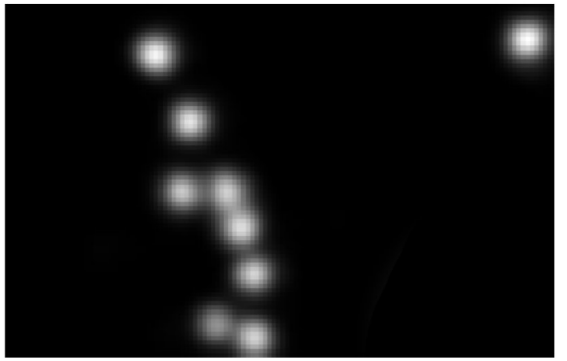
- A post-processing steps derives the object locations by detecting all responses whose value is greater or equal to its 8-connected neighbors followed by a non-maximum suppression.

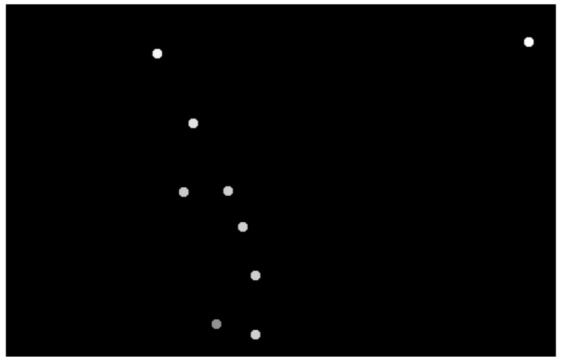
# Network Training

- Output: Heatmap $H$

- Pixel Position $x, y$ in $H$
- Variance $\sigma$ of Gaussian

$$H(x,y) = \sum_{n=1}^{N} \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}$$





- Loss Function- Euclidian distance in the heatmap for training and Adam as optimizer

- Transfer Learning
  - trained on down sampled aerial WAMI (WPAFB'09 dataset), which contains over 160,000 annotated moving vehicles

  - fine-tuned on satellite video

# Experimental Setup

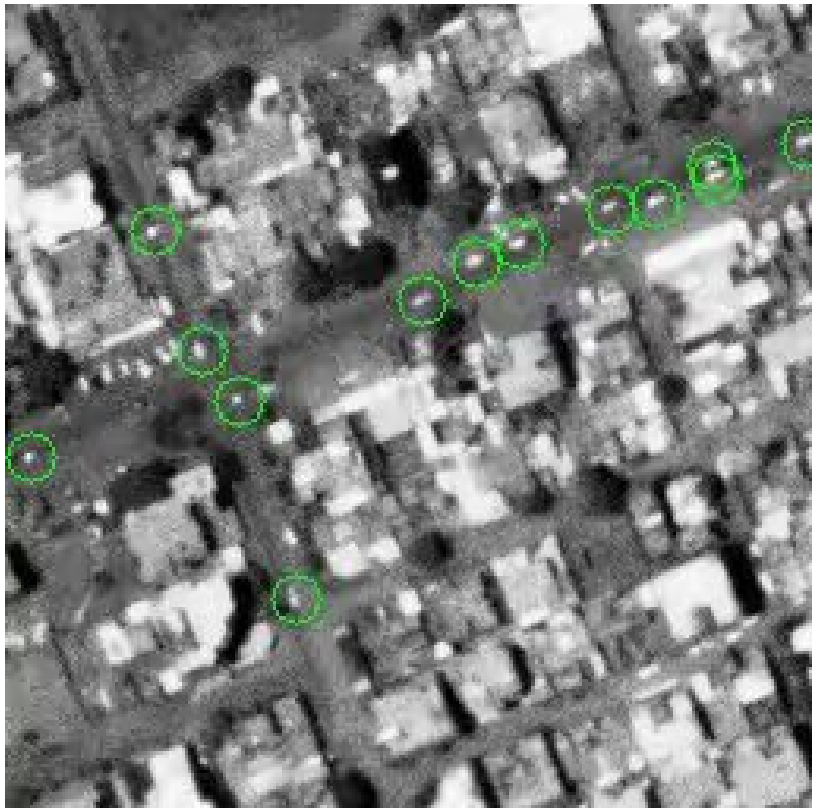
- Choose AOIs used for training (green) and evaluation (blue)
- Split data - two AOIs for training, one for validation
- Precision, recall, F1 measures

# Results for Tiny Vehicles

- Foveanet on Khartoum Video (Planet Inc.)
- 12,087 Training Samples, 3,436 Test Samples

# A closer look

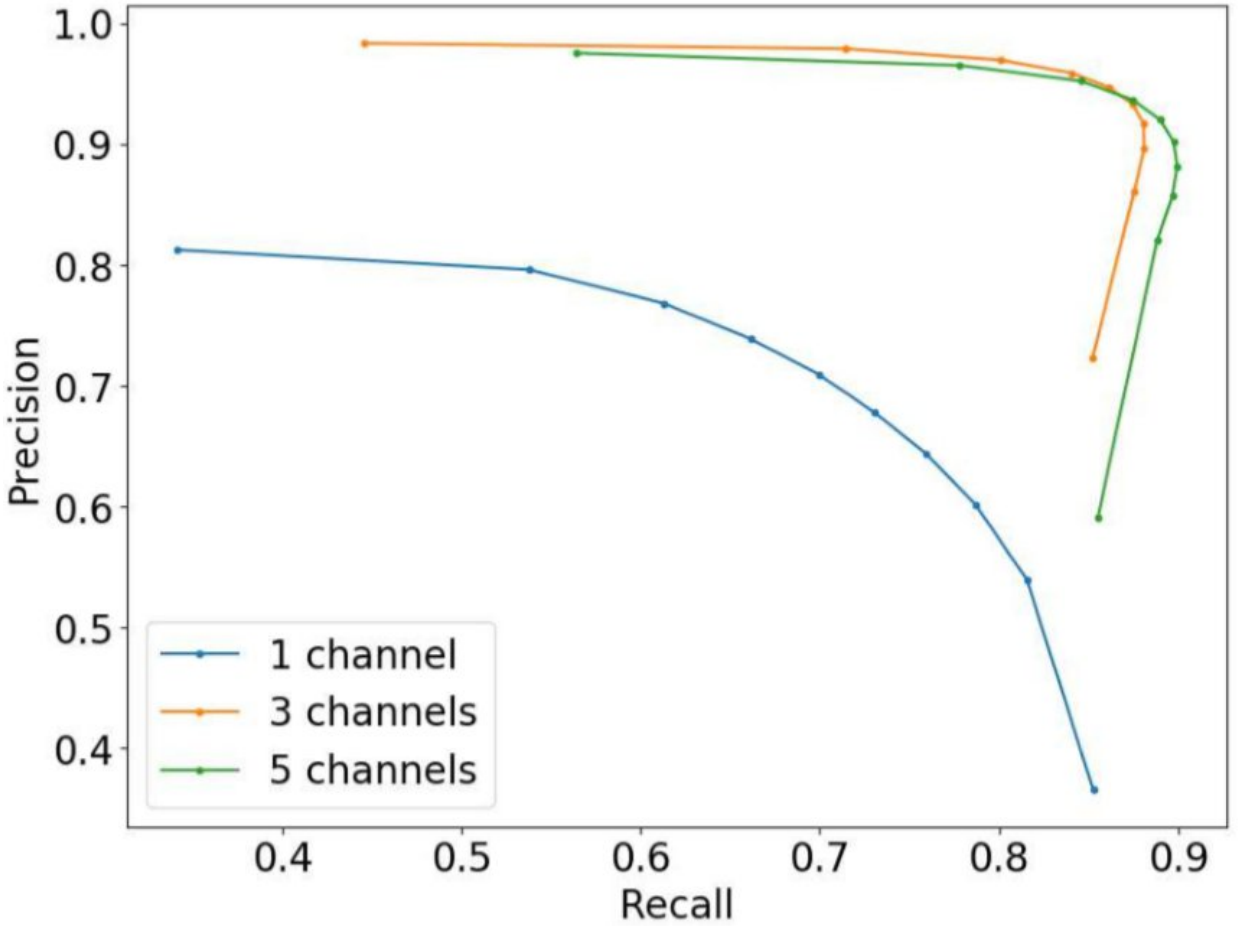
- Colour Coding (Green = True Positive; Blue = False Positive; Red = False Negative)
- Khartoum



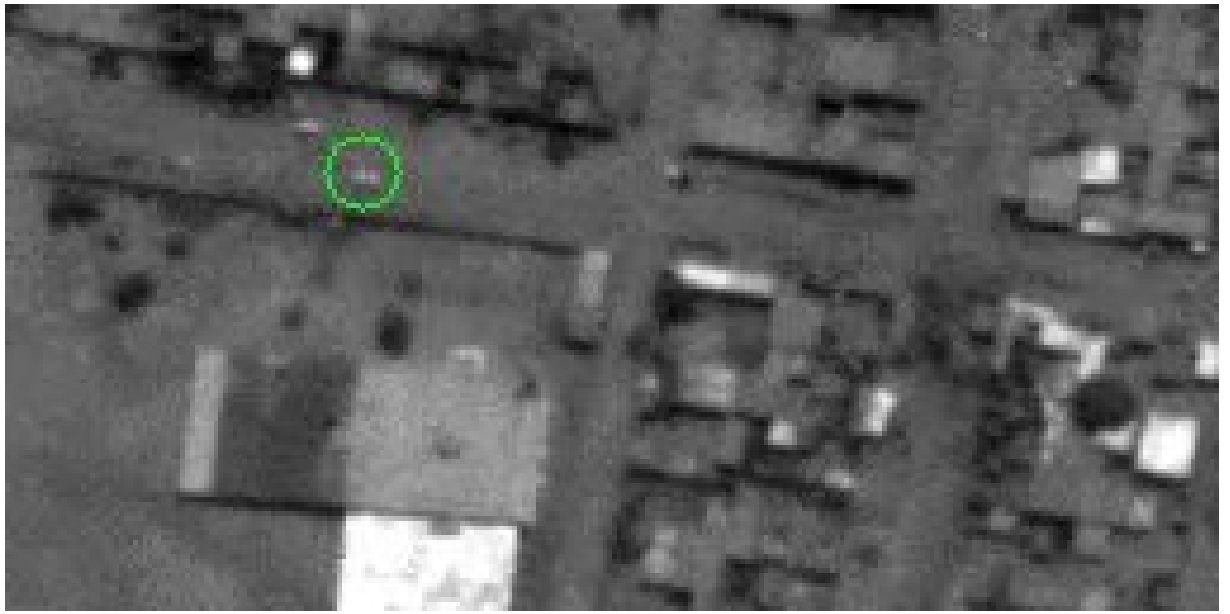
| Measures | Khartoum | Agadez |
|---|---|---|
| Precision | 0,92 | 0,96 |
| Recall | 0,81 | 0,70 |
| F1 | 0,86 | 0,81 |
| Training Exempels | 12,087 | 9,144 |
| Test Exempels | 3,436 (1/4) | 2,291 (1/4) |

# Importance of motion

- Recall-Precision curve of Foveanet on Khartoum Video




Agadez



# Las Vegas Video

- Quantitative study and comparison to other methods

| Methods | Prec. | Recall | F1 |
|---|---|---|---|
| ViBe | 0.58 | 0.17 | 0.26 |
| GMMv2 | 0.65 | 0.27 | 0.38 |
| GMM | 0.46 | 0.5 | 0.48 |
| F-RCNN-LPR | 0.58 | 0.44 | 0.5 |
| GoDec | 0.95 | 0.36 | 0.52 |
| RPCA-PCP | 0.94 | 0.41 | 0.57 |
| Decolor | 0.77 | 0.59 | 0.67 |
| LSD | 0.87 | 0.71 | 0.78 |
| E-LSD | 0.85 | 0.79 | 0.82 |
| **FoveaNet** | **0.86** | **0.82** | **0.84** |
| **FoveaNet4Sat** | **0.85** | **0.92** | **0.88** |

Ground truth provided by Junpeng Zhang et al. "Error bounded foreground and background modeling for moving object detection in satellite videos". In: TGARS. 58.4. 2019.

# Conclusions

- Spatiotemporal convolutional neural network (CNN) very suitable for detecting tiny moving objects and very good results are achieved
- Foveanet4Sat yields 0.88 $F1$ comparable to E-LSD (J. Zhang, X. Jia, J. Hu, TGARS'19) 0.82 $F1$ on Las Vegas
  - $F1$ of 0.86 in Khartoum
  - F1 of 0.81 in Agadez
- Challenge: Lack of public datasets to enhance SOTA

Future work:
Sparsity of vehicle occurrences not considered
  - need of attention/clustering mechanism

What does the spatiotemporal neural network learn?
  - Ideally the network learns slopes of trajectories in spacetime -> further insights needed