FOVEA: Foveated Image Magnification for Autonomous Navigation

Carnegie Mellon University

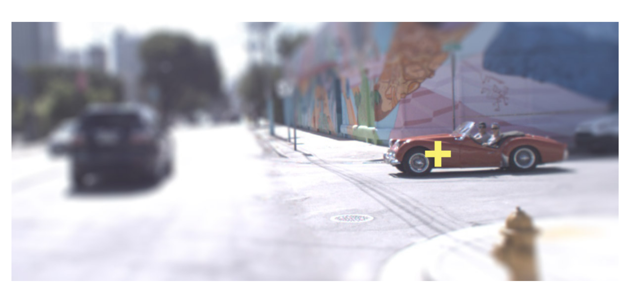
Chittesh Thavamani[1]* Mengtian (Martin) Li[1]* Nicolas Cebron[2] Deva Ramanan[1,2]

[1]Carnegie Mellon University [2]Argo AI

The Problem: Real-Time High-Res Object Detection

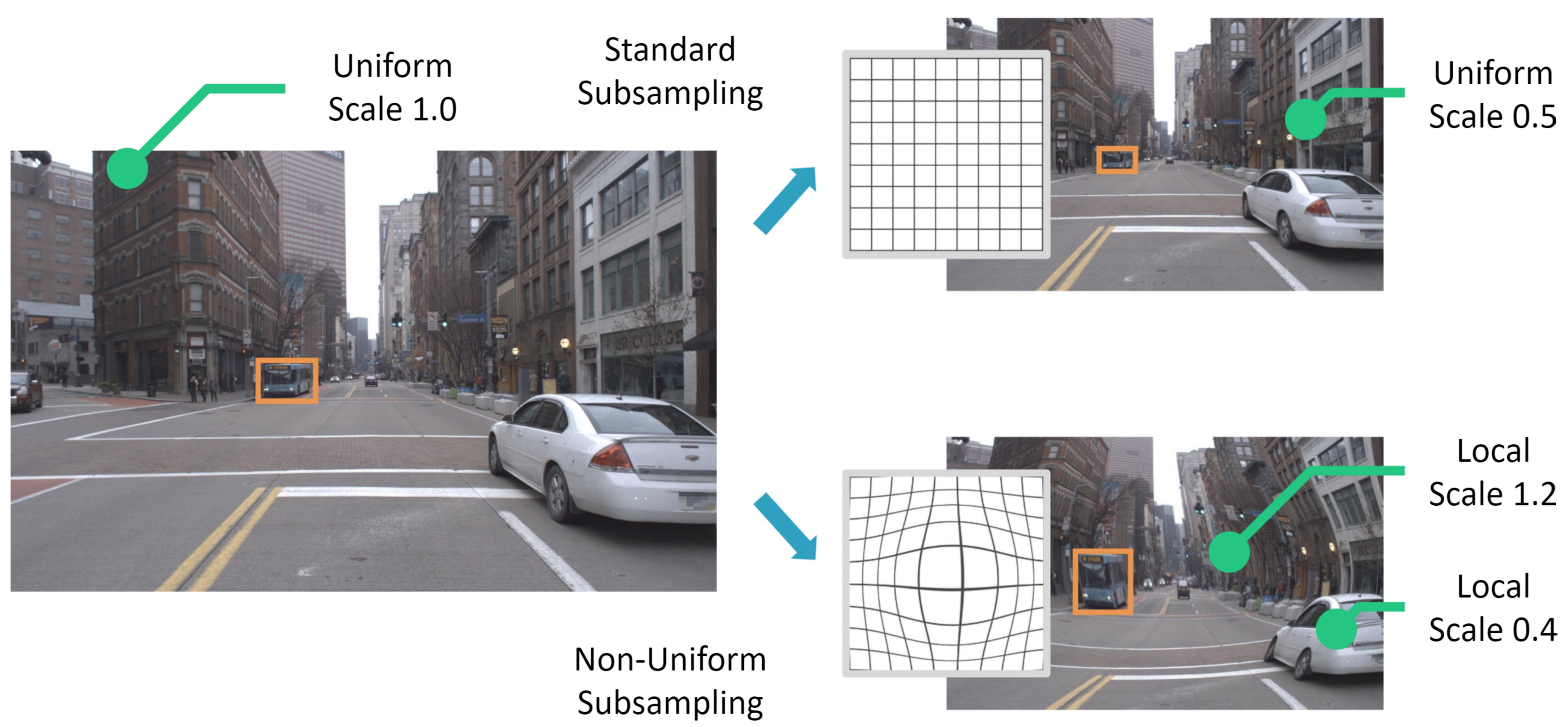
Efficiency becomes unprecedentedly important in the era of *"oversensing"*: multiple high-resolution cameras and LiDAR sensors can be found on a single autonomous vehicle. Such an *overwhelming* amount of data coming at a high-frame rate call for novel approaches to make use of high-res footage beyond the *conventional downsampling and frame dropping* [1].

Intuition: Attentional Foveation

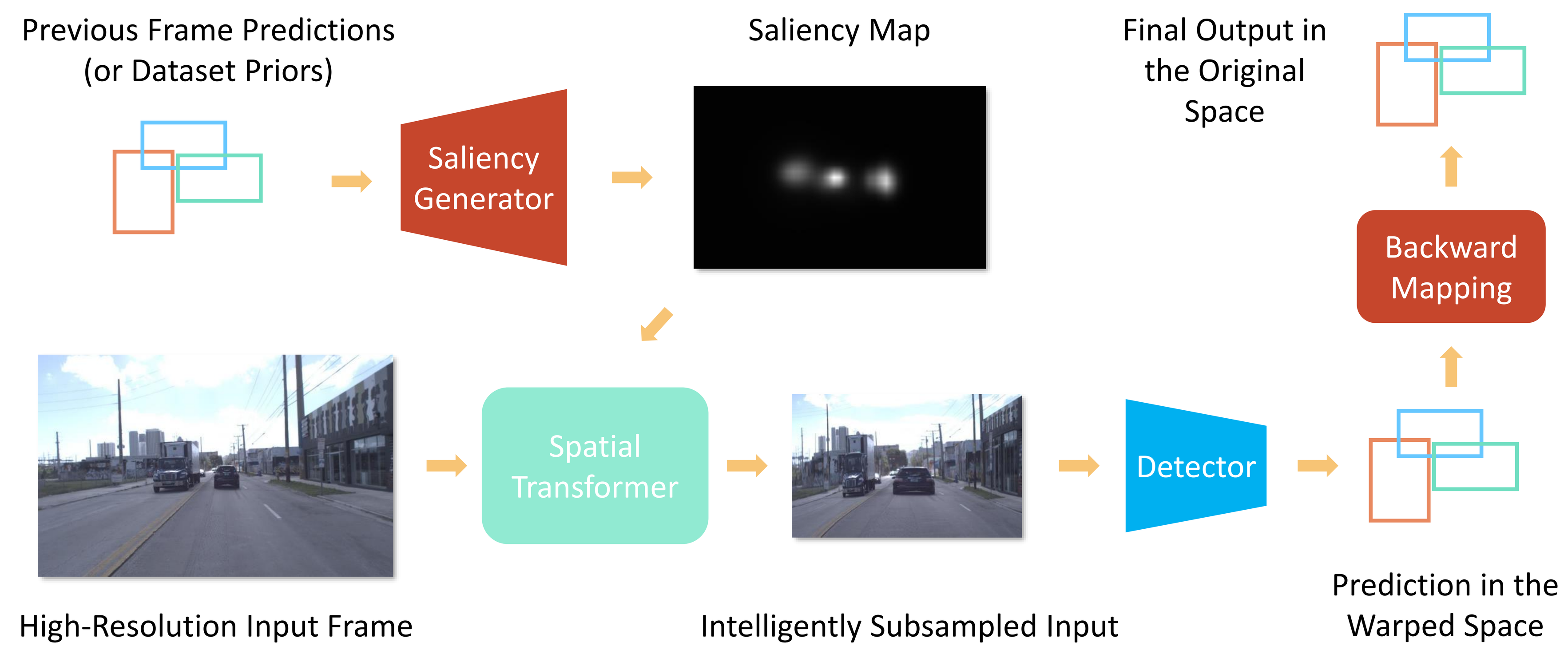


Inspiration: In the human visual system, the center (fovea) has a much higher resolution than the periphery.

Key Insight: Learn to intelligently subsample the input. We adaptively downsample the high-resolution raw image such that the original resolution is better preserved for *salient* areas. For example, we might contract the background and bigger objects to make room for smaller objects while maintaining a small canvas size.



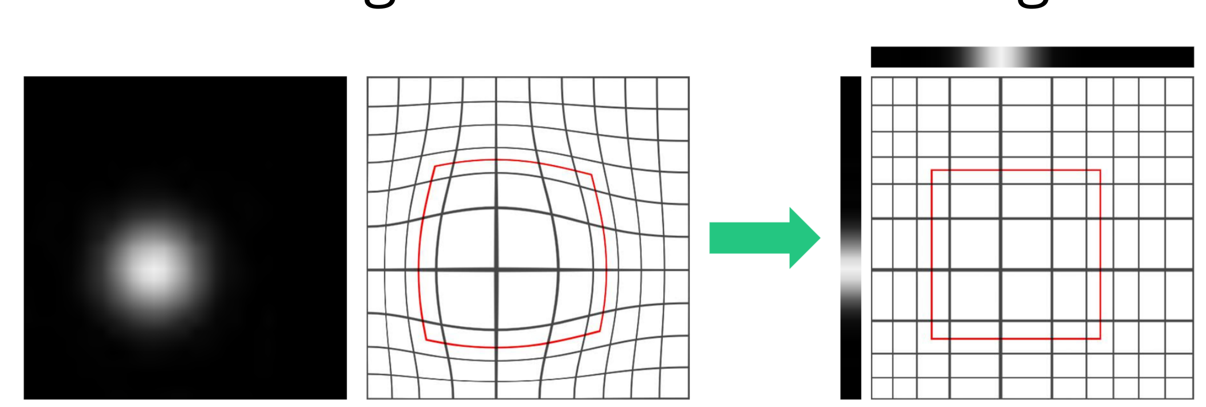
Approach



Note that our method is *largely agnostic to a specific detector* since we modify the input/output of a given detector. All components are *differentiable*, and can be trained end-to-end!

Key Components and Contributions

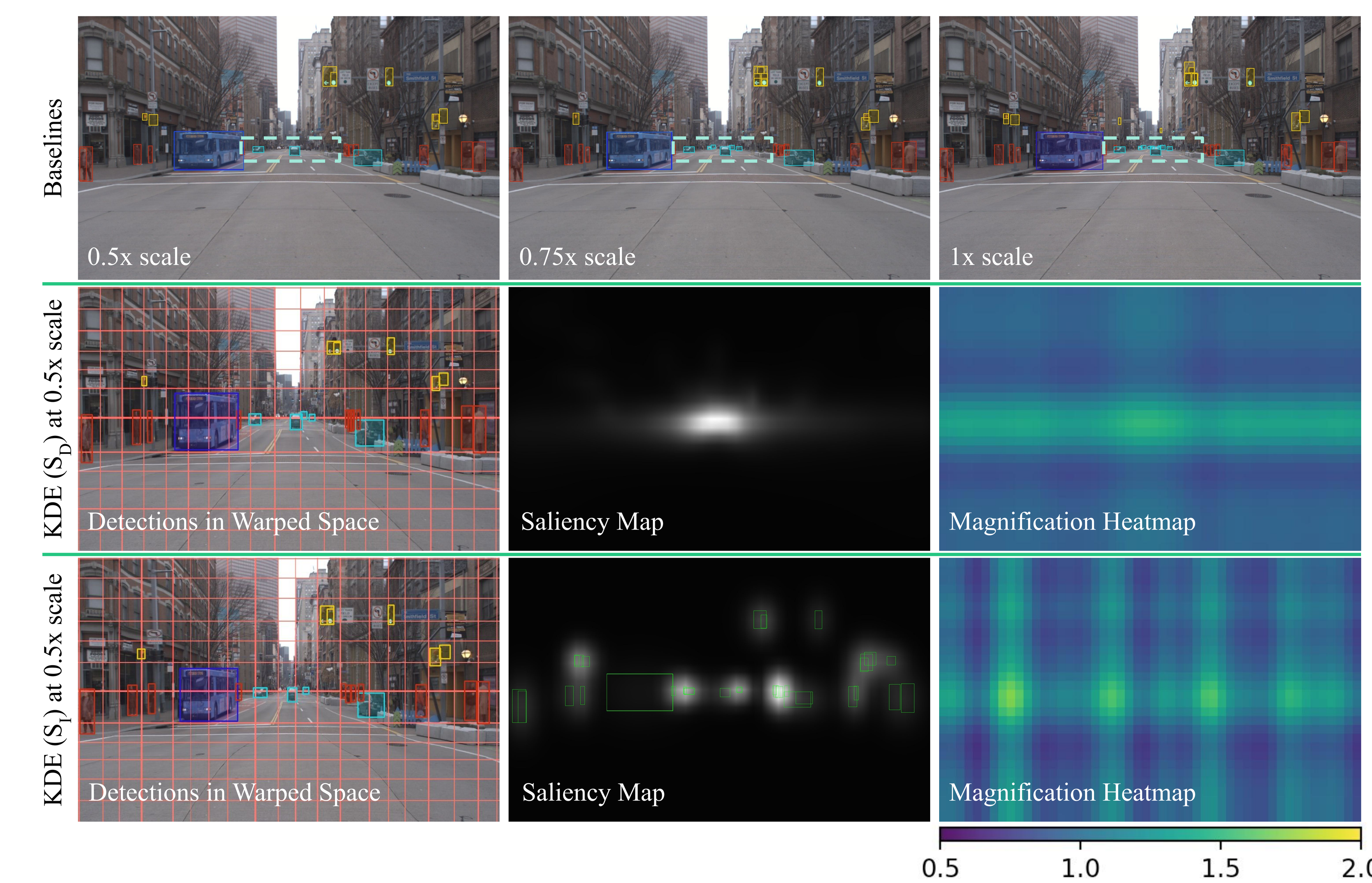
- **Saliency from Temporal and Spatial Object Priors**
  The *saliency generator* maps bounding boxes from previous prediction or over the entire dataset to a soft saliency map.
- **Bounding-Box Backward Mapping**
  Warped image leads to warped boxes and need to be unwarped! It turns out to be an existing transformation.
- **Anti-Cropping Regularization**
  The spatial transformer in [2] tends to crop the input, undesirable for object detection. We reflectively pad the saliency map to prevent cropping.
- **Separable Warps**
  Ensure that bounding boxes remain axis-aligned!



Results

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	Latency (ms)
Baseline (FRCNN FPN)	24.2	38.9	26.1	4.9	29.0	50.9	50.9 ± 0.9
KDE (S_D)	26.7	43.3	27.8	8.2	29.7	54.1	50.8 ± 1.2
KDE (S_I)	28.0	45.5	**29.2**	**10.4**	**31.0**	**54.5**	52.2 ± 0.9
LKDE (S_I)	**28.1**	**45.9**	28.9	10.3	30.9	54.1	50.5 ± 0.8
Upper Bound (0.75x)	29.2	47.6	31.1	11.6	32.1	53.3	87.0 ± 1.4
Upper Bound (1x)	33.3	53.9	35.0	16.8	34.8	53.6	135.0 ± 1.6

Table 1. Offline detection on Argoverse-HD after finetuning.



ID	Method	AP	AP_S	AP_M	AP_L
1	Prior art [1]	17.8	3.2	16.3	33.3
2	1 + Better implementation	19.3	4.1	18.3	34.9
3	1 + Train with pseudo GT	21.2	3.7	**23.9**	43.8
4	2 + KDE (S_I)	19.3	5.2	18.5	39.0
5	3 + KDE (S_I)	**23.0**	**7.0**	23.7	**44.9**

Table 2. Streaming detection [1] on Argoverse-HD.

References

[1] Mengtian Li, Yuxiong Wang, and Deva Ramanan. Towards streaming perception. In *ECCV*, 2020.

[2] Adria Recasens, Petr Kellnhofer, Simon Stent, Wojciech Matusik, and Antonio Torralba. Learning to zoom: a saliency-based sampling layer for neural networks. In *ECCV*, pages 51–66, 2018.