Implicit Occupancy Flow Fields for Perception and Prediction in Self-Driving

Ben Agro*, Quinlan Sykora*, Sergio Casas*, Raquel Urtasun

TUE-AM-131

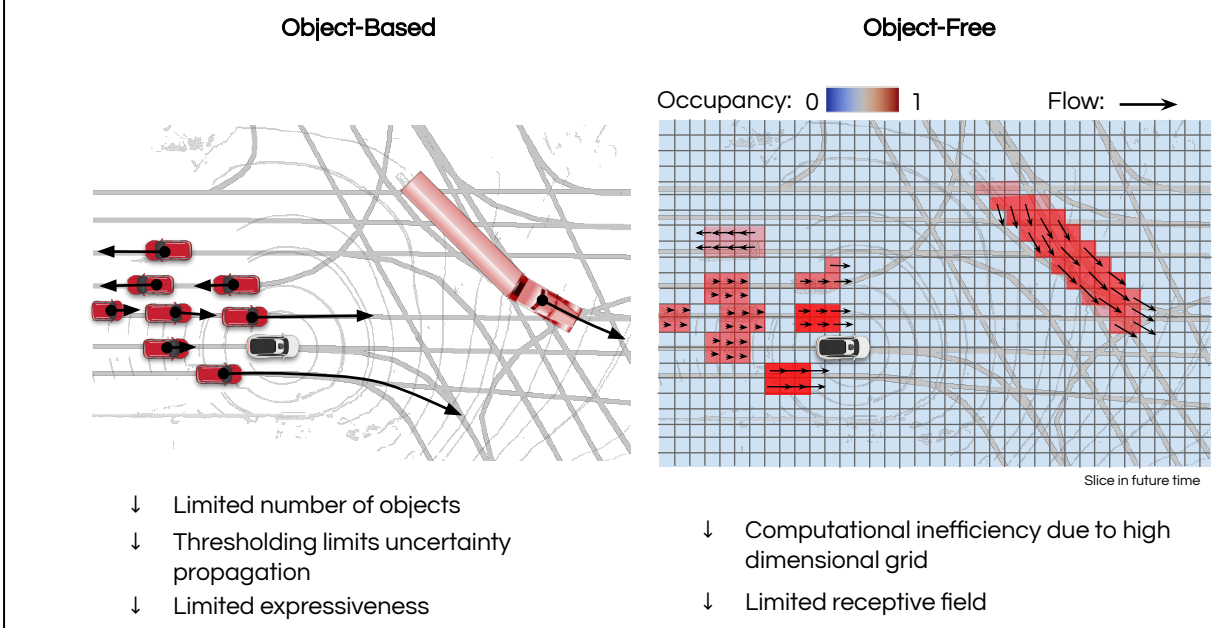
This video contains audio narration, please keep the volume on

5 second static frame displaying the paper title, authors, and session-poster ID

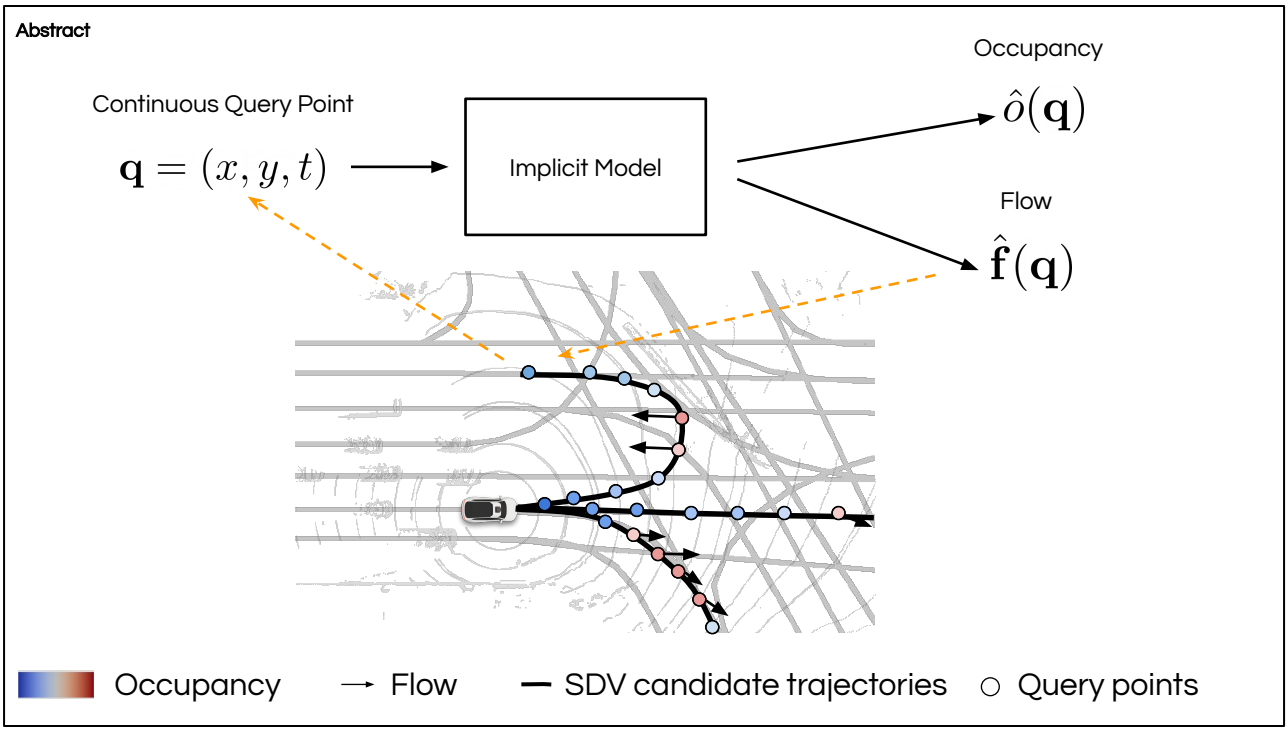
I am Ben Agro and today I'll be presenting Implicit Occupancy Flow Fields for Perception and Prediction in Self-Driving

Begin 55 s Summary

Abstract

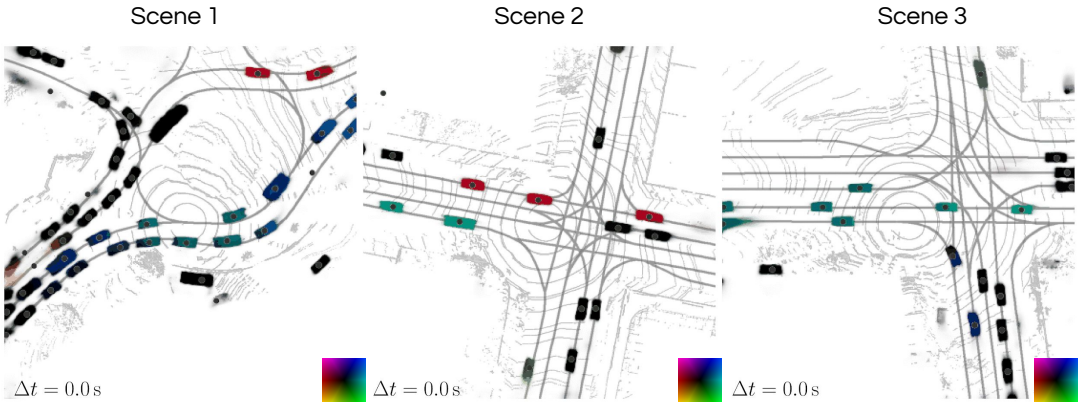


Traditional object-based autonomy detects discrete (click) objects by thresholding confidence scores.
This can be unsafe, because (click) we need to limit the number of detections for efficiency, (click) and thresholding reduces uncertainty propagation.
(click) Alternatively, object-free methods output (click) dense occupancy and (click) flow grids for the whole scene over future time.
These methods are (click) computational inefficient due to high dimensional grids, and (click) inaccurate due to the limited receptive field inherent to fully convolutional networks.



This motivates ImplicitO, a unified perception and prediction approach which employs an efficient global attention mechanism to implicitly represent occupancy and flow over time with a single neural network.
ImplicitO avoids unnecessary computation as it can be (click) directly queried by a motion planner at a continuous point in space and future time for (click) the probability that point is occupied at that time, and a 2D velocity vector at that point if it were occupied.

Abstract

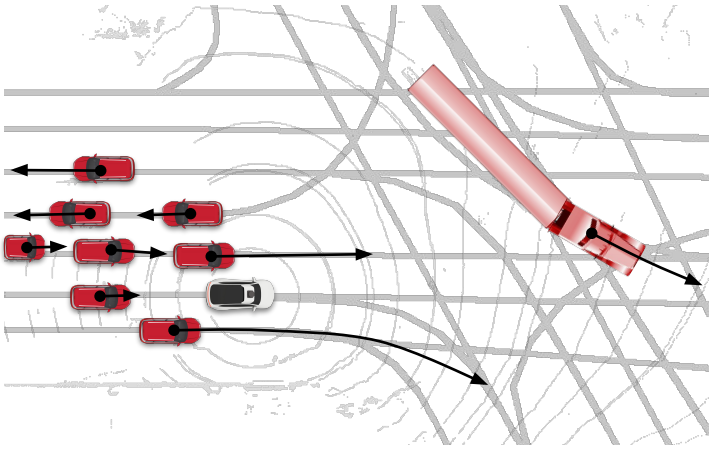


Through extensive experimentation in urban and highway settings, we demonstrate ImplicitO outperforms the current state of the art.

End 55 s Summary

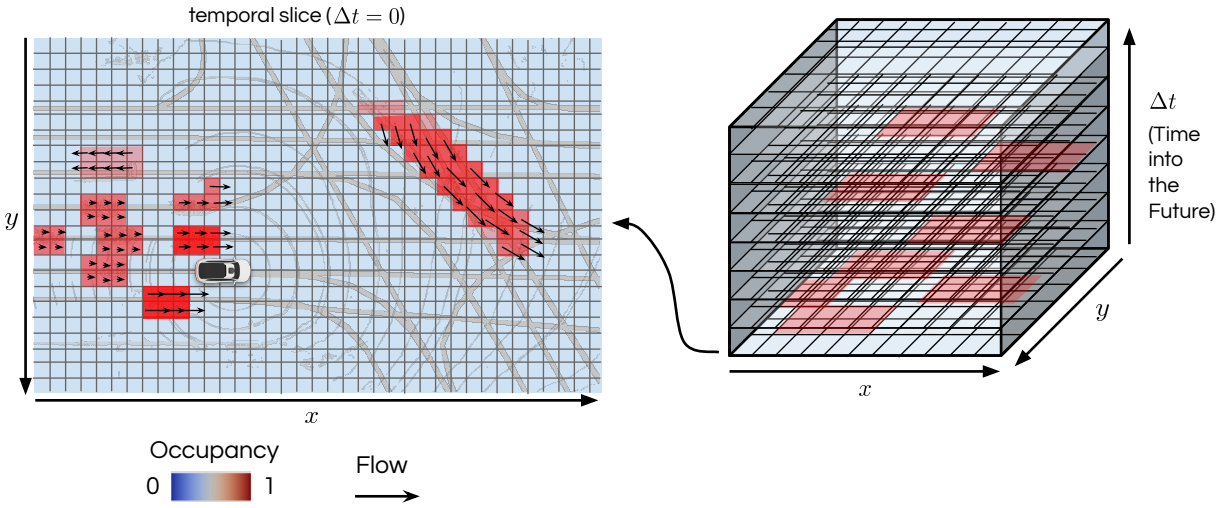
Traditional Object-Based Autonomy Stack

1. Perception
2. Prediction
3. Motion Planning



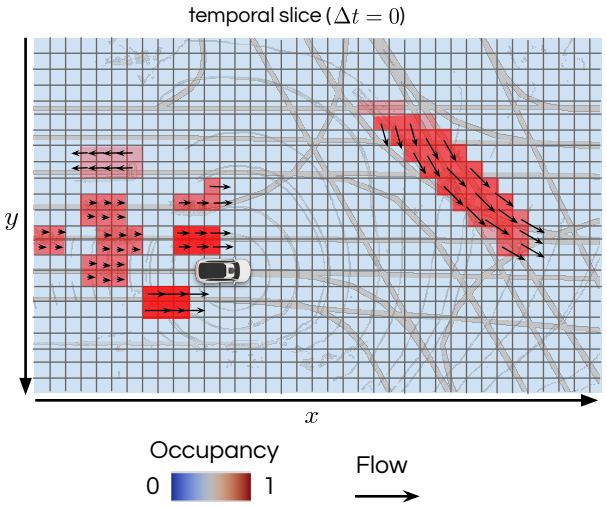
Now, we will present our method in more detail.
The traditional object-based autonomy stack of a self-driving vehicle takes in (click) sensor evidence, (click) and offline data such as a map
(click) to first detect objects in the scene,
(click) and then predict the trajectories of those objects.
(click) Then a motion planner uses those predictions to decide on a plan for the self-driving vehicle.
(click) Our work focuses on improving perception and prediction.

Object-Free Perception and Prediction



Alternatively, object-free perception and prediction discretize space (click) and future time (click) to create a spatio-temporal grid.
(click) Visualizing a single temporal slice from this grid, at each grid cell occupancy (click) and motion (click) are predicted.

Object-Free Perception and Prediction

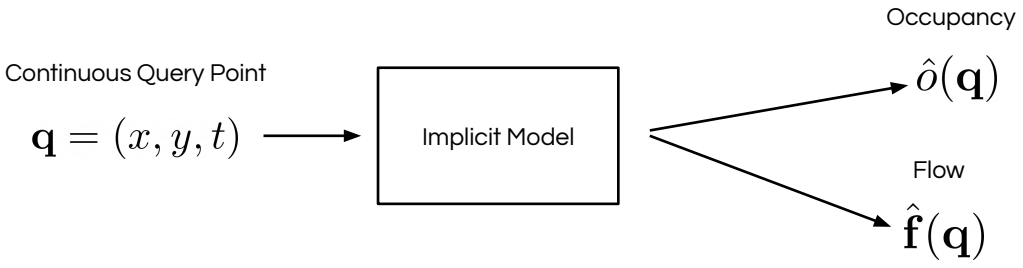


+ No thresholding
+ Uncertainty propagation
+ Expressive future trajectories
– Computationally expensive
– Wasted computation

This addresses the shortcomings of object-based perception and prediction:
(click) No detection confidence thresholding is required,
(click) uncertainty is propagated between perception and prediction as they are now framed as a single task,
(click) the distribution over possible future behaviors is more expressive.

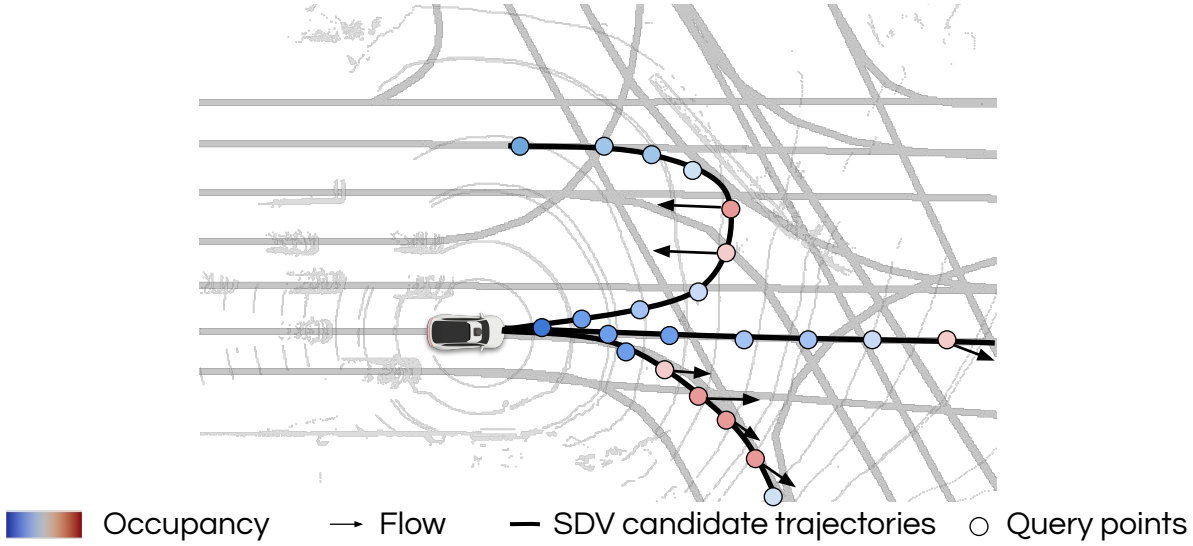
However, (click) this approach is computationally expensive because the spatio-temporal grid must be high-dimensional to mitigate discretization errors,
(click) and much of the computation is performed in regions irrelevant to motion planner.

Our Approach: **ImplicitO**



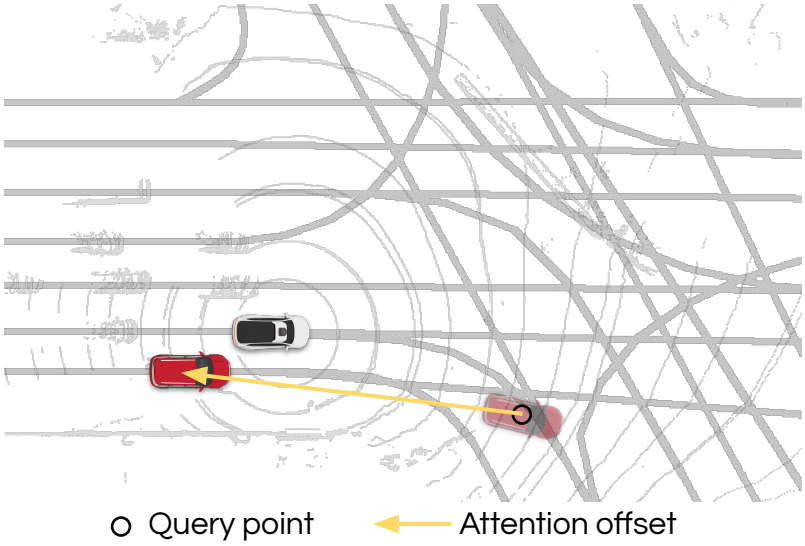
This motivates our approach: **ImplicitO,** which utilizes an implicit representation to predict occupancy probability and a 2-dimensional flow vector in bird's eye view at any continuous spatio-temporal query point, (x, y, t).

Our Approach: **ImplicitO**



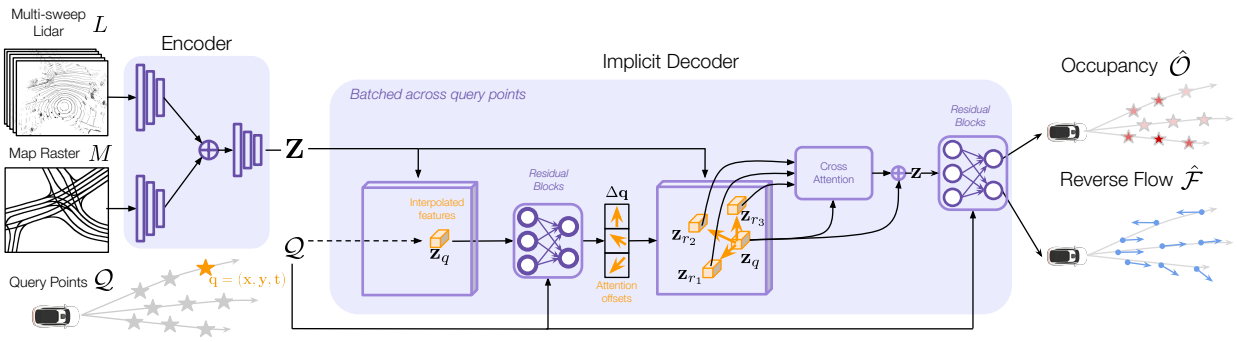
For instance, (click) given sensor and map data, (click) ImplicitO can be queried by the motion planner around candidate trajectories,
improving efficiency by computing occupancy and flow only where it matters downstream.

ImplicitO Method



We design ImplicitO with an efficient global attention mechanism that increases its effective receptive field for accurate occupancy prediction.
(click) Consider a car travelling at 10 m/s observed by the SDV. Two seconds into the future, (click) this car will likely be 20 meters from where it was observed.
This motivates our architecture, which uses local features at the query point to predict where to (click) “look next” in the scene for information relevant to occupancy and flow prediction,
for example, back to where the LiDAR evidence is.

ImplicitO architecture



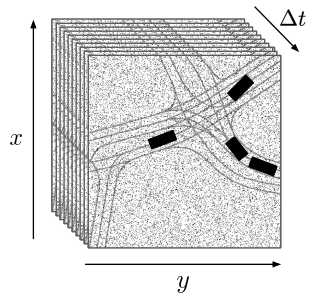
(click) ImplicitO takes as input a rasterization of the map, and voxelized birds-eye view of past LiDAR frames.
(click) These are processed by a convolutional encoder, to produce a feature map.
(click) The decoder also takes as input a set of query points.
(click) Considering a single query,
(click) the global attention mechanism predicts K offsets from this point to locations in the feature map with relevant information.
(click) This information is used to predict occupancy and flow,
(click) and runs in parallel on all query points.

ImplicitO Training

$$\mathcal{L} = \mathcal{L}_o + \lambda_{\mathbf{f}} \mathcal{L}_{\mathbf{f}}$$

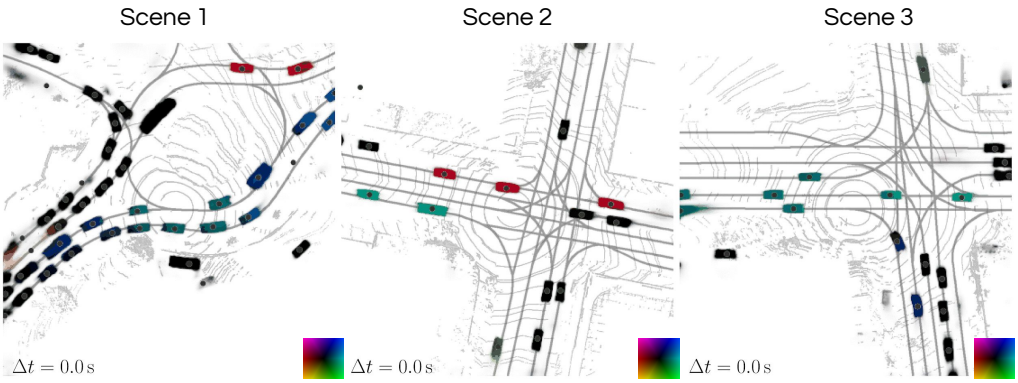
$$\mathcal{L}_o = \frac{1}{|\mathcal{Q}|} \sum_{\mathbf{q} \in \mathcal{Q}} \mathcal{H}(o(\mathbf{q}), \hat{o}(\mathbf{q}))$$

$$\mathcal{L}_f = \frac{1}{|\mathcal{Q}|} \sum_{\mathbf{q} \in \mathcal{Q}} o(\mathbf{q}) ||\mathbf{f}(\mathbf{q}) - \hat{\mathbf{f}}(\mathbf{q})||_2$$



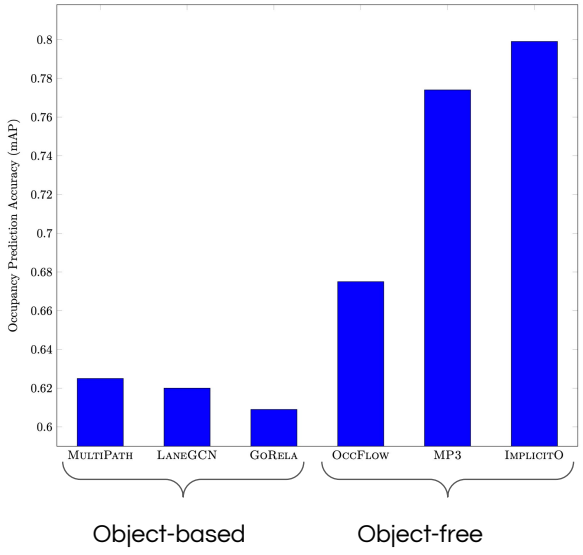
We train our implicit network to minimize (click) a linear combination of an occupancy loss and a flow loss.
Occupancy is supervised with a (click) binary cross entropy between the predicted and ground struth occupancy at each query point.
(click) We supervise the flow with an L2 loss only for query points that are occupied, and in doing so, the model learns to predict the motion of a query point should it be occupied.
(click) During training, the continuous query points are sampled (click) uniformly randomly from the spatio-temporal volume.

ImplicitO Predictions



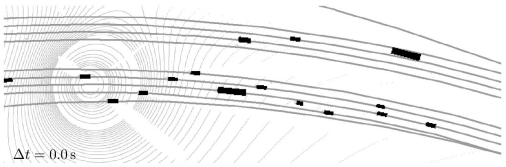
Here, we show predictions of ImplicitO when we query on a spatio-temporal grid. The hue and value represent the direction and magnitude of flow, respectively, while the alpha channel represents the occupancy probability.

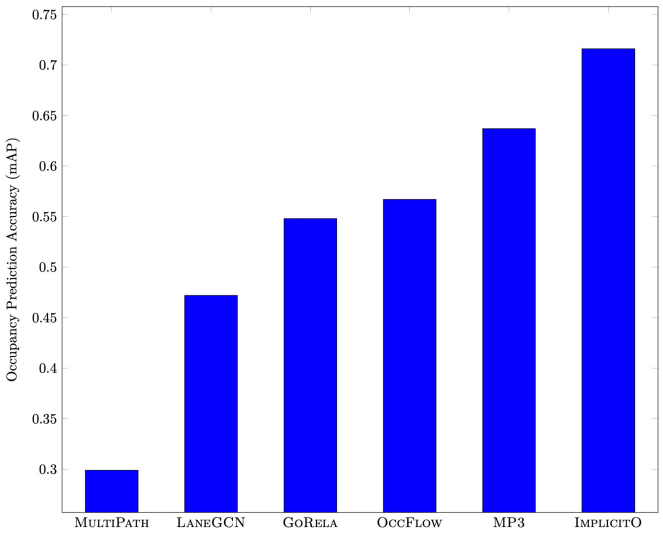
Quantitative Comparison: Argoverse



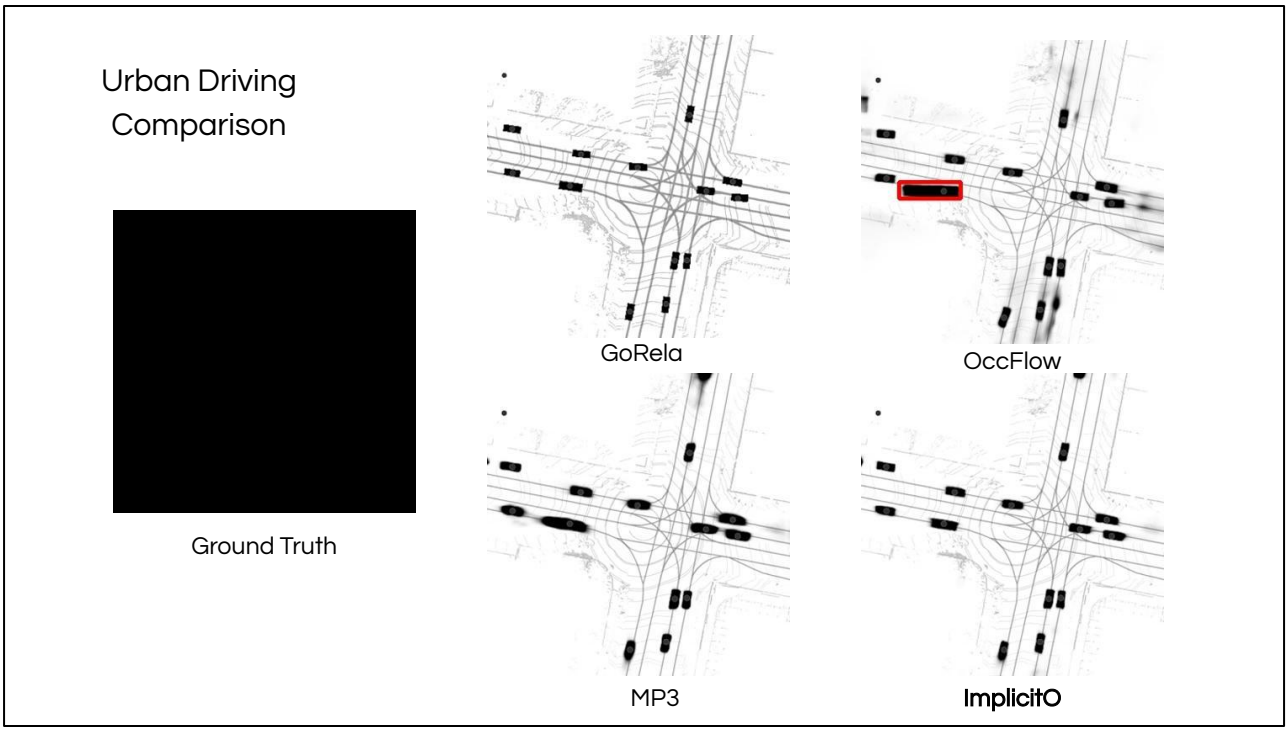
We evaluate our method against the state of the art on two datasets. The first is Argoverse 2, an urban driving dataset collected in US cities.
(click) This is a scene from argoverse, where the LiDAR is at observation time, and the black boxes denote vehicle occupancy 5 seconds into the future.
(click) We evaluate occupancy prediction accuracy against (click) three object-based baselines, and (click) two object-free baselines.
(click) ImplicitO with one predicted offset outperforms the baselines.

Quantitative Comparison: HighwaySim



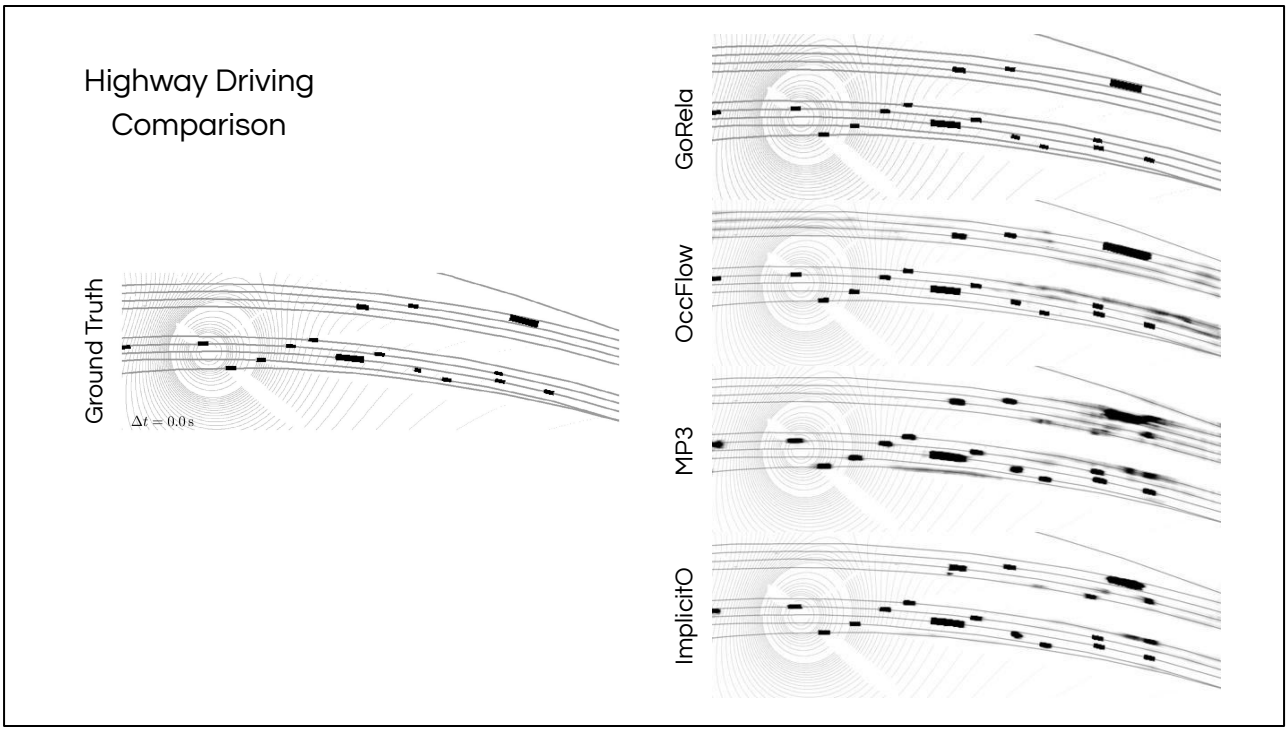


The second evaluation dataset is Highway Sim, which is a highway driving dataset generated with a state of the art simulator.
Again, ImplicitO outperforms the baselines on occupancy prediction accuracy.



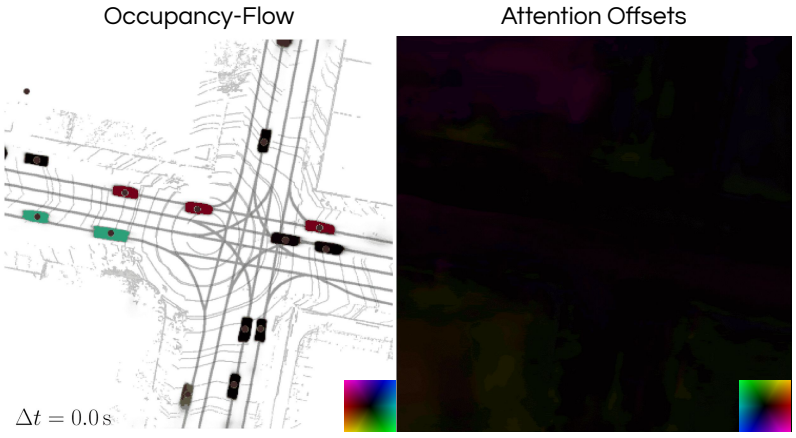
These animations show the occupancy predictions, in black, on argoverse. The alpha channel denotes to the occupancy probability.

(play) GoRela, an object-based model, predicts occupancy that is inconsistent across the actors in the scene. OccFlow shows a detection of the wrong size,
(play) and predicts a spread of occupancy in the later timesteps, caused by the limited receptive field of its convolutional decoder.
(play) MP3 exhibits disjoint-pixel occupancy predictions caused by its forward-flow warping mechanism.
(play) Our model, ImplicitO, overcomes these problems, and demonstrates realistic multi-modal occupancy predictions.



Here we show occupancy predictions on HighwaySim.
(click) We observe that ImplicitO outperforms the baselines with more accuracy occupancy predictions, particularly at later timesteps. Why is this?

Model Introspection

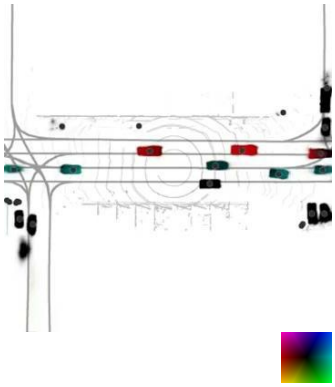


In the right plot, we visualize the implicit decoder's attention offsets, and observe that ImplicitO has learned a simple, yet effective policy to find the most relevant features: looking backwards along the lane, backtracking an object's trajectory to the LiDAR evidence, allowing it to outperform the state-of-the-art baselines.

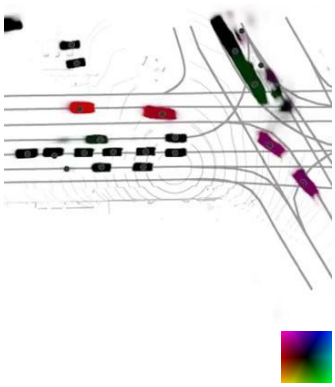
Unrolled Predictions on Argoverse

$\Delta t = 0.0$ s

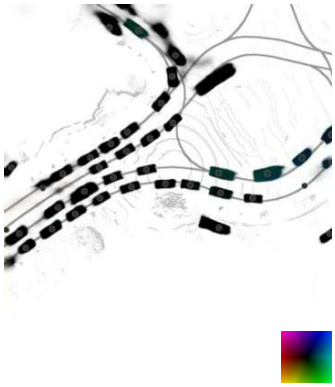
Sequence 1



Sequence 2



Sequence 3



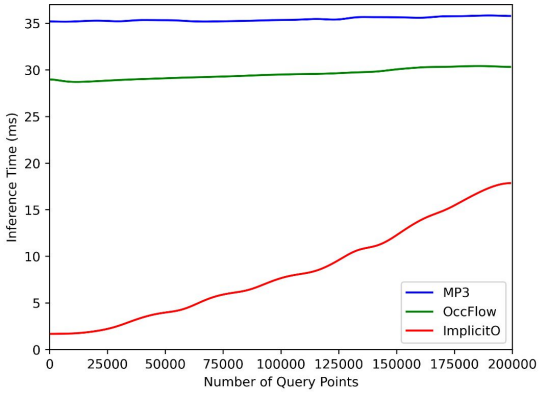
Here (click) we visualize the predictions of ImplicitO at the present time, unrolled across an entire sequence from argoverse 2. (wait)
(click) ImplicitO produces accurate occupancy-flow perceptions across each sequence, capturing vehicles of all shapes and sizes (wait)
(click) It also models uncertainty in occupancy in occluded regions (wait)

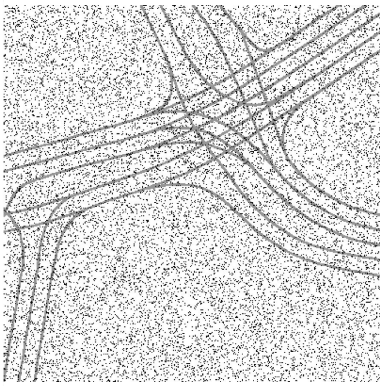
Unrolled Predictions on Argoverse

$\Delta t = 0.0$ s $\Delta t = 2.5$ s $\Delta t = 5.0$ s

Here we visualize the unrolled predictions of ImplicitO on an entire sequence, at different prediction horizons (click)
At later time horizons, ImplicitO captures the uncertainty inherent in drivers actions (wait)

Inference Time vs Number of Query Points





20k query points per timestep

(200 K total for 10 timesteps)

We observe that for a realistic number of query points, ImplicitO has a lower inference time than the other object-free models because it only predicts occupancy and flow where it is queried.

For illustration, we visualize 20,000 randomly-placed query points to show that they cover the majority of the scene.

Conclusion

Introduced **ImplicitO**, a state-of-the-art perception and prediction model:

- **Object-free**, unified perception and prediction
- Queryable, **implicit** continuous representation of occupancy flow
- **Outperforms prior work** in both urban and highway settings
  - Prediction quality
  - Inference time

In conclusion, this work has introduced ImplicitO, a novel object-free perception and future prediction model that provides a queryable implicit representation of occupancy-flow. We have shown that ImplicitO outperforms the state of the art methods in both urban and highway settings on the task of occupancy-flow prediction.

Please see our paper and supplementary for more details.