Autonomous Driving Trajectory Prediction

Aman Sharma, Vyshnav Achuthan, Neha Madhekar, Rishikesh Jadhav, Xiyang Wu

Problem Statement

Bird's Eye View (BEV) is less explored in prediction, especially under only a multi-camera setup. PowerBEV is one of the models using BEV. It outperforms state-of-the-art multi-camera baselines on the NuScenes dataset. This project aims to analyze its performance on other datasets and check the generalizability of such models.

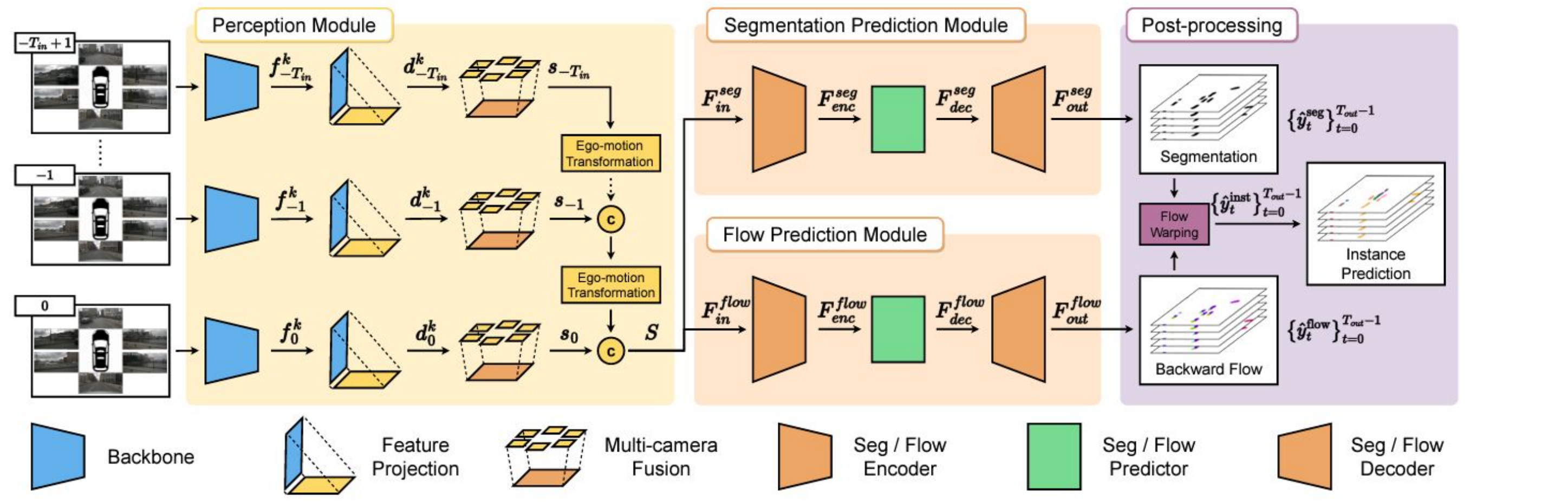


Figure 2: **Architecture of our Proposed End-to-End Framework:** In PowerBEV, the perspective features extracted by the perception module (yellow area) from surrounding camera images of each frame are projected into the BEV plane and then fused and stacked into the current global dynamic state. Subsequently, two independent prediction modules with the same structure (orange area) take the current state as input and predict the segmentation maps and centripetal backward flow for the future frames. Finally, future multi-frame instance predictions are generated by the flow warping post-processing (purple area).

Dataset

Dataset	NuScenes	Woven
# of Cameras	6	6
Camera Intrinsics and Extrinsics	✔	✔
Annotations	3D boxes	3D boxes
# of Scenes	850	180
Size	250 GB	50 GB
Other Sensors (not used for prediction)	Radar, Lidar	Lidar
# Scenes in Subsampled Dataset	85	84

Methodology

1. Use the NuScenes and Woven dataset for evaluation.
2. Use the model weights trained on NuScenes and evaluate on Woven directly without fine-tuning to show how transferable NuScenes-trained models are.
3. Train a model on Woven from scratch and evaluate it on itself to create a baseline and compare it with fine-tuned models using weights pre-trained on NuScenes.
4. Fine-tune on Woven using NuScenes pre-trained model weights to show the model's capacity in generalization after some fine-tuning.
5. Extract a smaller dataset (sub-dataset) by sampling from both datasets and evaluating it.

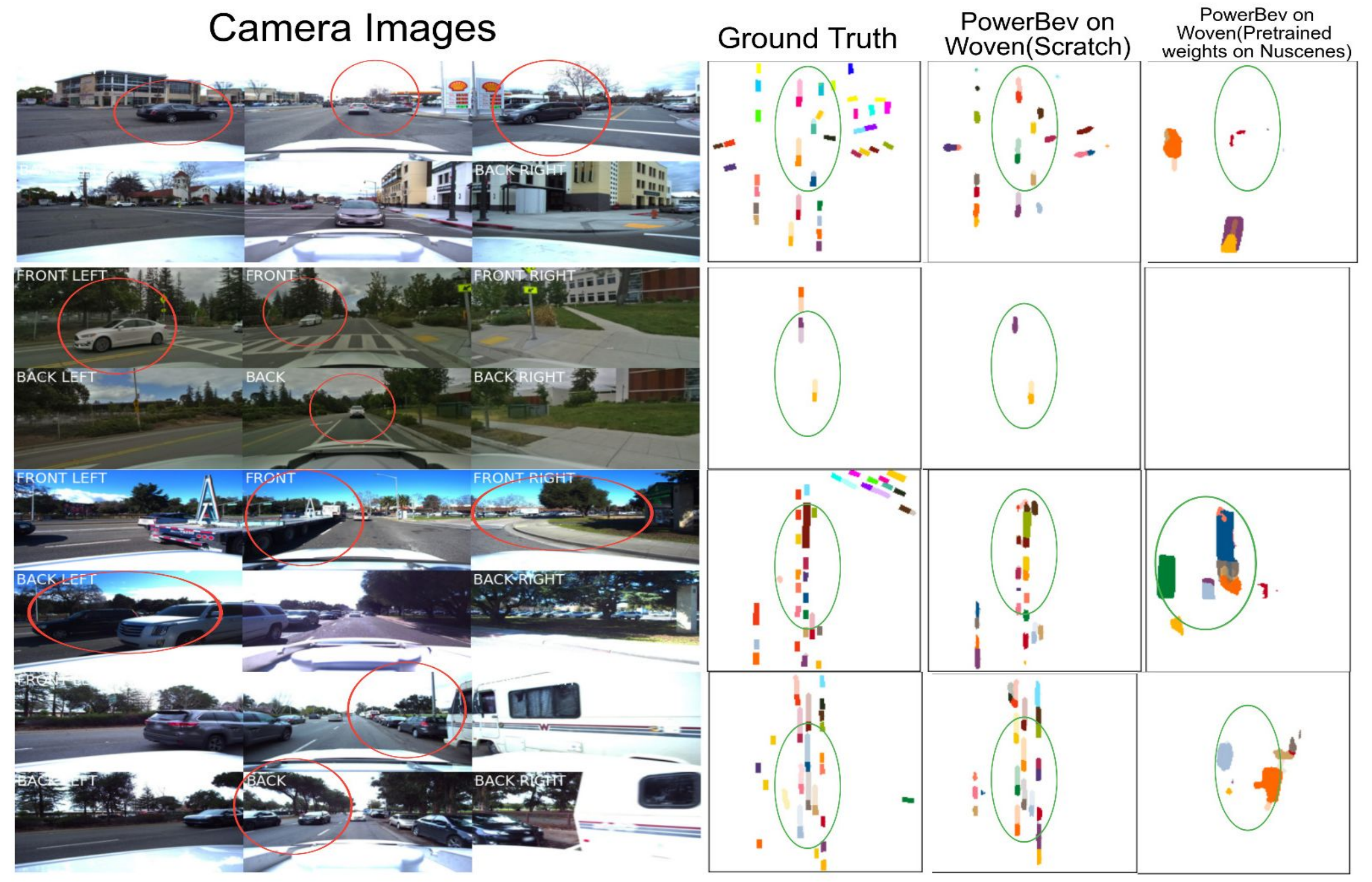
Results

Dataset	Version	Model Used for Eval	IOU	VPQ	Total Time	GPU
NuScenes	Complete Dataset (20 Hz)	Values Reported	0.625	0.555	-	3090
NuScenes	2 Hz sampled	Trained from Scratch	0.923	5.48E-06	2.299	3050 TI
Woven	Complete Dataset	NuScenes Model	0.918	0.008	2.136	3050 TI
Woven	Complete Dataset	Trained from Scratch	0.967	0.322	0.257	A100
Woven	Complete Dataset	Fine-Tuned from NuScenes	0.956	0.011	0.228	A100
Woven	Sub-Dataset	NuScenes Model	0.949	4.55E-07	0.229	A100
Woven	Sub-Dataset	From Scratch	0.958	0.238	0.524	A100

Findings

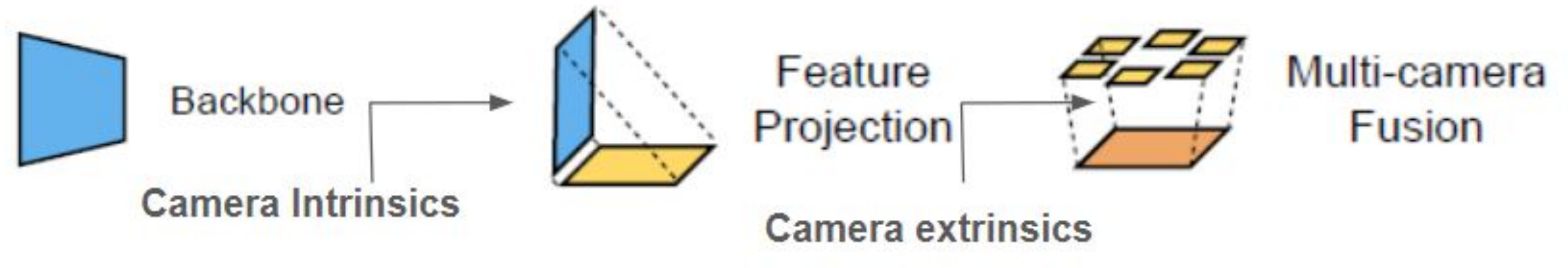
Image Size - We are using the image size that has not been compared in the paper but is in sync with other baselines. Using a smaller image size could be the reason for higher IOU values than that reported in the paper.

Model Efficiency: Perception Module took a 85-95% of the total time which mainly consists of backbone 'EfficientNet' used in the feature extractor.It could be replaced with much faster alternatives.


Camera Images
Ground Truth
PowerBev on Woven(Scratch)
PowerBev on Woven(Pretrained weights on Nuscenes)
FRONT LEFT
FRONT
FRONT RIGHT
BACK LEFT
BACK
BACK RIGHT

Analysis

Model (Perception Module) dependent on camera parameters. Evident when directly evaluated on different datasets.

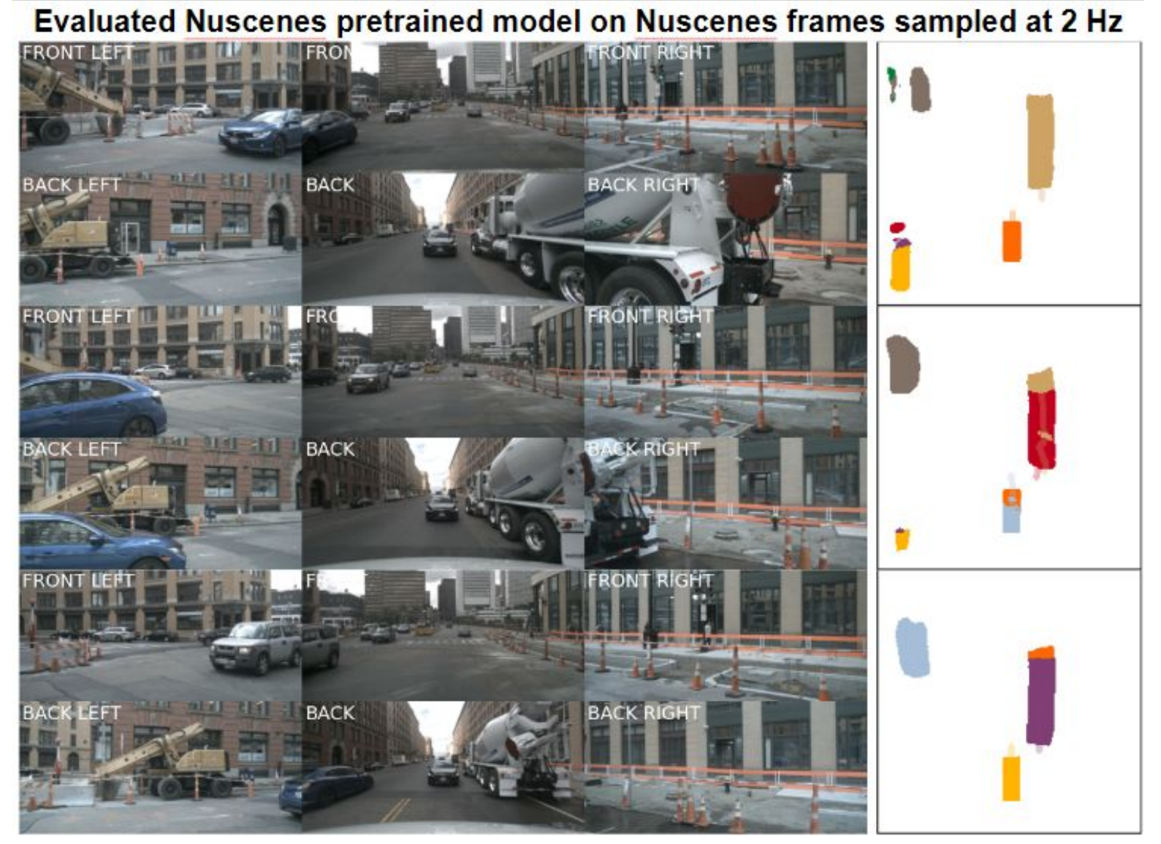


Perception Module: Model trained for specific Camera parameters.

Qualitative Analysis

Evaluated on woven dataset with a model trained from scratch

Evaluated Nuscenes pretrained model on Nuscenes frames sampled at 2 Hz



1. A model trained from scratch on the Woven dataset does not switch the IDs much in the scenarios having less number of vehicles which is the reason for a decent VPQ value.
2. However, it is not able to keep the IDs constant when there are comparatively more vehicles.

1. Keyframes samples at **2 Hz instead of 20 Hz** from NuScenes. The model does not translate well.
2. The quick ID switches reflect the low VPQ values due to the change in terms of the segmentation module even though IOU remains high.

Conclusion

1. PowerBEV is not generalizable on other datasets.

2. A model trained from scratch performs better than the other alternatives due to various reasons such as training done for specific camera parameters or variability in the dataset collected at different frequencies.

3. Link to GitHub Repository. Scan Me: