A. Supplemental material

A.1. Details on nuScenes and Argoverse 1 protocol

Preprocessing. We subsampled the datasets to 10 HZ with a tolerance of 0.1 HZ. We discard all scan sequences that violate this tolerance to ensure a high quality. While the datasets provide annotations for all, only a subset of frames was labeled by humans. For all remaining frames labeles were generated automatically. To ensure the highest possible quality, we evaluate only on scan sequences where at least one of the frames was annotated by a human. To ignore the ego vehicle, we discard all points within a radius of 3 meters around the origin. We also discard points higher than 4 meters and further away than 50 meters, as the point clouds become very sparse, which can result in noisy pseudo flow.

Before creating the pseudo ground truth flow, we discard points belonging to the ground plane; as for these points, the flow is commonly noisy (due to the circular scanning patterns on the ground plane). We discard ground points by estimating the ground plane using Progressive Morphological Filtering (PMF) [31]. PMF applies a series of filtering steps and progressively refines the ground plane estimate.

To compensate for the ego-motion, we use the remaining points after discarding the ground points and filter all points labeled as potentially dynamic points in the datasets. The remaining points are neither dynamic nor do they belong to the ground. We use these points to estimate a transformation between the point clouds by applying KISS-ICP [26]. We apply this transformation to all non-ground points in the point cloud (including the dynamic points) to align the static parts of the two point clouds. After this transform, we compute the flow for each point as described below.

Ground truth scene flow for 3D objects. After compensating for the ego-motion described above, we use the bounding box annotations of the datasets to estimate the flow of the dynamic objects. We compute the transformation from each bounding box to the corresponding annotation in the next frame and use the corresponding transformation to calculate the ground-truth flow of all points within the 3D bounding boxes. The annotated objects might not be moving but can be static, e.g., a parked car. To automatically label points as “static” or “dynamic” we compute the mean motion over all points of each potentially dynamic object. If this motion is larger than $0.5\ \mathrm{m/s}$ we label them as “dynamic”.

A.2. Baseline details and hyperparameters

Early stopping. For all experiments with NSFP and FNSF with early stopping activated, we keep the default parameters as published by the authors[1,2]. We set the early stopping patience to 100 epochs and early stopping minimum delta to 0.0001. The maximum number of epochs is set to 5000.

Learning rate. All baseline models are trained with the Adam optimizer. We set the learning rate for FNSF-8 to its default value of 0.001. In the case of NSFP-8 and NSFP-16, we diverge from the default value and set the learning rate to 0.0008. Finally, staying in line with the default settings, we do not use weight decay.

MLP hidden units. For all baseline experiments, we keep the number of hidden units to its default value of 128.

A.3. Floxels details and hyperparameters

Early stopping. For experiments with Floxels, we train for a maximum of 500 epochs, set the early patience to 250 epochs, and set the early stopping minimum delta to 0.01. Whenever it is stated that early stopping is identical to the baselines, we use the early stopping as described in Sec. A.2 and set the maximum epochs to 5000 for consistency.

A.4. Complete quantitative results

In the experiments section, we omit results for static points. Tables 6, 7, 8, and 9 show the full results. The official leaderboard for Argoverse 2 (2024) Scene Flow Challenge can be found here[3].

A.5. Qualitative comparison of FNSF, FNSF with Floxel losses and Floxels

For the qualitative results in Sec. 4.2 and Sec. 4.3, we use a slight variation of FNSF. In particular, we extended the losses of FNSF with a rigidity loss (similar to our clustering loss) and trained an MLP with 16 instead of 8 layers. We use this variant as our primary qualitative baseline as it performs on average better on our qualitative comparison dataset. For completeness, we show the qualitative results for the original FNSF in Fig. 8 and Fig. 9.

We show further examples with a qualitative comparison of FNSF and Floxels in Fig. 10.

A.6. A closer look at the neural prior

In Sec.4.2, we compare qualitatively the influence of an MLP-based method and Floxels. To further separate the effects of the MLP vs. the voxel grid, we train an MLP using the Floxels losses. Fig. 8 reveals that both slower convergence and windmill artifacts are consequences of the MLP

[1] https://github.com/Lilac-Lee/Neural_Scene_Flow_Prior

[2] https://github.com/Lilac-Lee/FastNSF

[3] https://eval.ai/web/challenges/challenge-page/2210/leaderboard/5463

Table 5. **Static/Dynamic Normalized EPE on Argoverse 2 (2024) Scene Flow Challenge test set** [8]. Baseline scores from challenge leaderboard.

	Method	BG	car	other vehicle	pedestrian	wheeled VRU	mDEPE
Superv.	Flow4D [9]	0.005	0.087	0.150	0.216	0.127	0.145
	TrackFlow [8]	0.002	0.182	0.305	0.358	0.230	0.269
	DeFlow [32]	0.005	0.113	0.228	0.496	0.266	0.276
Unsupervised	NSFP [11]	0.034	0.251	0.331	0.722	0.383	0.422
	Fast NSF [12]	0.091	0.296	0.413	0.500	0.322	0.383
	Zeroflow XL 5x [23]	0.013	0.238	0.258	0.808	0.452	0.439
	Liu et al. 2024 [14]	0.106	0.310	0.559	0.509	0.276	0.413
	SeFlow [33]	0.006	0.214	0.291	0.464	0.265	0.309
	Euler Flow [24]	0.053	0.093	0.141	0.195	0.093	0.130
	Floxels 5 0.5 m (ours)	0.024	0.119	0.194	0.243	0.113	0.168
	Floxels 9 0.5 m (ours)	0.018	0.108	0.202	0.208	0.100	0.155
	Floxels 13 0.5 m (ours)	0.015	0.112	0.213	0.195	0.096	0.154

Table 6. **Results on nuScenes validation set.** Models trained without early stopping (5000 epochs) denoted with “*”. “-N” indicates the number of layers. For a fair comparison we provide timings only when using early stopping.

Method	Dynamic Points				Static Points			
	EPE ↓	Acc_5 ↑	Acc_{10} ↑	angle error ↓	EPE ↓	Acc_5 ↑	Acc_{10} ↑	Time (s)
	Supervised							
DifFlow3D	0.089	0.554	0.823	0.325	0.044	0.748	0.935	0.48
	Self-supervised test-time optimization							
NSFP-8	0.141	0.316	0.636	0.471	0.068	0.613	0.861	**3.43**
NSFP-8*	0.139	0.315	0.641	0.470	0.067	0.613	0.861	–
NSFP-16	0.148	0.322	0.647	0.488	**0.061**	0.664	**0.883**	8.99
NSFP-16*	0.145	0.384	0.679	0.460	0.087	0.619	0.818	–
FNSF-8	0.266	0.211	0.501	0.628	0.137	0.439	0.723	5.93
FNSF-8*	0.372	0.122	0.361	0.757	0.241	0.241	0.531	–
Floxels (5s)	**0.102**	**0.464**	**0.786**	**0.430**	0.063	**0.755**	0.881	4.35

and are primarily solved by the voxel grid. Nevertheless, it can be seen in Fig. 8c that windmill artifacts are less pronounced when using the Floxels losses to train the MLP in comparison to the FNSF losses. Furthermore, equipped with the multi-frame Floxels loss, the MLP can predict the flow in the challenging occluded region. Together, these results highlight that both the voxel grid and the Floxels losses contribute to the superior performance of Floxels and solve different failure cases.

A.7. Convergence speed and optimization videos

Also visually, the convergence speed is much faster and more stable. We provide various videos of the optimization progress here: https://www.youtube.com/playlist?list=PLCtNe14NZWtVjaoW_KDc19Kb-oThHNA2S. We would like to explicitly highlight the differences between “Flow Field evolution for Floxels” and “Flow Field Evolution FNSF”. Further, we would like to highlight the difficulties in removing the windmill artifacts, from MLP + Floxels losses (“Flow Field Evolution Custom MLP with Floxel Losses”).

Table 7. **Results on Argoverse test set.** Models trained without early stopping (5000 epochs) denoted with "*". "-N" indicates the number of layers.

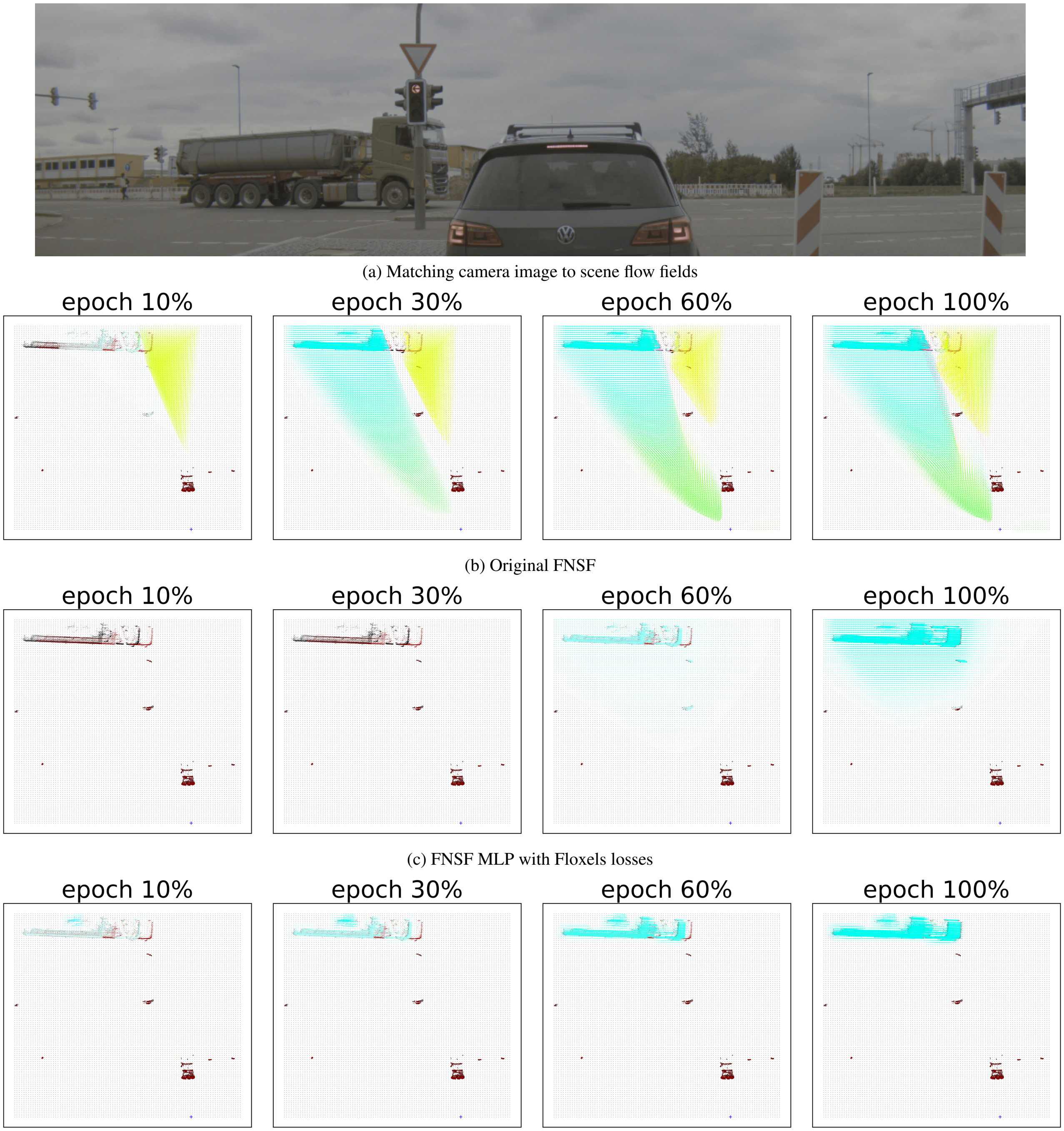
Method	Dynamic Points				Static Points			Time (s)
	EPE ↓	Acc$_5$ ↑	Acc$_{10}$ ↑	angle error ↓	EPE ↓	Acc$_5$ ↑	Acc$_{10}$ ↑	
	Supervised							
DifFlow3D	Out of Memory							
	Test-time optimization with same early stopping							
NSFP-8	0.200	0.288	0.521	0.468	0.046	0.745	0.933	63.01
NSFP-16	0.226	0.280	0.498	0.530	0.045	0.772	0.942	72.54
FNSF-8	0.282	0.281	0.518	0.588	0.065	0.704	0.905	20.66
Floxels (5s)	**0.104**	**0.537**	**0.755**	**0.420**	**0.024**	**0.919**	**0.962**	**4.38**
	Test-time optimization.							
NSFP-8*	0.202	0.272	0.508	0.478	0.047	0.743	0.931	-
NSFP-16*	0.203	0.336	0.541	0.495	0.043	0.815	0.948	-
FNSF-8*	0.370	0.215	0.458	0.651	0.148	0.463	0.723	-
Floxels (5s)	**0.109**	**0.526**	**0.739**	**0.423**	**0.024**	**0.912**	**0.962**	-

Table 8. **Influence of the number of scans.** Using nuScenes mini.

Method	Dynamic				Static			
	EPE ↓	Acc$_5$ ↑	Acc$_{10}$ ↑	angle error ↓	EPE ↓	Acc$_5$ ↑	Acc$_{10}$ ↑	Time (s) ↓
3 scans	0.095	0.468	0.799	0.524	0.064	0.719	0.887	**2.47**
5 scans	0.085	**0.537**	0.833	0.489	0.057	0.797	0.909	3.52
7 scans	0.082	0.533	0.839	0.474	0.051	0.816	0.916	4.61
9 scans	0.078	0.516	0.852	0.460	0.048	0.830	0.923	5.69
11 scans	**0.076**	0.486	**0.864**	**0.447**	**0.045**	**0.840**	**0.929**	6.72

Table 9. **Influence of different loss components.** Results obtained on nuScenes mini. All models use five scans. "-" indicates that the respective component got removed.

Method	Dynamic Points				Static Points		
	EPE ↓	Acc$_5$ ↑	Acc$_{10}$ ↑	angle error ↓	EPE ↓	Acc$_5$ ↑	Acc$_{10}$ ↑
Floxels	0.085	0.537	0.833	0.489	0.057	0.797	0.909
- flow norm	0.084	0.528	0.833	0.487	0.069	0.735	0.879
- cluster loss	0.201	0.133	0.413	0.802	0.153	0.205	0.493
- cluster loss and - flow norm	0.206	0.123	0.401	0.793	0.182	0.153	0.420



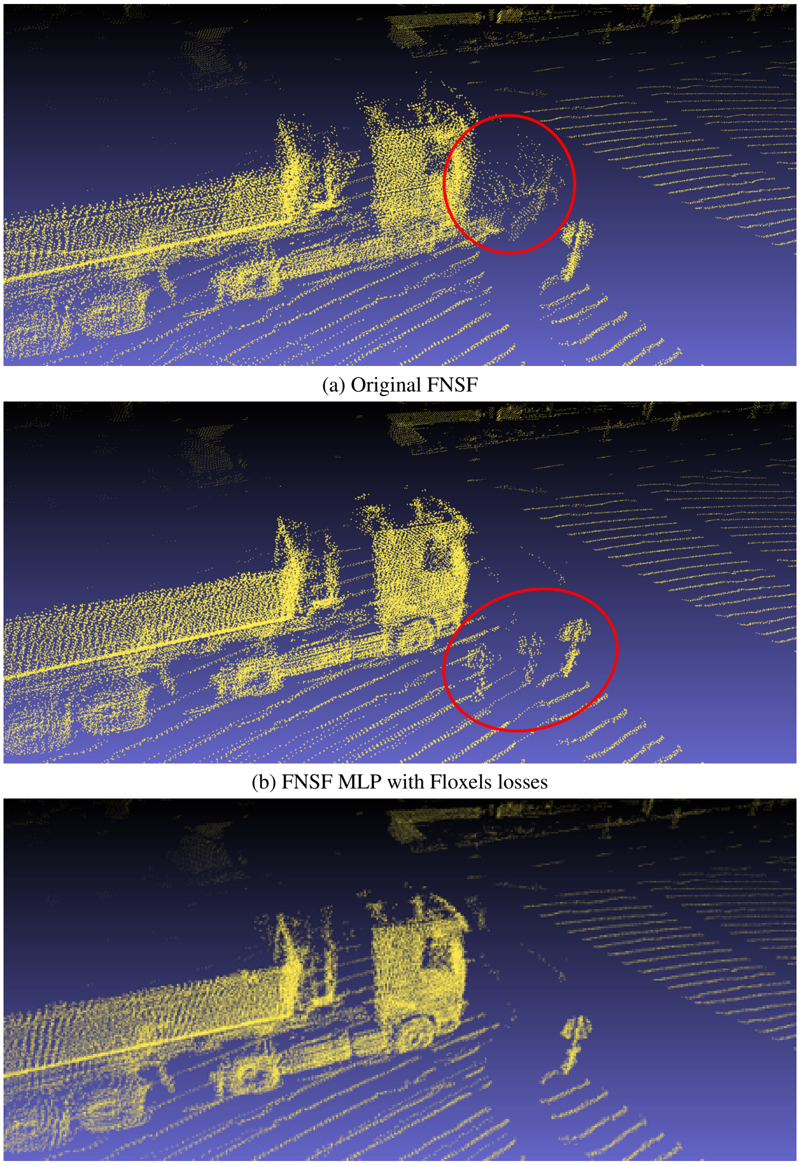
(a) Matching camera image to scene flow fields

(b) Original FNSF

(c) FNSF MLP with Floxels losses

(d) Floxels

Figure 8. **Evolution of scene flow comparison between FNSF, FNSF with Floxels losses and Floxels**. We show a birds-eye view of the estimated flow during optimization. FNSF exhibits problems in occluded regions and strong "windmill artifacts". For FNSF MLP with Floxels losses we observe that multi-frame and cluster losses help in occluded regions. Full Floxels also predicts zero-flow in empty regions and converges faster. Points at time t are black and $t+1$ are red. Other colors are scene flow.

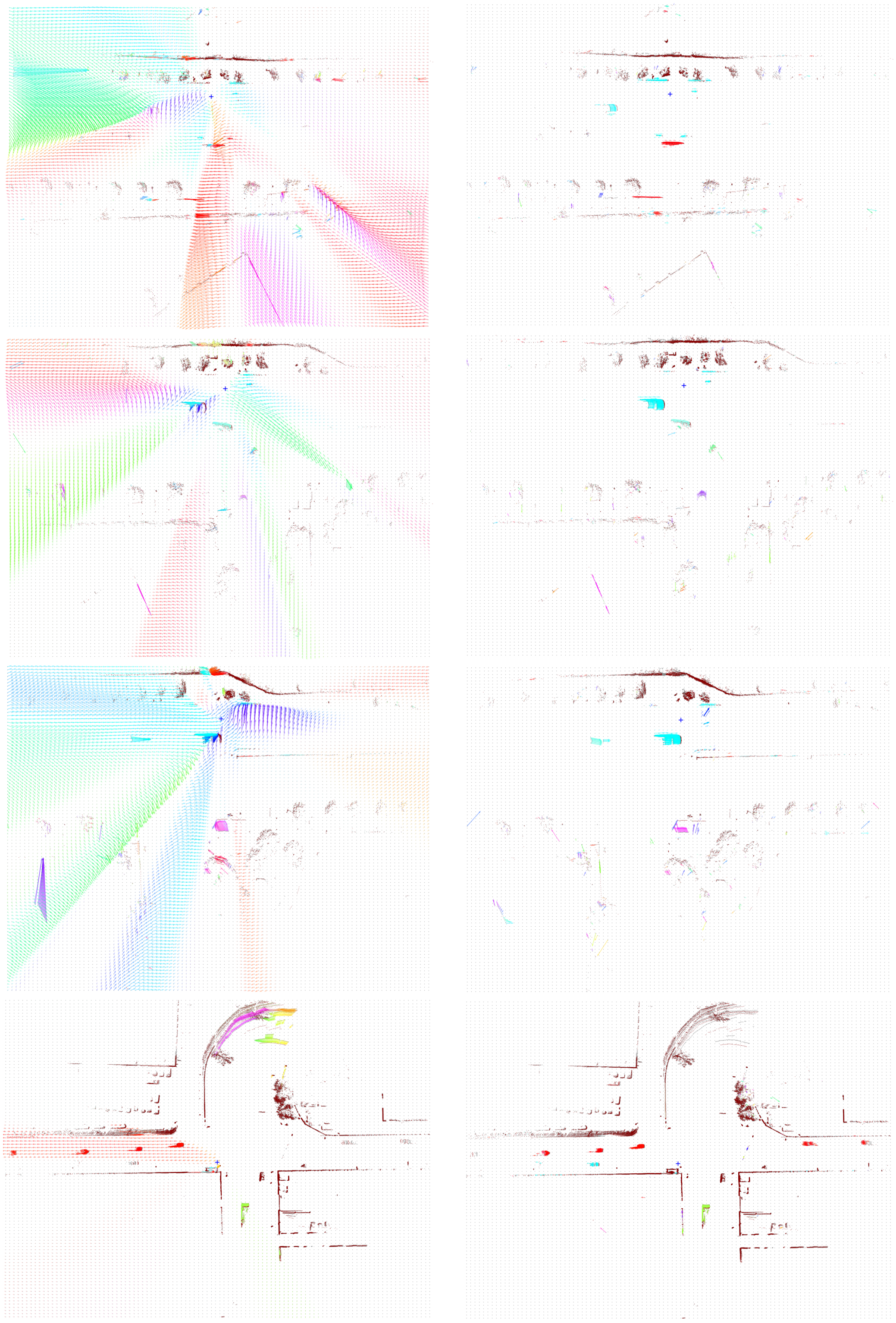


(a) Original FNSF

(b) FNSF MLP with Floxels losses

(c) Floxels

Figure 9. **Accumulation over time compared between FNSF, FNSF with Floxels losses and Floxels**. We accumulate five point clouds t-2, t-1, t, t+1, and t+2. For FNSF parts of the front of the truck are moved too far forward. For FSNF MLP with Floxels loss the traffic sign is falsely affected by the scene flow field, which makes it appear 3 times in the accumulated point cloud. Floxes shows a much cleaner accumulated point cloud with more details.



(a) FNSF

(b) Floxels

Figure 10. **Comparison of scene flow fields after convergence**. We show a birds-eye view of the scene flow fields for FNSF (left) and Floxels (right) after convergence. Points at time t are black, and $t + 1$ are red. Floxels does well at isolating the dynamic environment whereas FNSF struggles to do so. Consequently, FNSF sometimes predicts zero-flow on dynamic objects and noisy flow vectors in the static regions as depicted above.