TeFlow: Enabling Multi-frame Supervision for Self-Supervised Feed-forward Scene Flow Estimation

Supplementary Material

A. Datasets Description

We evaluate our method on three major autonomous driving benchmarks: Argoverse 2, nuScenes, and Waymo Dataset. The Argoverse 2 dataset is a primary benchmark for scene flow estimation, featuring 700 training, 150 validation, and 150 test scenes, totaling approximately 107,000 training frames. Our main evaluations are conducted on the official test split, with results compared against the Argoverse 2 Scene Flow Challenge leaderboard [15], which provides official baseline results. For the local validation, we follow [43] and apply dynamic motion compensation to generate ground-truth flow labels.

The nuScenes dataset [3] contains 700 training and 150 validation scenes. Since nuScenes does not provide official scene flow annotations, we follow the protocol of [42] to generate pseudo ground truth. To ensure a consistent temporal resolution, the native 20Hz LiDAR data is first downsampled to 10Hz, resulting in a standard 100ms interval between frames. For each object, a rigid transformation is estimated from its 3D bounding box annotations and instance ID. This transformation is then applied to all LiDAR points within the object to compute their displacements, which serve as pseudo ground-truth flow labels. These labels are generated only for the validation set for evaluation, while training uses the full 137,575 unlabeled frames, demonstrating the scalability of our self-supervised approach.

The Waymo dataset [14, 31] is captured by a custom 64-channel LiDAR and contains 798 training and 202 validation sequences, each recorded at 10 Hz for around 20 seconds. The training set consists of 155,000 frames.

For all datasets, ground points are removed prior to evaluation. In Argoverse 2 and Waymo, we use the provided HD maps following the official protocol, whereas in nuScenes we apply a line-fitting-based ground segmentation method [10]. As a result, all reported evaluation metrics are computed exclusively on non-ground points.

B. Additional Quantitative Analysis

State-of-the-art Comparison in Waymo Following our extensive evaluation on Argoverse 2 and nuScenes in the main text, we further validate TeFlow on the Waymo dataset to demonstrate the method's robustness across diverse scenarios and sensor configurations. As shown in Table 7, compared to the second-best feed-forward baseline (SeFlow++), TeFlow achieves a 14.9% reduction in Dynamic Bucket-Normalized EPE (0.275 vs. 0.323), with consistent improvements across all object categories. Notably, the error for VRU (e.g., cyclist) is reduced by 19.0% (0.198 vs. 0.247), highlighting the benefit of reliable multi-frame temporal modeling for tracking small, fast-moving agents.

In terms of absolute accuracy on the standard Three-way EPE metric, TeFlow maintains its lead over all baselines. It attains a mean EPE of 3.48 cm, surpassing SeFlow++ (3.63 cm), while achieving the lowest dynamic EPE (8.61 cm), significantly outperforming both feed-forward and per-scene optimization-based methods like NSFP (17.12 cm) and ICP-Flow (20.81 cm). These consistent results across three major datasets confirm that TeFlow establishes a significantly more effective self-supervised learning framework, capable of mining reliable multi-frame supervision across diverse scenarios.

Different Multi-frame Backbone Table 8 evaluates the generality of our self-supervised framework across different multi-frame backbones. We first adapt Flow4D, an architecture originally designed for supervised learning, to the self-supervised setting. Enabled by TeFlow's reliable supervision, Flow4D successfully learns to perform temporal reasoning without ground-truth labels and achieves a respectable mean dynamic normalized error of 0.330. Applying the same framework to the more advanced ΔFlow backbone yields a performance boost on all categories, reducing the overall dynamic error by 19.7% (0.330 to 0.265) and three-way EPE by 22.3% (5.70 to 4.43). This improvement aligns with the supervised results [43], where ΔFlow demonstrates stronger temporal representation and motion modeling than Flow4D. Together, these experiments show that TeFlow is backbone-agnostic: it unlocks the self-supervised potential of existing multi-frame architectures (e.g., Flow4D and ΔFlow) and can be seamlessly applied to future scene flow backbones as they emerge.

Ablation Study on Candidate Voting and Aggregation As detailed in Section 4.1, our consensus formulation operates in two stages: (a) voting, which scores candidates based on directional consistency agreement $\mathbf{M}$ and reliability

Table 7. Performance comparisons on the Waymo validation set. TeFlow achieves state-of-the-art accuracy in scene flow estimation. Runtime is reported per sequence (≈200 frames) using the same device. The best results are shown in **bold**.

Methods	#F	Runtime per seq	Dynamic Bucket-Normalized ↓				Three-way EPE (cm) ↓			
			Mean	CAR	PED	VRU	Mean	FD	FS	BS
Ego Motion Flow	-	-	1.000	1.000	1.000	1.000	17.10	50.85	0.46	0.00
Optimization-based										
NSFP [21]	2	96m	0.574	0.315	0.823	0.584	10.05	17.12	10.81	2.21
ICP-Flow [24]	2	11m	0.328	0.305	0.485	0.195	8.50	20.81	2.14	2.57
FastNSF [22]	2	6.7m	0.458	0.236	0.719	0.418	9.24	18.19	2.56	6.98
Feed-forward										
ZeroFlow [32]	2	11.5s	0.770	0.444	0.982	0.884	8.64	22.40	1.57	1.96
SeFlow [41]	2	12s	0.351	0.212	0.551	0.289	4.29	10.49	1.39	1.00
VoteFlow [25]	2	13.5s	0.347	0.197	0.548	0.298	3.89	9.65	1.12	0.88
SeFlow++ [42]	3	16s	0.323	0.201	0.521	0.247	3.63	9.30	**0.87**	**0.71**
TeFlow (Ours)	5	14s	**0.275**	**0.157**	**0.469**	**0.198**	**3.48**	**8.61**	0.97	0.86

Table 8. Ablation study on different multi-frame backbones within our self-supervised pipeline TeFlow. Results are evaluated on the Argoverse 2 validation set with five input frames. The results demonstrate that our self-supervised objective integrates seamlessly with different multi-frame backbones (Flow4D and ΔFlow), showing that the method is architecture-agnostic and applicable to future multi-frame scene flow designs.

Backbone	Dynamic Bucket-Normalized ↓					Three-way EPE (cm) ↓			
	Mean	CAR	OTHER	PED.	VRU	Mean	FD	FS	BS
Flow4D	0.330	0.254	0.326	0.329	0.411	5.70	12.98	2.67	1.46
ΔFlow	**0.265**	**0.198**	**0.275**	**0.295**	**0.293**	**4.43**	**10.36**	1.86	1.08

Table 9. Ablation of the candidate voting and aggregation pipeline. We evaluate three components: the directional consistency agreement matrix $\mathbf{M}$ in Equation (5), the reliability weights $\mathbf{w}$ in Equation (6), and the aggregation step in Equation (8). Removing either directional consistency or reliability weighting degrades performance, with $\mathbf{M}$ playing the dominant role in stabilizing candidate votes. Skipping aggregation further reduces accuracy by discarding supportive consistent candidates. Results on the Argoverse 2 validation set demonstrate that the full pipeline is essential for producing reliable multi-frame supervision.

Exp. Id	Ablation Variant	Dynamic Bucket-Normalized ↓					Three-way EPE (cm) ↓			
		Mean	CAR	OTHER	PED.	VRU	Mean	FD	FS	BS
1	w/o $\mathbf{M}$ (i.e., $\mathbf{M} = \mathbf{1}$)	0.349	0.301	0.408	0.292	0.394	5.81	15.33	1.34	0.74
2	w/o $\mathbf{w}$ (i.e., $\mathbf{w} = \mathbf{1}$)	0.271	0.199	0.300	**0.285**	0.302	5.70	12.98	2.67	1.46
3	w/o Aggregation	0.289	0.226	0.332	0.303	0.295	4.61	11.55	1.47	0.82
4	TeFlow (Ours)	**0.265**	**0.198**	**0.275**	0.295	**0.293**	**4.43**	**10.36**	1.86	1.08

weights $\mathbf{w}$, and (b) aggregation, which averages all candidates consistent with the winner. We ablate these components to quantify their contribution to the supervisory signal.

1) Directional consistency agreement Equation (5). The directional consistency agreement $\mathbf{M}$ is used to determine which candidates in the pool reinforce each other during voting. We ablate it by replacing $\mathbf{M}$ with an all-ones matrix, ignoring agreement among candidates. As shown in Table 9 (Experiment 1), removing directional consistency leads to a notable performance drop: the dynamic bucket-normalized error increases from 0.265 to 0.349, corresponding to a 31.7% relative degradation. This demonstrates that directional agreement is important for filtering out inconsistent motion hypotheses and stabilizing the temporal ensembling process. The only exception is pedestrians, where performance slightly decreases. Their motion frequently changes direction within short windows, so enforcing strict agreement can suppress valid short-term cues.

2) Reliability weighting Equation (6). The reliability weights $\mathbf{w}$ are used to emphasize candidates that provide clearer motion cues. We ablate this by setting all weights to one, removing both magnitude-based emphasis and temporal decay. As shown in Table 9 (Experiment 2), this leads to a moderate increase in dynamic bucket-normalized error (0.265 to 0.271). The impact is particularly evident in absolute accuracy, where the mean Three-way EPE degrades significantly from 4.43 cm to 5.70 cm (a 28.7% error hike). This demonstrates that reliability weighting serves as a refinement for precision: while directional consistency filters outliers, weighting refines the consensus by emphasizing

stronger motion cues. Similar to the directional-consistency ablation (Experiment 1), pedestrians exhibit a slightly different trend, suggesting that strictly favoring larger motions may occasionally suppress short-term but valid cues for rapidly changing agents.

3) Flow aggregation Equation (8). In our flow aggregation, we use not only the consensus winner but also supporting candidates that agree with it. We ablate this aggregation step by supervising with the consensus winner alone, i.e., $\bar{\mathbf{f}}_{\mathcal{C}_j} = \mathbf{f}_{a^\dagger}$. As shown in Table 9 (Experiment 3), removing flow aggregation leads to a drop in accuracy: the dynamic bucket-normalized error increases from 0.265 to 0.289, a 9.1% relative increase. Although the consensus winner achieves the highest vote score, it remains a single estimation sample. Consequently, relying on it directly leaves the supervision susceptible to the specific noise or quantization artifacts of that individual candidate. In contrast, aggregating all directionally consistent flows reinforces stable temporal evidence and suppresses spurious single-frame deviations, leading to smoother supervision and consistently better overall accuracy.

Together, these ablations show that our proposed voting pipeline is necessary for producing stable and reliable multi-frame supervision. Directional consistency filters contradictions, reliability weights prioritize trustworthy cues, and aggregation consolidates evidence to mitigate isolated noise. Combining all three components yields the best self-supervised training performance by providing a reliable supervisory signal.

Analysis on Hyperparameter Selection We further analyze the sensitivity of TeFlow to its key hyperparameters by varying one parameter at a time while keeping the others fixed at their optimal values. Results are reported in Table 10 and provide additional insight into the functioning of the temporal ensembling strategy.

Table 10. Ablation study on the key hyperparameters of TeFlow, evaluated on the Argoverse 2 validation set. The default and best-performing configuration is cosine similarity $\tau_{cos} = 0.7$ (45°), Top-K = 5, and time decay $\gamma = 0.9$. In each row, only the specified parameter is varied from this setting.

TeFlow Setting	Dynamic Bucket-Normalized ↓					Three-way EPE (cm) ↓			
	Mean	CAR	OTHER	PED.	VRU	Mean	FD	FS	BS
Default	**0.265**	**0.198**	**0.275**	0.295	**0.293**	**4.43**	**10.36**	1.86	1.08
$\tau_{cos} = 0$ (90°)	0.307	0.239	0.365	0.291	0.332	5.19	13.02	1.60	0.95
$\tau_{cos} = 0.9$ (20°)	0.289	0.207	0.356	0.294	0.297	4.42	10.41	1.80	1.04
$K = 20$	0.353	0.283	0.355	0.312	0.463	5.88	14.97	1.68	1.00
$K = 10$	0.307	0.241	0.314	0.296	0.377	5.11	12.39	1.83	1.12
$\gamma = 1$	0.303	0.224	0.348	0.311	0.330	4.73	11.55	1.66	0.98
$\gamma = 0.5$	0.285	0.232	0.308	**0.290**	0.311	4.92	11.65	1.98	1.12

1) *Top-K* This parameter controls the number of external candidates in the candidate pool. A small, high-quality set proves most effective, with the best performance at $K = 5$. Larger values introduce noise from less reliable geometric matches and degrade accuracy.

2) *Cosine Similarity* This threshold determines which candidates are included in the consensus matrix. The optimal value of 0.707 (45°) strikes the right balance: looser thresholds allow inconsistent motions, while stricter ones discard valid candidates too early.

3) *Time Decay* This factor weights candidates by their temporal distance, giving higher importance to recent frames. Our default of $\gamma = 0.9$ outperforms both no decay ($\gamma = 1.0$) and stronger decay ($\gamma = 0.5$). Without decay, distant frames are treated equally and introduce noise, while overly strong decay underutilizes longer-term consistency that benefits large, predictably moving objects.

C. Qualitative Results

The qualitative results in the main paper are derived from the scenes '8749f79f-a30b-3c3f-8a44-dbfa682bbef1' and 'scene-0104' in the Argoverse 2 and nuScenes validation set, respectively.

Here, we present additional qualitative results comparing our TeFlow with top self-supervised feed-forward methods, namely SeFlow [41], VoteFlow [25], and SeFlow++ [42]. All visualizations use a standard color-coding scheme, where hue indicates motion direction and saturation encodes speed.

Figure 5 shows two complex multi-agent scenes from Argoverse 2. In the left scene, three oncoming vehicles are captured. While the ground truth indicates consistent forward motion, all baseline feed-forward methods occasionally

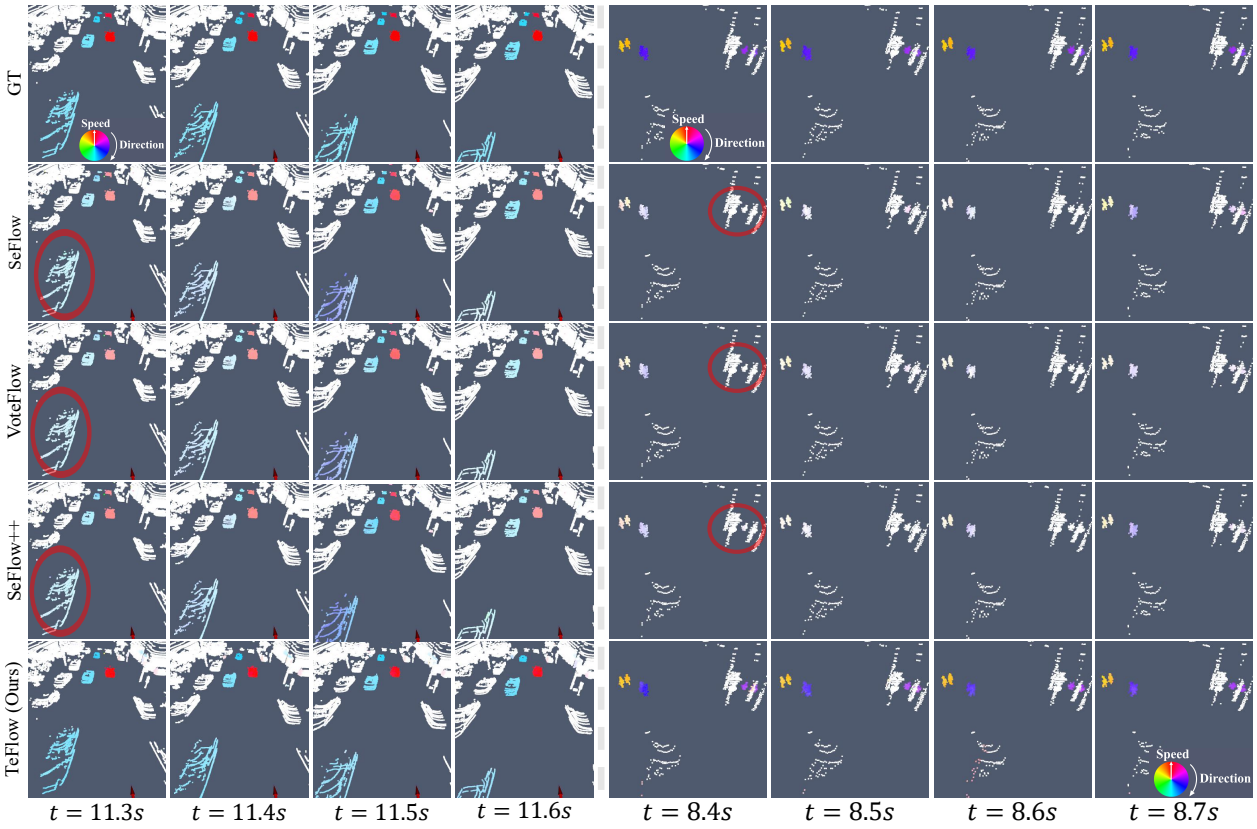

Figure 5. Qualitative comparisons on the Argoverse 2 validation set. Left: A multi-vehicle scene. Right: A vehicle stopping for pedestrians. Our method robustly handles both scenarios, unlike the baseline. (Best viewed in color.) The scenes correspond to scene IDs 'c85a88a8-c916-30a7-923c-0c66bd3ebbd3' and 'b6500255-eba3-3f77-acfd-626c07aa8621'.

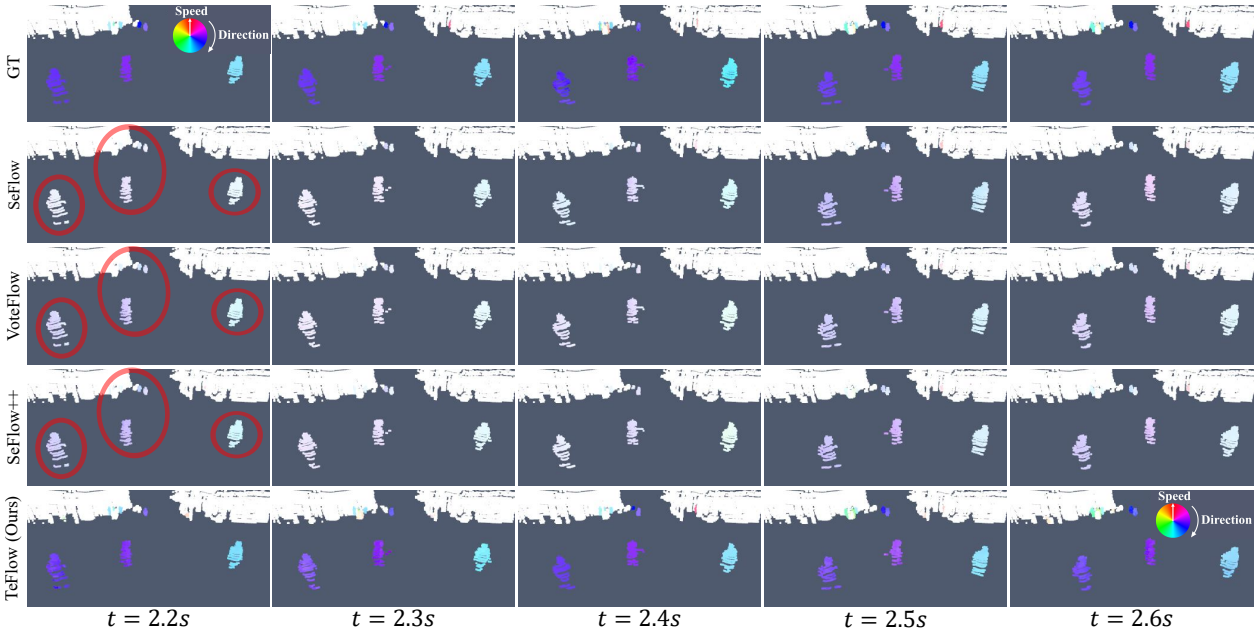

Figure 6. Qualitative results on the Argoverse 2 validation set. Our method accurately captures the motion of multiple pedestrians, while all feed-forward baselines underestimate the flows of moving pedestrians. (Best viewed in color.) The scenes correspond to scene IDs '9f871fb4-3b8e-34b3-9161-ed961e71a6da'.

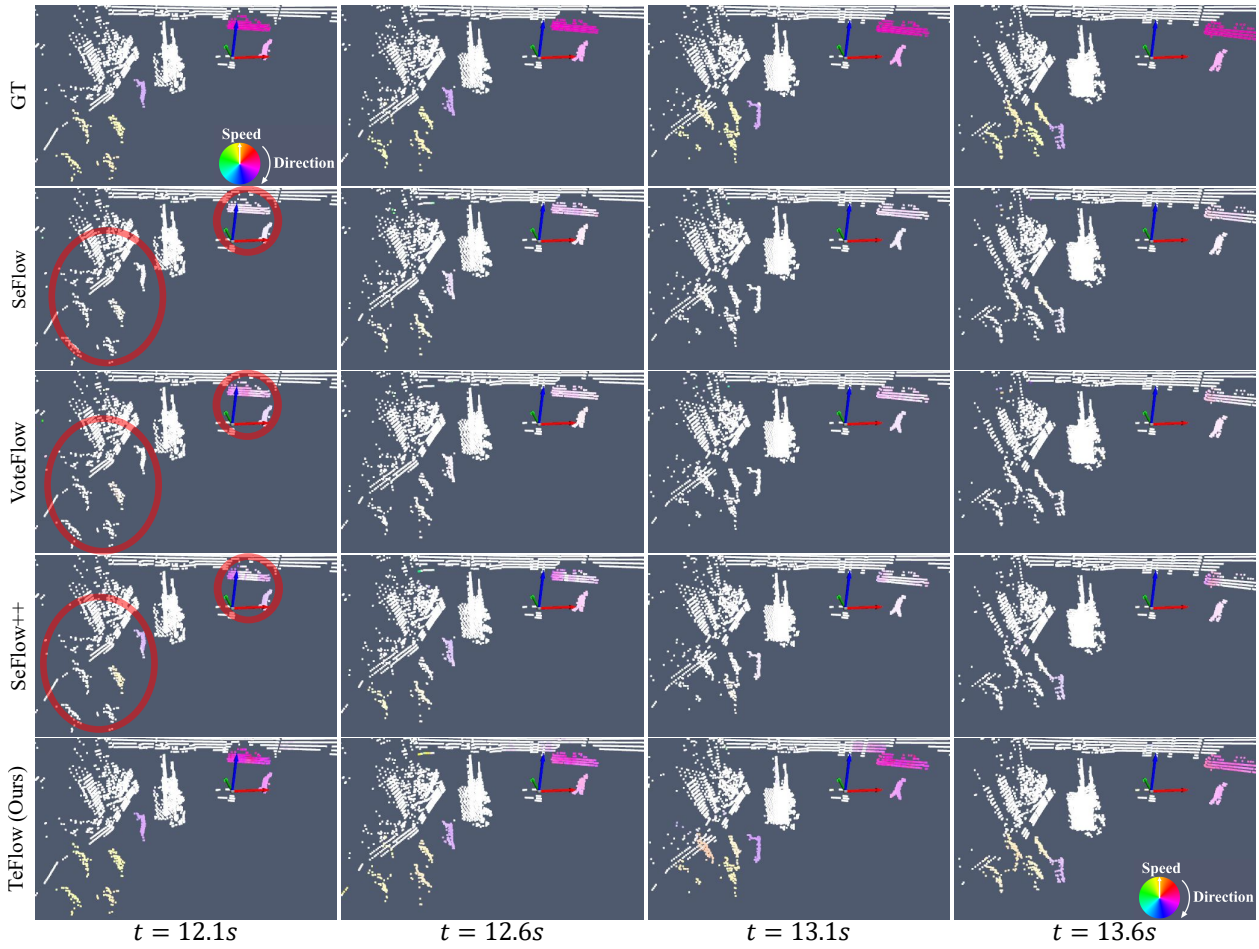

Figure 7. Qualitative results on the nuScenes validation set. On this sparser data, TeFlow provides complete motion for the vehicle and detects the pedestrians, whereas the baseline underestimates the car's flow and misses the smaller actors. (Best viewed in color.) The scenes correspond to the scene IDs 'scene-0025'.

predict conflicting directions (e.g., flows shift from blue to purple around $t = 11.3$–11.5), reflecting the instability of two-frame supervision. In contrast, our TeFlow maintains coherent motion across time, producing stable and accurate flow for each vehicle. The right scene highlights another common failure case: pedestrians motion. The ground truth reveals clear trajectories, including a distant pedestrian partially occluded by a lamp post. Baseline methods consistently underestimate the flow magnitudes of these small or occluded agents, resulting in weak or inconsistent predictions. While our TeFlow captures their motion with the correct magnitude and direction.

Figure 6 presents a challenging scene with three pedestrians crossing the road simultaneously. In the ground truth, all pedestrians exhibit clear motion, yet baseline feed-forward methods underestimate their flow magnitudes due to noisy two-frame supervision, resulting in weak and inconsistent predictions under such dynamic motion. In contrast, the model trained with our TeFlow objective produces flow fields that are both spatially coherent and temporally stable. Each motion of pedestrian is captured with accurate magnitude and direction, closely matching the ground truth across the time window. Furthermore, TeFlow also preserves reliable estimates for other small or distant dynamic objects, highlighting its robustness under challenging scenarios with sparse observations.

Figure 7 shows a challenging scene from the nuScenes validation set. In the lower-left corner, five pedestrians are walking together, while a vehicle and another pedestrian are passing in front of the ego car. The ground truth indicates clear motion for both the vehicle and pedestrians. However, baseline feed-forward methods significantly underestimate the vehicle's flow magnitude and often fail to detect the motions of the smaller pedestrians. In contrast, TeFlow produces a smooth and complete flow field for the vehicle and successfully captures the individual motions of the pedestrians, even under the sparse point density of nuScenes.

Figure 8 illustrates a complex roundabout scene from the Argoverse 2 validation set. Multiple vehicles are moving

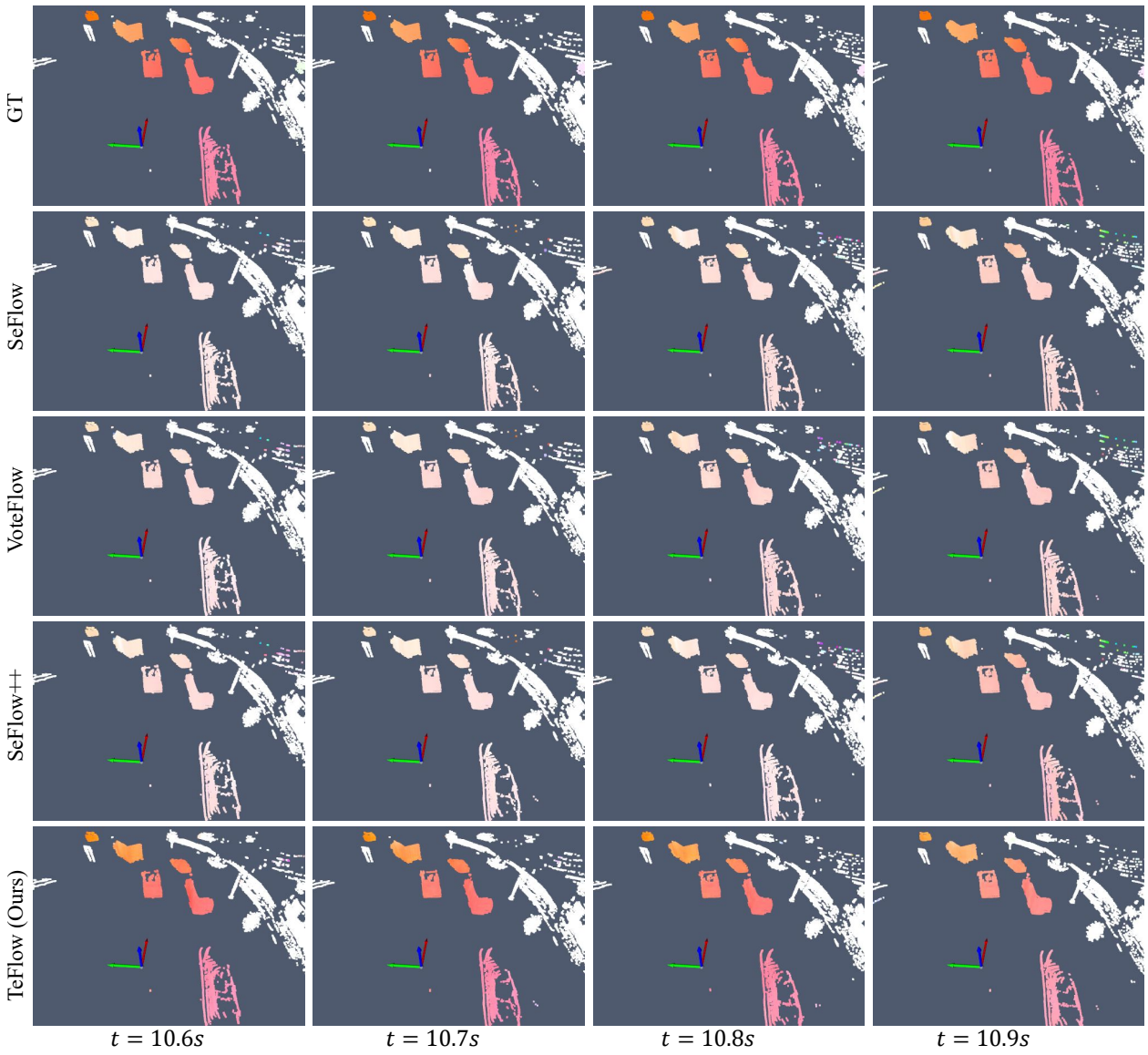

Figure 8. Qualitative results on the Argoverse 2 validation set. Our method accurately captures the motion of vehicles in complex roundabout scenarios. (Best viewed in color.) The scenes correspond to scene IDs 'bdb9d309-f14b-3ff6-ad1f-5d3f3f95a13e'.

along curved trajectories. The baseline methods fail to provide consistent estimates, often underestimating the motion or producing fragmented flows, especially for vehicles entering or exiting the roundabout. While, TeFlow produces coherent and smooth flow fields that closely follow the ground-truth directions, demonstrating its ability to handle complex multi-agent interactions in curved motion scenarios.