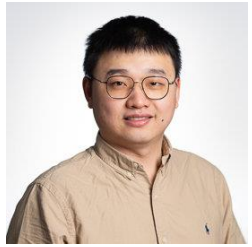
VoteFlow: Enforcing Local Rigidity in Self-Supervised Scene Flow

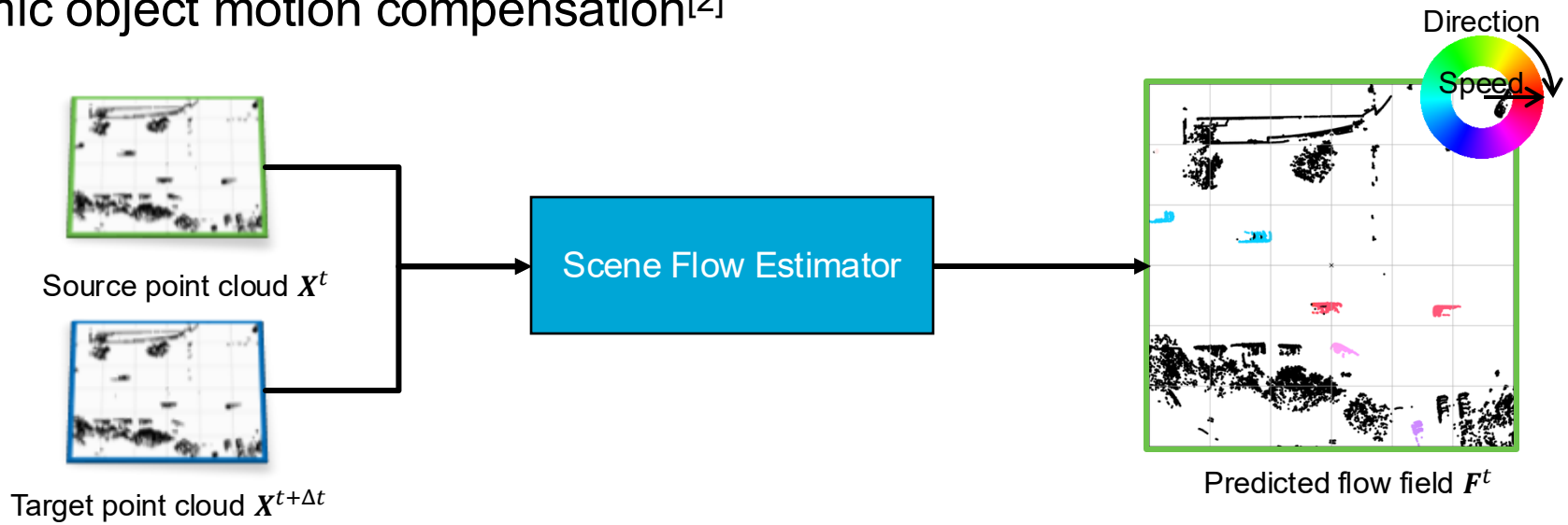




Yancong Lin ❖,†,✶, Shiming Wang ❖,✶, Liangliang Nan ❖, Julian Kooij ❖ and Holger Caesar ❖

❖ TU Delft † ETH Zurich ✶ Equal Contribution

Introduction: Scene Flow

- **Input:** two consecutive LiDAR scans $\boldsymbol{X}^t$ and $\boldsymbol{X}^{t+\Delta t}$
- **Output:** a point-wise 3D motion field from t to $t + \Delta t$ $\boldsymbol{F}^t$
- **Application:**
  - Unsupervised object discovery[1]
  - Dynamic object motion compensation[2]

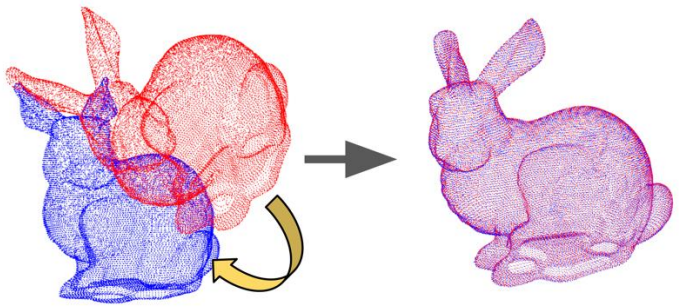


[1] Lentsch et al, UNION, NeurIPS'24; [2] Zhang et al, HiMO

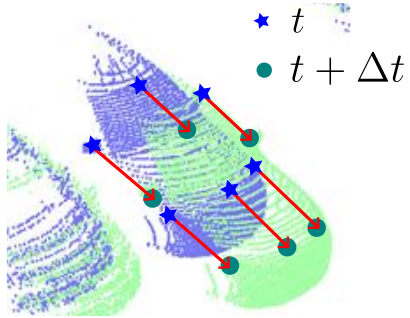
Introduction: Rigid Motion

- Scene flow estimation methods commonly adopt the **Rigid Motion** assumption.
- **Rigid Motion** indicates that nearby points on rigid objects share the same motion
- SoTA methods enforces this assumption via:
  - Extra loss function or regularizer [1,2]
  - Post-processing [3]

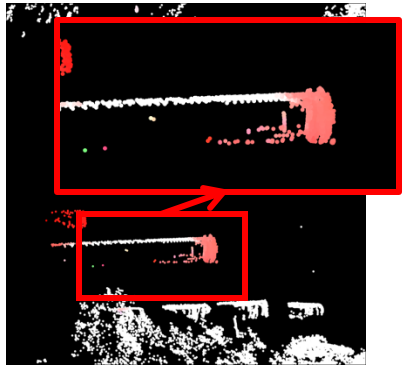
$\Longrightarrow$ However, none of them encode motion rigidity ***by design***



Motion in a rigid object



Rigid motion in scene flow

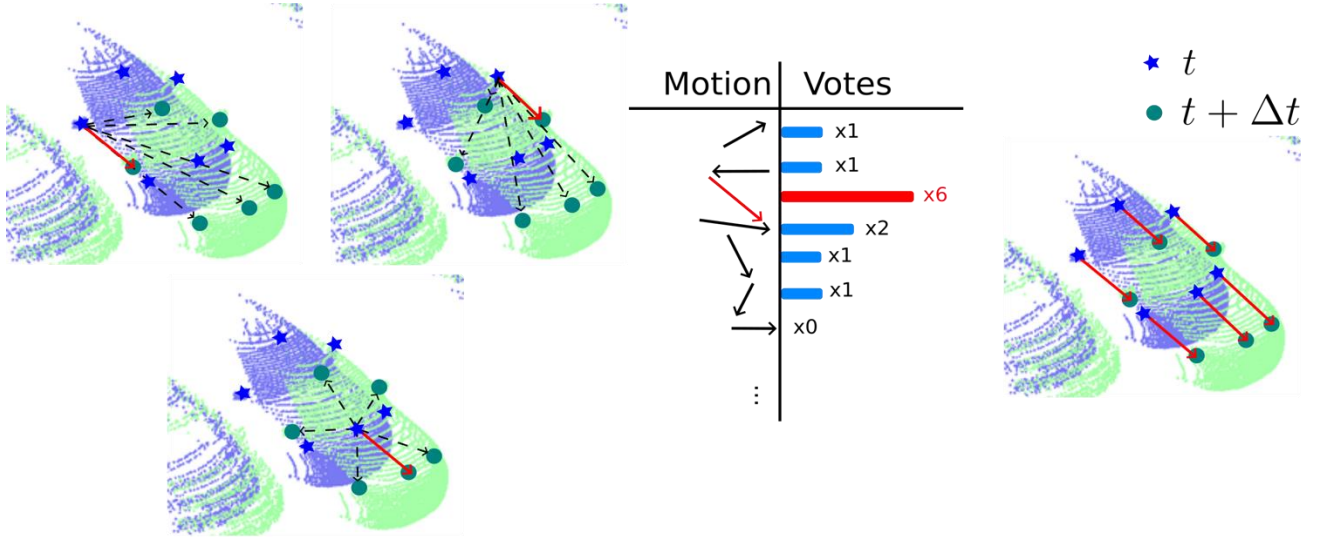


Lack of motion rigidity in model design results in inconsistent flow

[1] Zhang et al, Seflow, ECCV'24; [2] Lin et al, ICP-Flow, CVPR'24; [3] Chodosh et al, Re-Evaluating LiDAR Scene Flow for Autonomous Driving, WACV'24

Introduction: Our Idea

- Our idea: Use **voting** to identify shared rigid motion

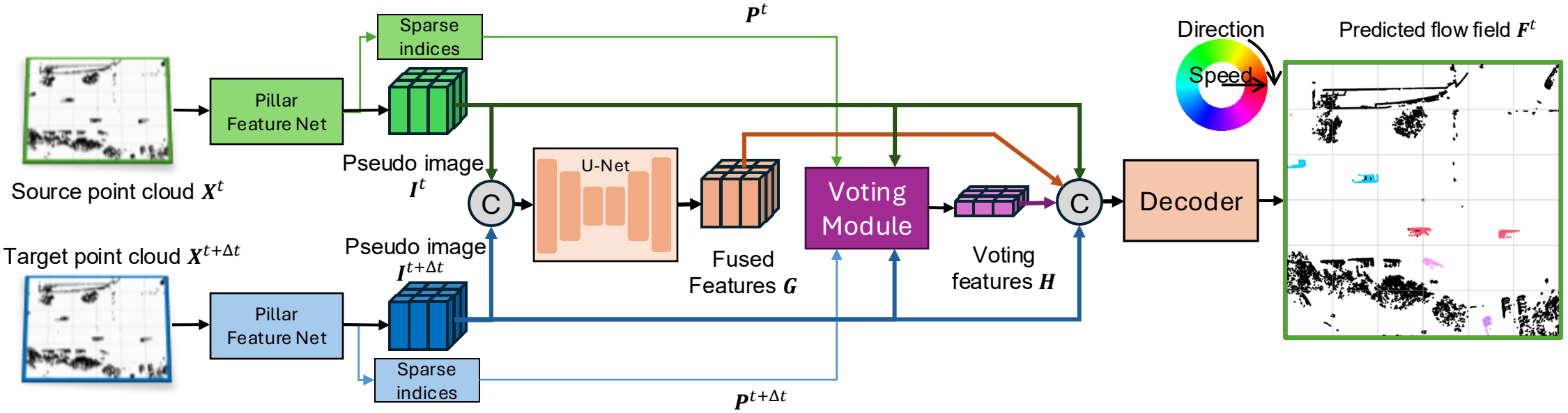


(a) Translation calculation (b) Voting (c) Scene flow

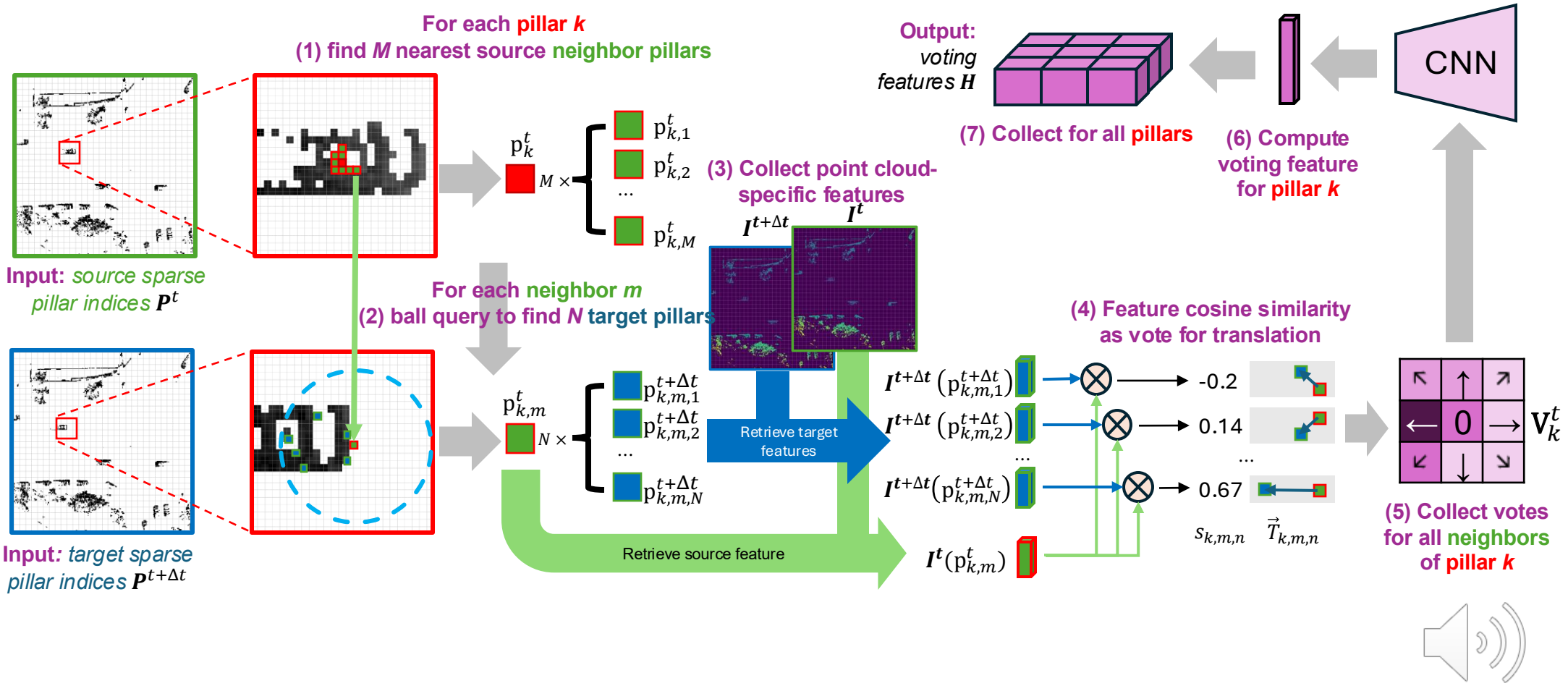
Introduction: Main Contributions

- **Motion rigidity**
  - ***Local rigid motion as an inductive bias*** in the self-supervised model
- **Plug & Play design**
  - Core component: a differentiable, light-weight and add-on **Voting Module**
- **Leading results, better generalization, and fast inference**
  - **In-domain evaluation:** outperforms on the Argoverse 2 dataset
  - **Cross-domain evaluation:** excels on the Waymo Open dataset
  - **Inference latency:** 40FPS (real-time)

Methodology: VoteFlow



Methodology: Voting Module



Experiments: Main Results

Method	Labels	Argoverse2 test					Waymo val (in-domain)			Waymo val (cross-domain)		
		Dynamic ↓ (normalized EPE)					EPE ↓ (in meters)			EPE ↓ (in meters)		
		Mean	Car	O.V.	Pd	W.V.	FD	FS	BS	FD	FS	BS
Flow4D [1]	✓	0.174	0.096	0.167	0.278	0.155	-	-	-	-	-	-
NSFP [2]	✗	0.422	0.251	0.331	0.723	0.383	0.171	0.108	0.022	-	-	-
SeFlow [3]	✗	0.309	0.214	0.292	0.463	0.267	0.151	0.018	**0.011**	0.155	0.018	0.013
VoteFlow	✗	**0.289**	**0.202**	**0.288**	**0.417**	**0.249**	**0.117**	**0.015**	0.016	**0.142**	**0.014**	**0.012**

- Flow4D is the state-of-the-art supervised method, showing the upper bound here
- O.V.: Other Vehicles, Pd: Pedestrians, W.V.: Wheeled VRUs
- FD: foreground dynamic, FS: foreground static, BS: background static
- In-domain: trained and valuated with Waymo, Cross-domain: trained on Argoverse 2 but tested on Waymo

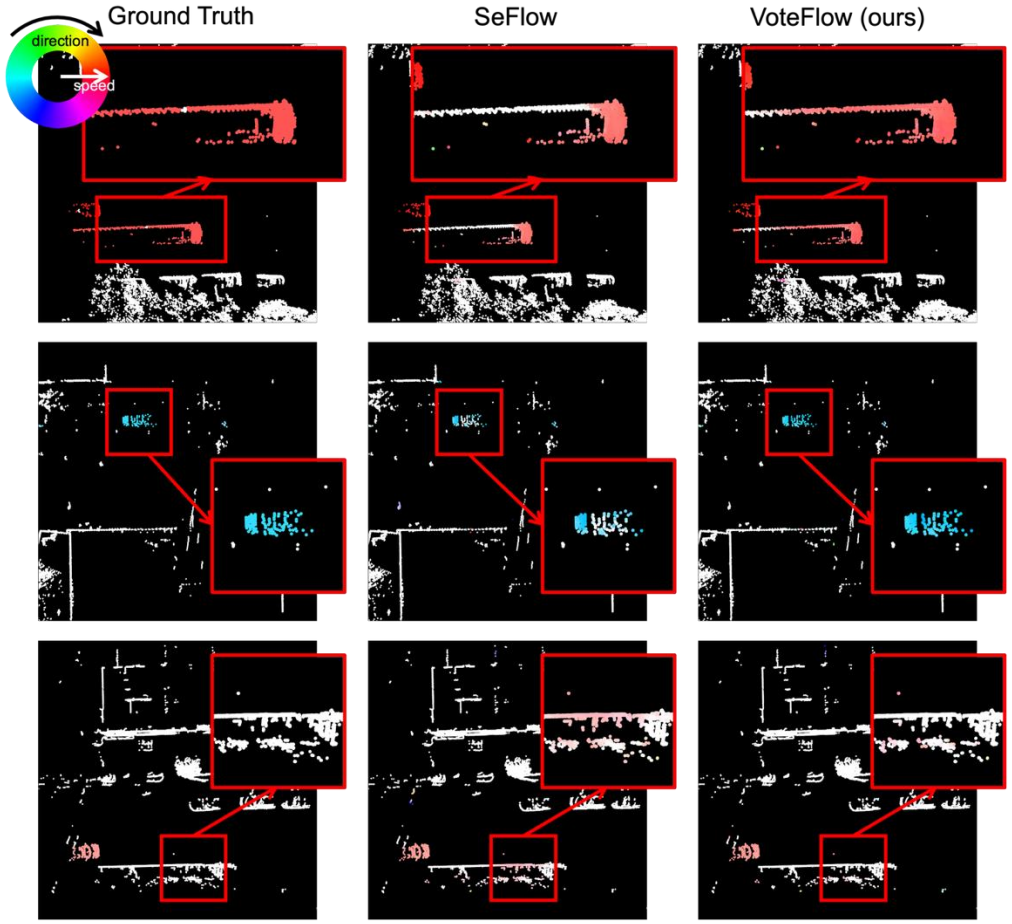
[1] Kim et al, Flow4D, RA-L, 2025; [2] Li et al, NSFP, NeurIPS, 2021; [3] Zhang et al, Seflow, ECCV, 2024.

Experiments: Plug & Play design

Method	Labels	Argoverse2 val				
		Dynamic ↓ (normalized EPE)				
		Mean	Car	O.V.	Pd	W.V.
FastFlow3D[1]	✓	0.487	0.268	0.351	0.812	0.517
FastFlow3D+Voting	✓	**0.389**	**0.164**	**0.292**	**0.674**	**0.429**
SeFlow [2]	✗	0.371	0.221	0.385	0.527	0.352
SeFlow+Voting	✗	**0.354**	**0.221**	**0.374**	**0.475**	**0.344**

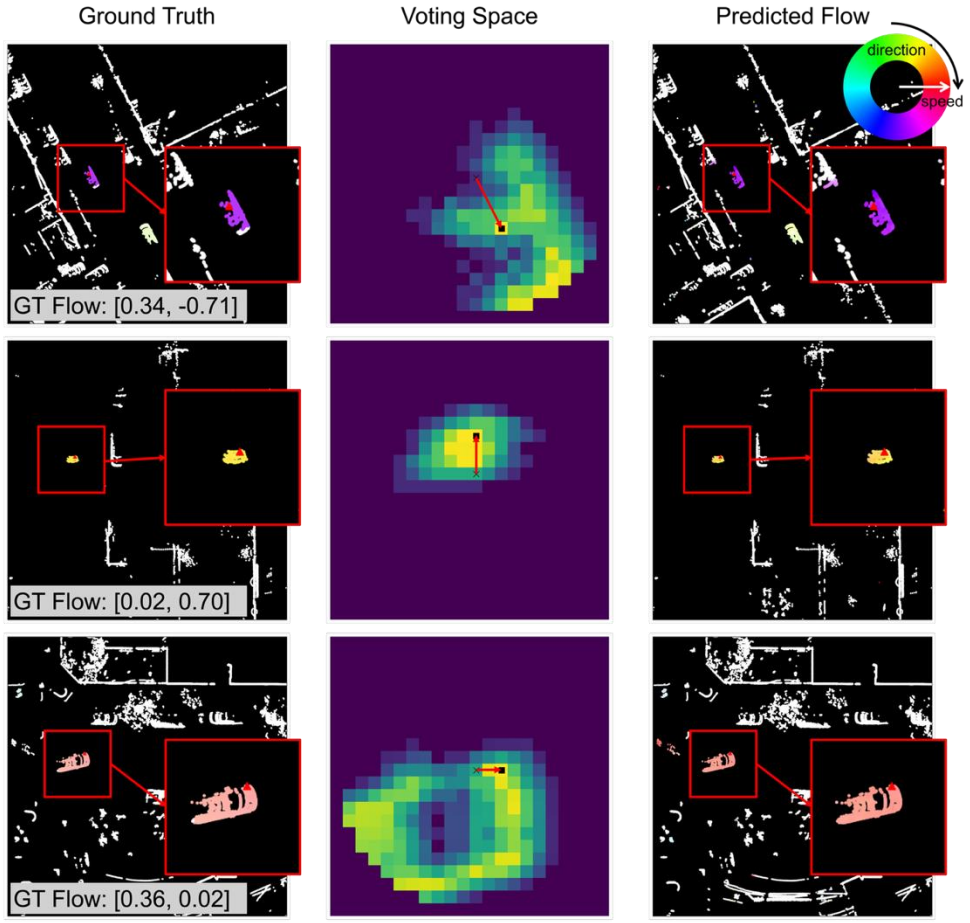
[1] Jund et al, FastFlow3D, RA-L, 2021; [2] Zhang et al, Seflow, ECCV, 2024.

Experiments: Qualitative Results



- VoteFlow generates more consistent predictions over the entire bus.
- VoteFlow achieves more uniform coloring overall, indicating the local motion rigidity was well captured.
- VoteFlow predicts less false positives and remains largely consistent with the ground truth.

Experiments: The Voting Space



- The middle column shows the voting space for an anchor pillar (▲ in the ground truth plot).
- We show an arrow (→) that points from the plot center (x) to the *argmax* bin (■), indicating the translation vector voted by the voting module.
- The center (x) indicates zero translations and boundaries indicate minimal and maximal translations along both dimensions.
- As shown in the voting space, → aligns with the ground truth flow.

Conclusions

- **Main takeaways:**
  - Motion rigidity, as an inductive bias, is crucial for scene flow estimation.
  - Our differential Voting Module Integrates motion rigidity into network design improving the performance and generalizability of the model.
- **Limitations:**
  - When Δt increases, the maximum possible translation grows, expanding the voting space and increasing the computational burden.
- **Future work:**
  - Long-horizon flow prediction
  - Multi-modal scene flow

TU Delft ETH zürich

CVPR Nashville JUNE 11-15, 2025

Thank you for watching!

Arxiv

Code

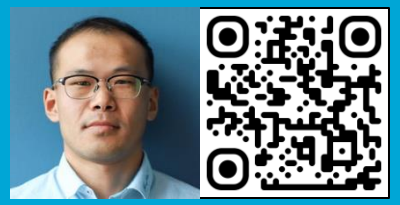
IV Group

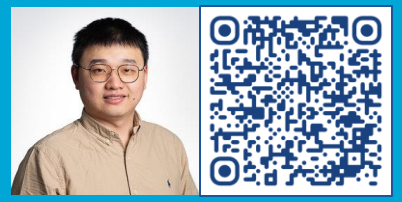
11-6-2025

TU Delft ETH zürich

CVPR Nashville JUNE 11-15, 2025

Thank you for watching.

Yancong Lin 

Shiming Wang 

Liangliang Nan 

Julian Kooij 

Holger Caesar