# SeFlow Supplementary Material

Qingwen Zhang[1], Yi Yang[1,2], Peizheng Li[3,4], Olov Andersson[1], and Patric Jensfelt[1]

[1] RPL, KTH Royal Institute of Technology, Stockholm, Sweden
[2] Scania CV AB, Södertälje, Sweden
[3] University of Tübingen, Tübingen, Germany
[4] Mercedes-Benz AG, Sindelfingen, Germany
{qingwen,yiya,olovand,patric}@kth.se, peizheng.li@mercedes-benz.com

In this supplementary material, we begin by detailing the implementation aspects, including datasets and hyperparameters for dynamic classification, clustering, and training, as outlined in Sec. 1.1 and Sec. 1.2. Following this, we present additional results in several key areas:

- Section 1.3 (Loss Terms): This section explores various applications of static and dynamic classification in designing loss functions. To validate the effectiveness of our proposed upper bound flow in the object cluster, we also incorporate common strategies such as averaging or maximizing the estimated flows within clusters. Furthermore, an additional ablation study table is provided to further elucidate the impact of our loss terms.
- Section 1.4 (Different Model Backbones): This section demonstrates that SeFlow's effectiveness is not limited to a specific model backbone. We show that also with the same backbone as FastFlow3D, SeFlow outperforms both self-supervised and supervised baselines, underscoring the strength of our self-supervised pipeline.
- Section 2 (Qualitative Results): In addition to the sequence scenes discussed qualitatively in the main paper, we present two more qualitative results showcasing SeFlow's performance on both Argoverse 2 and Waymo datasets, including some failure cases.

## 1 Appendix A - Experiment

### 1.1 Implementation Details

**Our Method** The resolution of network voxelization is set as 0.2 m, consquently, the $[512, 512]$ grid corresponds to a 102.4 m $\times$ 102.4 m map. DUFOMap [2] is used for the dynamic classification in our method and its resolution is consistent with the voxelization setting of 0.2 m. For DUFOMap's parameters $d_p$ and $d_s$, we keep them as default which is 1 and 0.2 respectively. HDBSCAN only clusters dynamic points inside $\mathcal{P}_{t,d}$ where the minimum cluster size is set to 20 and cluster selection $\epsilon$ is set to 0.7. Our SeFlow is trained for 50 epochs with a batch size of 80 without any ground truth labels. We employ Adam optimizer with a $2 \times 10^{-6}$ learning rate. The code is open-sourced at `https://github.com/KTH-RPL/SeFlow`.

**Other Methods** In our main comparison on the Argoverse 2 test set (as shown in Table 1 of our main paper), we directly reference results from the online leaderboard [1] and Table I in DeFlow [9], with result files available in their discussion thread[5]. For the Waymo validation set results (Table 2 in our main paper), we present the outcomes for FastFlow3D [3], ZeroFlow [7], and NSFP [4] as reported in Table 2 of ZeroFlow [7], which were trained on the Waymo train set. FastNSF [5] results were obtained by running their model[6] with default Waymo settings. For DeFlow [9] and our method SeFlow, we conducted our own training on the Waymo training set, adhering to the same training strategy outlined earlier. Regarding all ZeroFlow entries in our ablation study, we utilized the pre-trained weights available in their official repository[7].

**Setup** For inference, to measure the complexity and computational cost of our model and other methods, all experiments are executed on a desktop powered by an Intel Core i9-12900KF CPU and equipped with a GeForce RTX 3090 GPU.

**Process Time** Regarding the processing time on labeling DUFOMap and HDBSCAN steps, taking the Argoverse 2 dataset as an example, the average runtimes of DUFOMap and HDBSCAN are 50ms/frame and 500ms/frame respectively. It take approximately 16.35 CPU hours for the whole dataset in the setup we mentioned above. The DUFOMap and HDBSCAN steps are pre-processed before training to avoid redundant computations. SeFlow's training time compared to DeFlow is 30 hours vs 21 hours on 8 A100 GPUs.

### 1.2 Datasets

In this section, we present the datasets we use. For the convenience of the reader and to make these presentations self-contained there are some repetitions from the main paper.

**Argoverse 2 [8]** It contains two subdataset - *Sensor* and *Lidar*. The *Sensor* dataset encompasses 700 training and 150 validation scenes. Each scene is approximately 15 seconds long in 10 Hz, complete with annotations for evaluation. The *LiDAR* dataset lacks imagery and any other annotations containing 16,000 training, 2,000 validation, and 2,000 test scenes, respectively. Each scene is approximately 30 seconds long in 10 Hz. The *LiDAR* dataset is designed to support research into self-supervised learning in the lidar domain, as well as point cloud forecasting. All of the above datasets are collected using two 32-channel LiDARs. The average number of points in one frame is around 52,871 after ground removal. The 200% data (214k frames in total) in the main paper means 100%

[5] `https://github.com/KTH-RPL/DeFlow/discussions/2`
[6] `https://github.com/Lilac-Lee/FastNSF`
[7] `https://github.com/kylevedder/zeroflow_weights`

*Sensor* dataset which contains 107k frames plus another 107k frames from *LiDAR* dataset, selected via the same process as in [7]. [7] uniformly sampled 12 pairs of frames from each scene of the *LiDAR* dataset first, followed by random sampling to get another 107k frames, i.e., another 100% of data.

**Waymo Open Dataset [6]** The dataset contains 798 training and 202 validation sequences. Each sequence contains 20 seconds of 10Hz point clouds collected using a custom LiDAR mounted on the roof of a car. The total number of training frames is 155k. The average number of points in one frame is around 79,327 after ground removal. Since it does not have a public leaderboard or official evaluation scripts, in this paper, we follow the same setting and process steps as ZeroFlow [7] to make fair comparisons. The evaluation follows the same Argoverse 2 official evaluation scripts and evaluates flow performance on points that do not belong to the ground and are within a 100m $\times$ 100m range centered on the origin.

### 1.3 Additional Ablation Studies in Loss Terms

**Dynamic Chamfer and Static Loss** We investigate different alternatives for the design of dynamic Chamfer and static losses. In addition to the standard Chamfer loss $\mathcal{L}_{\text{cham}}$, both FastFlow3D [3] and DeFlow [9] propose different losses and weights for static and dynamic points based on ground truth classification labels. In the comparison, we reformat the dynamic weight loss formulas of these two methods, making use of the classification results:

$$[3] : \mathcal{L}_{d,s} = \frac{1}{|\mathcal{P}_t|} \sum_{p \in \mathcal{P}_t} \sigma(p)\mathrm{S}(p), \text{ where } \sigma(p) = \begin{cases} 0.9 & \text{if } p \in \mathcal{P}_{t,d} \\ 0.1 & \text{if } p \in \mathcal{P}_{t,s} \end{cases}. \tag{1}$$

$$[9] : \mathcal{L}_{d,s} = \frac{1}{|\mathcal{P}_{t,d}|} \sum_{p \in \mathcal{P}_{t,d}} \mathrm{S}(p) + \frac{1}{|\mathcal{P}_{t,s}|} \sum_{p \in \mathcal{P}_{t,s}} \mathrm{S}(p). \tag{2}$$

In the above formulas, $\mathrm{S}(\cdot) = \mathrm{D}(\cdot)^2$, and $\mathrm{D}(p, \mathcal{P}_{t+1})$ denotes the distance between point $p$ and its nearest neighbor in $\mathcal{P}_{t+1}$. Since we do not use any labels, the dynamic points $\mathcal{P}_{t,d}$ and static points $\mathcal{P}_{t,s}$ in Eq. (1) (FastFlow3D strategy) and Eq. (2) (DeFlow strategy) are obtained from the dynamic classification results. As a comparison, our dynamic and static losses can be represented as:

$$\mathcal{L}_{d,s} = \mathcal{L}_{\text{dcham}} + \mathcal{L}_{\text{static}}. \tag{3}$$

Table 1 shows that under the self-supervised strategy, our static and dynamic loss design with $\mathcal{L}_{\text{dcham}} + \mathcal{L}_{\text{static}}$ is the best solution according to EPE 3-way. Looking at the individual EPE components, EPE FD is similar to the three strategies and the main difference is in the static EPE components (FS and BS) where our strategy results in significantly lower errors. We attribute this to targeted loss selection rather than just loss weight balancing.

**Table 1:** Ablation study on different static dynamic usage in loss design. Our design is $\mathcal{L}_{\text{cham}} + \mathcal{L}_{\text{dcham}} + \mathcal{L}_{\text{static}}$, while others are $\mathcal{L}_{\text{cham}} + \mathcal{L}_{d,s}$.

| Stategy | EPE | | | |
|---|---|---|---|---|
| | 3-way | FD | FS | BS |
| Eq. (1) [3] | 0.094 | **0.192** | 0.057 | 0.034 |
| Eq. (2) [9] | 0.099 | 0.211 | 0.053 | 0.033 |
| Eq. (3) Ours | **0.078** | 0.220 | **0.012** | **0.002** |

**Cluster Loss** To show the superiority of our cluster loss design, we experimented with different designs to determine $\tilde{f}_{c_i}$. Table 2 presents a comparison of different cluster flow loss configurations under the following definitions:

$$\text{avg} : \tilde{f}_{c_i} = \frac{1}{|\mathcal{P}_{c_i}|} \sum_{p \in \mathcal{P}_{c_i}} \hat{\mathcal{F}}(p). \tag{4}$$

$$\max : \tilde{f}_{c_i} = \max_{p \in \mathcal{P}_{c_i}} \hat{\mathcal{F}}(p). \tag{5}$$

$$\text{Ours} : \tilde{f}_{c_i} = p'_\kappa - p_\kappa, \text{ where } \kappa = \arg\max\{\mathrm{D}(p_k, \mathcal{P}_{t+1,d}) | p_k \in \mathcal{P}_{c_i}\}. \tag{6}$$

In the above formulas, $\hat{\mathcal{F}}(p)$ represents the estimated flow of point $p$ and $p'_\kappa$ is the nearest neighbor of $p_\kappa$ in $\mathcal{P}_{t+1,d}$. We explored the average of the estimated flow (Eq. (4)), the maximum from the estimated flow (Eq. (5)), and our proposed method as detailed in Eq. (6).

**Table 2:** Ablation study on different cluster flow consistency. All variations utilize four losses, and the only difference is the choice of $\tilde{f}_{c_i}$.

| $\tilde{f}_{c_i}$ | EPE | | | |
|---|---|---|---|---|
| | 3-way | FD | FS | BS |
| Eq. (4) avg | 0.078 | 0.221 | **0.012** | 0.002 |
| Eq. (5) max | 0.092 | 0.262 | 0.013 | **0.001** |
| Eq. (6) Ours | **0.064** | **0.160** | 0.029 | 0.004 |

The results in Tab. 2 demonstrate that our method decreases the EPE of foreground dynamics the most among the three definitions, which contributes significantly to the reduction of 3-way EPE. Compared to the huge improvement in the foreground dynamic (FD) estimation, the resulting fluctuation in the flow estimation of the static points (FS and BS) is minor.

**Different Loss Combinations** In this section, as detailed in Tab. 3, we present additional ablation studies where we deactivate one of the four losses to analyze

the impact of each loss's absence. Experiment A2 demonstrates that our model, even without $\mathcal{L}_{\text{cham}}$, achieves results comparable to using all loss terms (A1) as suggested in our paper. Our dynamic and static losses can replace the general chamfer distance loss to a large extent, but the overall scene-level consideration is still beneficial. Our three proposed losses, which are based on dynamic classification and divided into static, dynamic, and object-level aspects, still effectively reduce the EPE 3-way when combined with the Chamfer distance as a foundational constraint.

**Table 3:** Ablation study in different loss combinations. Results are evaluated on the Argoverse 2 validation set with 20 training epochs.

| Exp. Id | $\mathcal{L}_{\text{cham}}$ | $\mathcal{L}_{\text{dcham}}$ | $\mathcal{L}_{\text{static}}$ | $\mathcal{L}_{\text{dcls}}$ | EPE ↓ | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | 3-way | FD | FS | BS |
| A1 | ✓ | ✓ | ✓ | ✓ | **0.0643** | 0.160 | 0.029 | 0.004 |
| A2 | | ✓ | ✓ | ✓ | 0.0651 | 0.162 | 0.030 | 0.003 |
| A3 | ✓ | | ✓ | ✓ | 0.0717 | 0.175 | 0.037 | 0.003 |
| A4 | ✓ | ✓ | | ✓ | 0.0890 | **0.150** | 0.077 | 0.040 |
| A5 | ✓ | ✓ | ✓ | | 0.0779 | 0.220 | **0.012** | **0.002** |

Comparing experiments A3 and A5 with A1 in Tab. 3, it's evident that omitting $\mathcal{L}_{\text{dcham}}$ or $\mathcal{L}_{\text{dcls}}$ leads to a decline in dynamic flow estimation accuracy (FD). This highlights the significance of both dynamic and object-level self-supervised losses in assisting networks to understand object motion patterns. Notably, $\mathcal{L}_{\text{dcls}}$ (A5) has a more substantial impact than $\mathcal{L}_{\text{dcham}}$ (A3). A similar trend is observed for static aspects; comparing A4 with A1 reveals that the absence of $\mathcal{L}_{\text{static}}$ results in increased errors in both EPE FS and EPE BS, underscoring its importance in static error reduction.

**Table 4:** Ablation study in difference model backbones, where FF and DF mean different model backbones from supervised methods FastFlow3D [3] and DeFlow [9], respectively. We **bold** the best results and underline the second best results.

| BackBone | EPE | | | |
|---|---|---|---|---|
| | 3-way | FD | FS | BS |
| FastFlow3D | 0.081 | 0.222 | 0.020 | 0.002 |
| ZeroFlow (FF) | 0.088 | 0.231 | 0.022 | 0.011 |
| Ours (FF) | 0.065 | 0.164 | 0.028 | 0.002 |
| Ours (DF) | **0.059** | 0.147 | 0.026 | 0.004 |

### 1.4 Ablation Study in Difference Model Backbones

In this section, we examine the effects of varying the model backbone on performance. We replaced the DeFlow backbone with the FastFlow3D backbone, aligning our model structure with that of ZeroFlow and FastFlow3D. The results, presented in Tab. 4, show that even with the same backbone (Ours (FF)), our method still surpasses both ZeroFlow (ZF) and FastFlow3D (FF). This outcome underscores that the strength of our approach lies not in a specific model backbone.

## 2 Appendix B - Qualitative Results

In this section, we present additional qualitative results from the Argoverse 2 and Waymo validation datasets, including some failure cases. In each figure, unless otherwise specified, different colors represent different motion directions, and more saturated colors indicate larger flow estimations. The qualitative results in the main paper are derived from the scene 'b5a7ff7e-d74a-3be6-b95d-3fc0042215f6' in the Argoverse 2 validation set. Here, we include two more scenes for further illustration from the Waymo and Argoverse 2 validation set.

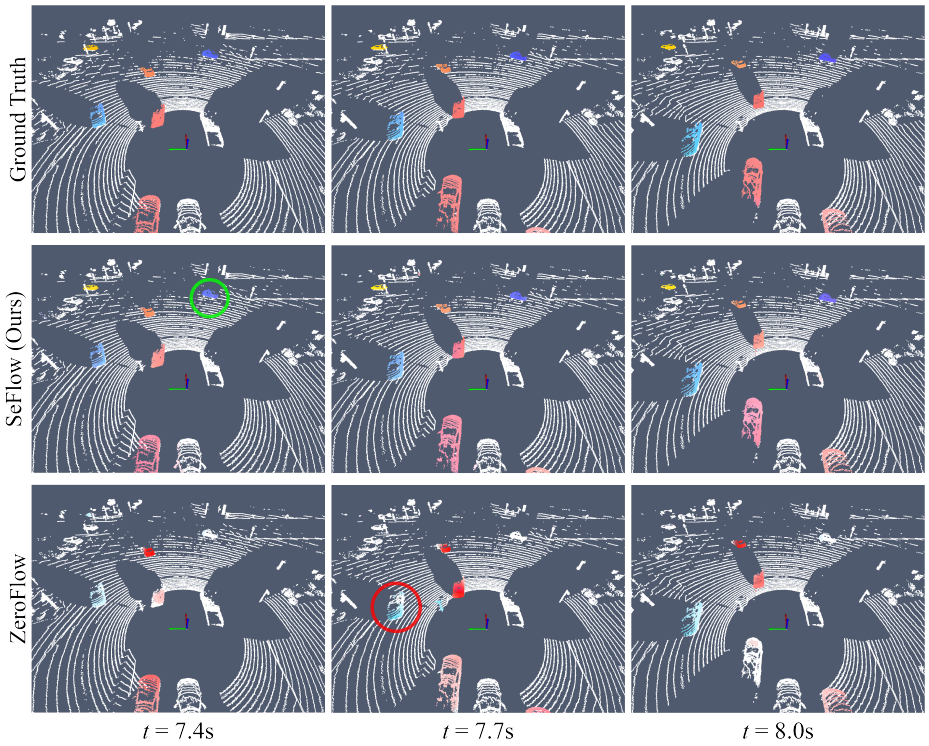



**Fig. 1:** Qualitative results from Waymo validation set (scene id '14081240615915270380_4399_000_4419_000'). The top row displays the ground truth flow, the middle row presents the SeFlow result, and the bottom row showcases the result of another self-supervised method ZeroFlow.

In terms of flow estimation accuracy, our SeFlow method demonstrates superior performance compared to ZeroFlow in the Waymo dataset, as depicted in Fig. 1. The flows estimated by our method closely align with the ground truth in both direction and magnitude, whereas there are flows from vehicles or parts of vehicles that are ignored in the ZeroFlow results. In Fig. 2, the ZeroFlow results exhibit noticeable issues with no flow estimation in small-scale objects like

pedestrians. In contrast, our SeFlow method maintains consistent and accurate flow estimation throughout the scene.

This additional qualitative analysis further validates the effectiveness of SeFlow in accurately capturing scene dynamics across diverse scenarios.

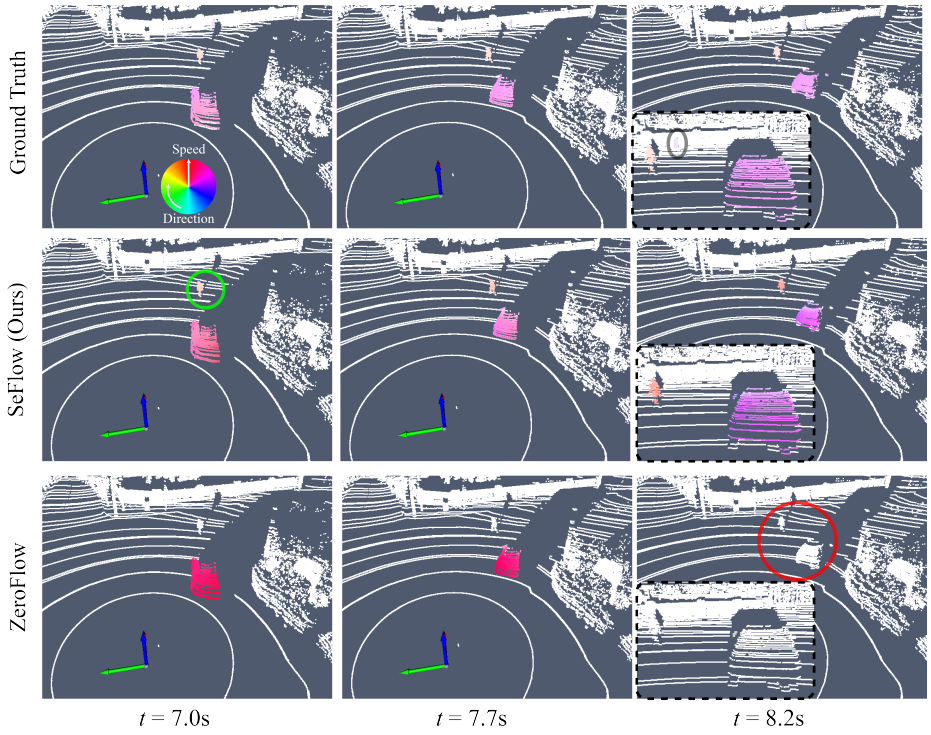
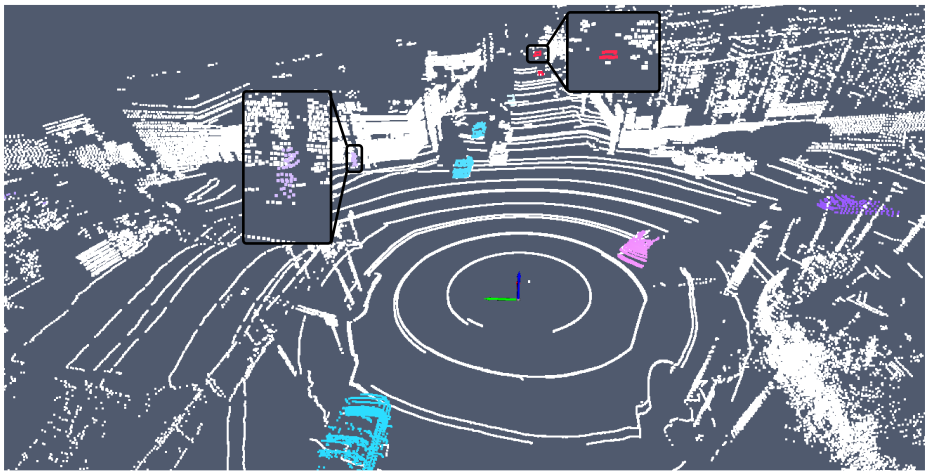
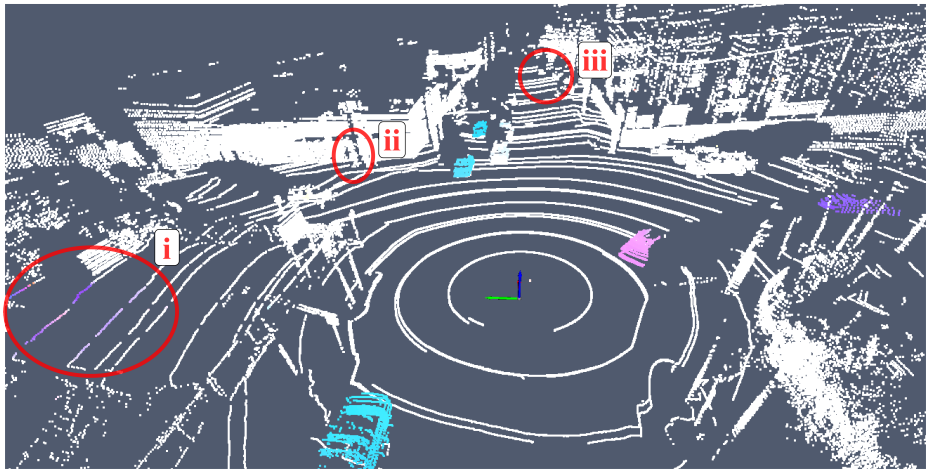



**Fig. 2:** Qualitative results from Argoverse 2 validation set (scene id '77574006-881f-3bc8-bbb6-81d79cf02d83'). Different colors represent different motion directions, and more saturated colors indicate larger flow estimations. The top row displays the ground truth flow, the middle row presents the SeFlow result, and the bottom row showcases the result of another self-supervised method ZeroFlow. The bottom right of the third column is the zoom-in view at the moment.

**Failure Cases** As illustrated in Fig. 3, our method also has a few deficiencies that need to be improved. One notable issue is the presence of false positive flow estimations, particularly when ground points are not completely removed (Fig. 3.b.i). Additionally, predicting the flow of pedestrians near static structures poses a challenge (Fig. 3.b.ii). Furthermore, accurately predicting the motion of distant objects proves difficult when relying solely on two consecutive point cloud inputs (Fig. 3.b.iii). These limitations highlight specific challenges in scene flow estimation and underscore the need for further refinement and development of our approach.



(a) Ground Truth Flow




(b) SeFlow Output

**Fig. 3:** Qualitative analysis of failure cases in SeFlow on Argoverse 2 validation set (scene id '22052525-4f85-3fe8-9d7d-000a9fffce36'). (a) displays the ground truth flow where black boxes are the zoom-in views. (b) presents the SeFlow result where the red circle means limitations in our estimation.

## References

1. 2, A.: Argoverse 2 scene flow online leaderboard. `https://eval.ai/web/challenges/challenge-page/2010/leaderboard/4759` (2024 Mar 4th)
2. Duberg, D., Zhang, Q., Jia, M., Jensfelt, P.: DUFOMap: Efficient dynamic awareness mapping. IEEE Robotics and Automation Letters **9**(6), 5038–5045 (2024). `https://doi.org/10.1109/LRA.2024.3387658`

3. Jund, P., Sweeney, C., Abdo, N., Chen, Z., Shlens, J.: Scalable scene flow from point clouds in the real world. IEEE Robotics and Automation Letters **7**(2), 1589–1596 (2021)
4. Li, X., Kaesemodel Pontes, J., Lucey, S.: Neural scene flow prior. Advances in Neural Information Processing Systems **34**, 7838–7851 (2021)
5. Li, X., Zheng, J., Ferroni, F., Pontes, J.K., Lucey, S.: Fast neural scene flow. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9878–9890 (2023)
6. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al.: Scalability in perception for autonomous driving: Waymo open dataset. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2446–2454 (2020)
7. Vedder, K., Peri, N., Chodosh, N., Khatri, I., Eaton, E., Jayaraman, D., Ramanan, Y.L.D., Hays, J.: ZeroFlow: Fast Zero Label Scene Flow via Distillation. International Conference on Learning Representations (ICLR) (2024)
8. Wilson, B., Qi, W., Agarwal, T., Lambert, J., Singh, J., et al.: Argoverse 2: Next generation datasets for self-driving perception and forecasting. In: Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS Datasets and Benchmarks 2021) (2021)
9. Zhang, Q., Yang, Y., Fang, H., Geng, R., Jensfelt, P.: Deflow: Decoder of scene flow network in autonomous driving. arXiv preprint arXiv:2401.16122 (2024)