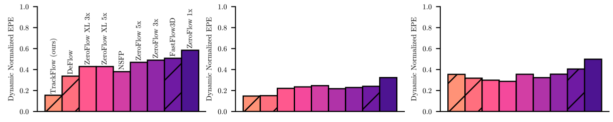
I Can't Believe It's Not Scene Flow!

Summary: Reviewers find our observations about small objects being overlooked by current scene flow metrics and methods to be: interesting (R1), sensible (R3), rarely studied before (R2), and valuable to the community (R3). Our simple baseline (TrackFlow) is effective, achieving state of the art results, and highlights issues in current methods (R1, R2, and R3). We address specific feedback below.

R1: Time cost of TrackFlow. Both the detector and tracker can run in real-time[1].

R1: Does TrackFlow depend on detector performance? Yes, see Section 5.3 for a discussion on what makes a detector better for our detect + track framework.

R2: Reliance on semantic classes prevents open-set evaluation. Thank you for highlighting this concern, as we think addressing it makes our paper stronger. Our paper used AV2's semantic taxonomy as a standard way to break down the object distribution into meaningful subsets; however, semantics are not *required* — any meaningful slicing of the distribution is sufficient. We use our evaluation with open-set flow labels (ground truth object motion is available, but there are no object semantic labels) and generate meaningful insights. We cluster AV2's ground-truth bounding boxes by *volume*[2], not semantic class, and re-run our evaluation protocol. The resulting evaluation still reveals that existing methods fail on small objects (Figures 1 & 2), something which is hidden by existing evaluations. We will include these results and a more detailed discussion on open set evaluation in our camera ready.



(a) SMALL (b) MEDIUM (c) LARGE

Figure 1. Per cluster Dynamic Normalized EPE. Methods are colored and ordered the same as in Figure 2.

R2: Comparison against unsupervised methods is unfair. We compare TrackFlow against both supervised (shown with hatching) and unsupervised methods (Figures 4-6) that run on *full point clouds*[3]. Both DeFlow, our strongest baseline, and FastFlow3D are supervised, making them directly comparable to TrackFlow. As discussed in Lines 22-41, although existing scene flow methods can generalize to the tail of object distributions *in theory*, our paper shows that *in practice* they are (shockingly) unable to even generalize to the head. Despite it's crude construction, TrackFlow's superior performance demonstrates the power of distribution awareness — it generalizes to the head far better than prior work because of the data augmentation used by 3D detectors — illuminating a fruitful path for future work.

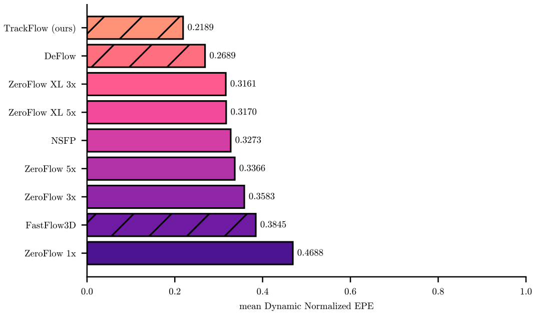


Figure 2. mean Dynamic Normalized EPE across open-set clustering of SMALL, MEDIUM, LARGE ground truth boxes.

R3: The proposed evaluation metric and method lack innovation. Although our metric and method are *post-hoc* obvious, reviewers agree (R1, R2, R3) that the community will benefit from our analysis. The simplicity of our proposed protocols and SotA baseline is our strength.

R3: TrackFlow is similar to 3D SORT. Yes, TrackFlow is a multi-object tracking method *applied to scene flow*. Our novelty comes from *applying it to scene flow* (and it's *SotA*).

R3: The evaluation seems tailored specifically for driving scenes. Figures 1 & 2 show our evaluation does not fundamentally require classes to extract useful insights and can generalize to any ground truth 3D objects labels.

R3: The proposed metrics and method seem difficult to apply to datasets such as FlyingThings3D. Our volume-based clustering can be applied to any dataset, including FT3D, and any object detector trained on FT3D can be used in our detect-and-track framework.

However, it should be noted that "[FlyingThings3D has] unrealistic rates of dynamic motion, unrealistic correspondences, and unrealistic sampling patterns. As a result, progress on these benchmarks is misleading and may cause researchers to focus on the wrong problems." [1]. Thus, our paper focuses on the (still unsolved!) problem of estimating scene flow for small, rare objects in real-world datasets.

References

[1] Chodosh, N., Ramanan, D., Lucey, S.: Re-Evaluating LiDAR Scene Flow for Autonomous Driving. arXiv preprint (2023) 1

[2] Liu, Z., Tang, H., Amini, A., Yang, X., Mao, H., Rus, D., Han, S.: Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In: IEEE International Conference on Robotics and Automation (ICRA) (2023) 1

[3] Vedder, K., Peri, N., Chodosh, N., Khatri, I., Eaton, E., Jayaraman, D., Liu, Y., Ramanan, D., Hays, J.: ZeroFlow: Scalable Scene Flow via Distillation. In: Twelfth International Conference on Learning Representations (ICLR) (2024) 1

[4] Weng, X., Wang, J., Held, D., Kitani, K.: 3D Multi-Object Tracking: A Baseline and New Evaluation Metrics. IROS (2020) 1

[1] For example, BEVFusion runs at 120ms (8fps) on an RTX3090, and AB3DMOT takes 5ms on a CPU allowing it to run at 200FPS. [2, 4]

[2] SMALL: $< 9.5m^3$, MEDIUM: $\geq 9.5m^3 \wedge < 40m^3$, LARGE: $\geq 40m^3$. These volume boundaries were determined via clustering.

[3] Many methods can't run on full point clouds; see [3] for details.

View Meta-Reviews

Paper ID

2757

Paper Title

I Can't Believe It's Not Scene Flow!

META-REVIEWER #1
META-REVIEW QUESTIONS

2. Paper final decision

Accept

4. Decision summary

Accept: The paper received mixed or borderline reviews. The area chairs considered the paper, reviews, and rebuttal, as well as further discussion, and decided to accept the paper. This decision has been confirmed by the AC panel. See comments below for details.

6. Comments on decision

The reviewers and AC's have gone through the submission and rebuttal, and participated in a subsequent discussion. All the reviewers are aware of the concerns. While not unanimous, the paper is seen as net-positive, and we appreciate that the authors have done the best that they could with data that predominantly focuses on driving. We were not clear what the authors meant about open-set clustering. We hope the authors will point this out (generalization to non-driving situations) as an area for further exploration.

View Reviews

Paper ID
2757
Paper Title
I Can't Believe It's Not Scene Flow!

Reviewer #1

Questions

2. Summary. In 5-7 sentences, describe the key ideas, experimental or theoretical results, and their significance.

    This paper addresses the problem of scene flow estimation. In particular, the paper first reveals that existing scene flow methods do not perform well on small objects, although existing scene flow evaluation metrics show these methods achieve high estimation accuracy on scene flow datasets. To address the limitation in existing scene flow evaluation metrics, a new scene flow evaluation metric named "Bucket Normalized EPE" is proposed, which takes classes and speed into account. Moreover, the paper proposes a scene flow method that is simple yet effective. The experimental results demonstrate the effectiveness of the proposed scene flow method.

3. Strengths. Consider the significance of key ideas, experimental or theoretical validation, writing quality, data contribution. Explain clearly why these aspects of the paper are valuable. Short bullet lists do NOT suffice.

    1. The paper shows the state-of-art methods that achieve high scene flow estimation accuracy do not perform well on small objects, which is interesting

    2. The paper proposes a new evaluation metric explicitly considering speed and classes, addressing the issue of existing scene flow evaluation metrics in assessing motion estimation accuracy for small objects.

    3. The paper proposes a novel method named TrackFlow which is simple yet effective. The experimental results show that the proposed TrackFlow outperforms state-of-art methods

4. Weaknesses. Consider the significance of key ideas, experimental or theoretical validation, writing quality, data contribution. Clearly explain why these are weak aspects of the paper, e.g., why a specific prior work has already demonstrated the key contributions, or why the experiments are insufficient to validate the claims, etc. Short bullet lists do NOT suffice. Be specific!

    1. The proposed method employs an object detection model and the tracing algorithm which increases the time cost. What is the time cost of the proposed? Does the proposed method consume much more time cost than the state-of-the-art methods?

2. The proposed TrackFlow heavily depends on the object detection model. Will the performance of the detection method significantly affect the scene flow results? Is the proposed method robust to the detection errors?

5. Paper rating (pre-rebuttal).

Weak Accept

7. Justification of rating. What are the most important factors in your rating?

The paper is interesting, and the experimental results show the effectiveness of the proposed method

8. Are there any serious ethical/privacy/transparency/fairness concerns? If yes, please also discuss below in Question 9.

No

10. Is the contribution of a new dataset a main claim for this paper? Have the authors indicated so in the submission form?

No dataset contribution claim

14. Final rating based on ALL the reviews, rebuttal, and discussion (post-rebuttal).

Borderline Accept

15. Final justification (post-rebuttal).

Thank the authors for addressing my questions. I agree with R2 that the proposed method relies on the detector and may not perform well on unseen categories. Considering novelty, my final rating is Borderline Accept

Reviewer #2

Questions

2. Summary. In 5-7 sentences, describe the key ideas, experimental or theoretical results, and their significance.

This paper studies scene flow estimation from point clouds. The authors observe that scene flow methods struggle with estimating the motion of small objects, and current metrics are inadequate for assessing the performance of scene flow methods on these objects. To address this issue, the authors propose a class-aware and speed-normalized evaluation metric for the task of scene flow estimation. Additionally, the authors also design a supervised scene flow model named TrackFlow, which integrates an object detector and a tracker. The designed model performs better than the prior scene flow methods on Argoverse 2 dataset.

3. Strengths. Consider the significance of key ideas, experimental or theoretical validation, writing quality, data contribution. Explain clearly why these aspects of the paper are valuable. Short bullet lists do NOT suffice.

> 1. This paper investigates how to evaluate the performance of scene flow methods on small objects, which has rarely been studied before.
>
> 2. The performance of TrackFlow is good on Argoverse dataset.

4. Weaknesses. Consider the significance of key ideas, experimental or theoretical validation, writing quality, data contribution. Clearly explain why these are weak aspects of the paper, e.g., why a specific prior work has already demonstrated the key contributions, or why the experiments are insufficient to validate the claims, etc. Short bullet lists do NOT suffice. Be specific!

> 1. The main contribution of this paper is a new evaluation metric, Bucket Normalized EPE. However, much information about this metric is presented in the supplemental material instead of the main text. Additionally, a clear formulation defining this metric is lacking in the paper, which hampers reader comprehension.
>
> 2. The applicability of Bucket Normalized EPE is constrained. Since it relies on point categories, this metric may fail to evaluate performance in open-set scenarios, where points from unseen categories will be ignored by this metric.
>
> 3. The comparison between the proposed TrackFlow and existing scene flow methods is unfair. First, many competing scene flow methods are unsupervised, while TrackFlow is supervised. Second, scene flow methods are class agnostic and capable of handling open-set scenarios, whereas TrackFlow relies on a pre-trained object detector, limiting its ability to estimate scene flow for points in specific categories. Although the authors mention in the supplemental material that TrackFlow can use a class agnostic open-world bounding box proposer for open-set scenarios, no experiments are provided to evaluate the feasibility and performance of this "class agnostic TrackFlow."

5. Paper rating (pre-rebuttal).

> Borderline

7. Justification of rating. What are the most important factors in your rating?

> My main concern is that the Bucket Normalized EPE and the TrackFlow may fail for open-set scenarios, while most of the current scene flow evaluation metrics and methods are class agnostic and feasible for open-set scenarios.

8. Are there any serious ethical/privacy/transparency/fairness concerns? If yes, please also discuss below in Question 9.

No

10. Is the contribution of a new dataset a main claim for this paper? Have the authors indicated so in the submission form?

No dataset contribution claim

14. Final rating based on ALL the reviews, rebuttal, and discussion (post-rebuttal).

Borderline Accept

15. Final justification (post-rebuttal).

I do appreciate the efforts that the authors made in the rebuttal. Their rebuttal has partially addressed some of my concerns, e.g., the applicability of Bucket Normalized EPE to open-set scenarios. However, I still doubt the feasibility of the TrackFlow using a class agnostic open-world bounding box proposer, which is mentioned in the third point of weaknesses. I tend to vote for Borderline Accept.

Reviewer #3

Questions

2. Summary. In 5-7 sentences, describe the key ideas, experimental or theoretical results, and their significance.

The paper addresses scene flow estimation, which analyzes the three-dimensional motion field of points in the world. It proposes two main contributions. First, it suggests revising the evaluation metric to make it class-aware and speed-normalized. Second, it introduces a new tracking baseline that follows a tracking-by-detection pipeline, utilizing a 3D object detector with a Kalman Filter for association. This new tracking baseline demonstrates strong performance in scene flow estimation on the Argoverse 2 dataset.

3. Strengths. Consider the significance of key ideas, experimental or theoretical validation, writing quality, data contribution. Explain clearly why these aspects of the paper are valuable. Short bullet lists do NOT suffice.

1.The paper is clearly written and easy to understand.

2. The observation that current scene flow evaluations prefer classes with larger volumes is sensible. The analysis reveals that although current state-of-the-art algorithms perform reasonably well, this is partly because small or dynamic objects, which have a lower percentage in evaluations, are underrepresented. This insight is valuable to the community.

2. The proposed baseline is straightforward and effective. It also highlights potential issues in current methods of acquiring scene flow ground truth, as ground truth is obtained by associating points inside 3D bounding boxes. A 3D SORT-style tracker can directly yield decent results for the task.

4. Weaknesses. Consider the significance of key ideas, experimental or theoretical validation, writing quality, data contribution. Clearly explain why these are weak aspects of the paper, e.g., why a specific prior work has already demonstrated the key contributions, or why the experiments are insufficient to validate the claims, etc. Short bullet lists do NOT suffice. Be specific!

1. The solutions proposed by the paper lack innovation in both the evaluation metric and the baseline. The paper does not introduce any new metrics for evaluation but rather conducts a more in-depth analysis by examining different factors such as classes and speed across various scene flow methods. While it is beneficial for the community to see this comprehensive analysis, it is not significant enough to claim as a novel evaluation protocol. The method resembles 3D SORT, possibly with ByteTrack enhancements for associating low-confidence objects, but it does not present significant innovation.

2. The evaluation seems tailored specifically for driving scenes, yet scene flow encompasses much more diverse scenarios, including non-driving scenes. In some cases, class definition might be difficult and debatable. For instance, while the car is a class, the car's window or wheel could also be considered separate classes. It is challenging to claim a general class-aware metric since it depends on class definitions. However, scene flow generally focuses more on low-level 3D point correspondences rather than on grouping strategies based on class definitions.

3. The proposed method appears also tailored to a specific context, as the paper only presents results on one dataset, reinforcing the above two points. The proposed metrics and method seem difficult to apply to datasets such as FlyingThings3D.

5. Paper rating (pre-rebuttal).

Weak Reject

7. Justification of rating. What are the most important factors in your rating?

My most concerns are that the proposed metric and method seems lack of novelty and are too specific for the driving scene. However, the scene flow has wider definition and application.

8. Are there any serious ethical/privacy/transparency/fairness concerns? If yes, please also discuss below in Question 9.

No

10. Is the contribution of a new dataset a main claim for this paper? Have the authors indicated so in the submission form?

No dataset contribution claim

14. Final rating based on ALL the reviews, rebuttal, and discussion (post-rebuttal).

Reject

15. Final justification (post-rebuttal).

Despite the detailed analysis provided, the paper lacks fundamentally new evaluation metrics or methods and is essentially a variant of 3D SORT. The method's limited applicability to non-driving scenarios are significant concerns. Additionally, the method's dependence on object detection performance poses a critical weakness for boarder applications. These unresolved issues require more revision for the paper, hence I do not recommend accepting the paper this time.