Rebuttal

2021年8月5日 17:16

审稿结果:

Rating: 4: Ok but not good enough - rejection
Rating: 5: Marginally below the acceptance threshold
Rating: 5: Marginally below the acceptance threshold
Rating: 5: Marginally below the acceptance threshold
Rating: 6: Marginally above the acceptance threshold
Rating: 6: Marginally above the acceptance threshold
Rating: 6: Marginally above the acceptance threshold

Rebuttal对象:

Official Review of Paper1226 by Reviewer BPY1

NeurIPS 2021 Conference Paper1226 Reviewer BPY1

Official Review19 Jul 2021 Program Chairs, Paper1226 Senior Area Chairs, Paper1226 Area Chairs, Paper1226 Reviewers Submitted, Paper1226 Authors

Summary:

The paper proposes Lite-FPN for monocular 3D object detection for autonomous driving applications. The authors take advantage of keypoint features of vehicles to obtain more accurate and reliable 3D bbox results. The proposed method can obtain very good performance-speed trade-off for real-time autonomous driving applications.

Main Review:

Originality: The idea of using keypoint for 3D object detection is not very novel. This work is upgraded from SMOKE with feature resampling from keypoint information. Another key contribution of this paper is the lightweight and real-time performance for autonomous driving applications.
Quality: Overall the paper has good quality, except for lack of innovation.
Clarity: The presentation of this paper is clear and easy to follow.
Significance: The problem to address in this paper is not very significant to me. 1) For autonomous driving purposes, monocular 3D object detection is not very practical since there are usually stereo or other sensors available on an autonomous vehicle. 2) Based on the keypoint, the method can only handle vehicle detection, but other objects are also important, e.g., pedestrians.

Limitations And Societal Impact:

Due to the limited contribution and significance of the work mentioned above, I would give a marginal reject to this paper.

Needs Ethics Review: No

Time Spent Reviewing: 3 hours

Rating: 5: Marginally below the acceptance threshold

Confidence: 5: You are absolutely certain about your assessment. You are very familiar with the related work and checked the math/other details carefully.

Code Of Conduct: While performing my duties as a reviewer (including writing reviews and participating in discussions), I have and will continue to abide by the NeurIPS code of conduct.

Official Review of Paper1226 by Reviewer bSwv

NeurIPS 2021 Conference Paper1226 Reviewer bSwv

Official Review18 Jul 2021 Program Chairs, Paper1226 Senior Area Chairs, Paper1226 Area Chairs, Paper1226 Reviewers Submitted, Paper1226 Authors

Summary:
This paper proposes to improve keypoint-based monocular 3D object detection using Lite-FPN and attention loss. Lite-FPN is a feature pyramid network that extracts features of the same location at different scales to regress bounding box information. Attention loss proposes to put different weights on the regression head according to its confidence score and localisation precision. The proposed methods show a consistent improvement across different state-of-the-art monocular 3D object detection algorithms.

Main Review:
This paper proposes two new components - Lite-FPN and attention loss. Both components are independent modules that can be integrated with any existing keypoint-based detection methods. I believe the design choices of the modules are reasonable. They successfully demonstrate that each of the proposed modules provide consistent improvement in the state-of-the-art methods.

If I were to raise a concern about this work, I am not sure if the novelty of this work is sufficient as a research paper. I find it difficult to agree with the novelty of Lite-FPN. Feature pyramid networks is a widely used network structure since 2015 and it is a very well known common knowledge that multi-scale features help multi-scale detection. The only major difference of the proposed method would be that the applied the multi-scale feature only on the regression head. I do not believe that there is a significant technical difference in reorganizing the regression phase from post process phase to the detection phase. I am not sure if attention loss by itself is a significant-enough novelty.

I do not see why the authors chose to not use multi-scale features on keypoint regression. I see a reasoning that keypoint detection is less of a bottleneck however I do not see how it would harm the performance and it would make sense for me to leverage full multi-scale features in both keypoint and bounding box regression.

I also wonder how the confidence score and 3D IoU impacts the significance of the attention. An ablation study of attention only based on each of these two would be an interesting experiment.

Limitations And Societal Impact:
The authors did not address the limitation of their work.

Needs Ethics Review: No

Time Spent Reviewing: 3hrs

Rating: 6: Marginally above the acceptance threshold

Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

Code Of Conduct: While performing my duties as a reviewer (including writing reviews and participating in discussions), I have and will continue to abide by the NeurIPS code of conduct.

Official Review of Paper1226 by Reviewer VCNY

NeurIPS 2021 Conference Paper1226 Reviewer VCNY

Official Review16 Jul 2021 Program Chairs, Paper1226 Senior Area Chairs, Paper1226 Area Chairs, Paper1226 Reviewers Submitted, Paper1226 Authors

Summary:
This paper proposes a new network architecture for Monocular 3D Detection task. Keypoint-based FPN-style detection backbone is used for feature extraction. A new attention loss is proposed for alleviating the false

attention issue.

Main Review:

Originality / Significance:

I think the method this paper proposes is based on several existing lidar detection methods, and lacks novelty. Therefore I am giving it a boarderline reject.

Quality / clarity:

I think the overall writing of this paper is of good quality.

Limitations And Societal Impact:

I don't think there is any "limitations and potential negative societal impact" in this work. This is a very general lidar/camera detection task/model.

Needs Ethics Review: No

Time Spent Reviewing: 2 hours

Rating: 5: Marginally below the acceptance threshold

Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

Code Of Conduct: While performing my duties as a reviewer (including writing reviews and participating in discussions), I have and will continue to abide by the NeurIPS code of conduct.

Official Review of Paper1226 by Reviewer pyuf

NeurIPS 2021 Conference Paper1226 Reviewer pyuf

Official Review14 Jul 2021 Program Chairs, Paper1226 Senior Area Chairs, Paper1226 Area Chairs, Paper1226 Reviewers Submitted, Paper1226 Authors

Summary:

This paper proposes a keypoint-based monocular 3D detection method. Compared to prior work, a multi-scale detection head is used to improve object detection across a large range of scales and distances. Also, a weighted loss is used in regression to improve localization and reduce false positives. Results are validated on KITTI, showing promising improvement over the baselines

Main Review:

Paper Strengths

1. The ablation experiments confirm that the idea of adding multi-scale detection head and attention loss help improve the performance consistently
2. Main experiments are validated over two baselines, SMOKE, and CenterNet, showing that the proposed techniques are generic to different keypoint-based monocular detection methods

Paper Weaknesses

1. The main concern I have for this paper is that the technique is really simple and there is no significant ML novelty. Essentially, the proposed two techniques (multi-scale feature extraction and weighted loss) are very standard engineering tricks used everywhere in CV/ML. So I would view this paper as more of an application paper, which might be better suited for ICRA/IROS rather than NeurIPS (stronger ML novelty expected)
2. Although the ablation experiments are great examples demonstrating the usefulness of the method, it would be even better to add some qualitative examples to show in which cases the proposed techniques help obtain

better boxes. This can help readers better interpret how the proposed method actually works. For example, is it really true that, after adding weighted loss, detection with high confidence but low IoU became high confidence and high IoU, i.e., localization is improved? Can we have some statistics about how the percentage of objects varies over the confidence and IoU? Moreover, is it really true that, after adding a multi-scale detection head, detection becomes better at a larger range of scales? If so, can we draw detection performance over scales and distances to analyze in which scales the performance is improved and in which scales the baseline without multi-scale detection is not good enough?

3. Would be great to validate the proposed method on another dataset as well because KITTI is too old, e.g., nuScenes
4. The experiments seem to be evaluated under 11 thresholds (old tradition of KITTI). However, KITTI changed the evaluation protocol to use 40 thresholds a year ago. Would be great to re-evaluate the proposed method under the current standard evaluation protocol (40 thresholds) so those future methods can compare against

Limitations And Societal Impact:
Would be great to add some failure case analysis with qualitative results to help readers understand the limitation. Currently, only some future directions are pointed which is not efficient. Also, no potential negative social impact is discussed in the paper

Needs Ethics Review: No

Time Spent Reviewing: 2

Rating: 6: Marginally above the acceptance threshold

Confidence: 5: You are absolutely certain about your assessment. You are very familiar with the related work and checked the math/other details carefully.

Code Of Conduct: While performing my duties as a reviewer (including writing reviews and participating in discussions), I have and will continue to abide by the NeurIPS code of conduct.

Official Review of Paper1226 by Reviewer B8kL

NeurIPS 2021 Conference Paper1226 Reviewer B8kL

Official Review10 Jul 2021 Program Chairs, Paper1226 Senior Area Chairs, Paper1226 Area Chairs, Paper1226 Reviewers Submitted, Paper1226 Authors

Summary:
The proposed framework is based on keypoint-based monocular 3D object detectors. First, they proposed a multi-scale feature fusion to address objects in multiple scales and second, they introduced an attention loss to handle the misalignment between classification and localization. The proposed technique is validated using two keypoint-based detectors on KITTI. The experimental results are promising on KITTI dataset.

Main Review:
Strengths:
The proposed idea is simple and intuitive. The technical part in general has merits. The experiments show promising results on KITTI (large improvements over the baselines). The paper is well written.
Weaknesses:

1. The technical part of this paper has merits. However, some pieces of ideas in their design are not totally new. Using multi-scale feature maps to capture objects in different scales is a popular strategy in object detection and segmentation works. Sampling key points in feature maps was

also investigated in 2D object detection methods (e.g., SaccadeNet: A Fast and Accurate Object Detector, CVPR2020) and 3D object detection methods (e.g., Infofocus: 3d object detection for autonomous driving with dynamic information modeling, ECCV2020).

2. Did you quantitatively analyze whether localization precision is the major failure case compared to classification? A straightforward way would be (1) using the groundtruth class for each prediction and keep the location prediction and (2) using the groundtruth location and keep the predicted class. Then, see which one improves the performance more.
3. KITTI is a relatively small dataset. Experiments on larger dataset like nuScenes would be more beneficial.
4. Missing the comparison with the SOTA methods in general.

Limitations And Societal Impact:
See limitations in the weaknesses.

Ethical Concerns:
N/A

Needs Ethics Review: No

Time Spent Reviewing: 2.5 hours

Rating: 6: Marginally above the acceptance threshold

Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

Code Of Conduct: While performing my duties as a reviewer (including writing reviews and participating in discussions), I have and will continue to abide by the NeurIPS code of conduct.

Official Review of Paper1226 by Reviewer 9nui

NeurIPS 2021 Conference Paper1226 Reviewer 9nui

16 Jul 2021NeurIPS 2021 Conference Paper1226 Official ReviewReaders: Program Chairs, Paper1226 Senior Area Chairs, Paper1226 Area Chairs, Paper1226 Reviewers Submitted, Paper1226 Authors

Summary:
The paper tackles the task of monocular 3D object detection by proposing a light-weight feature pyramid network (Lite-FPN) and a new attention loss. The main contributions of the manuscript are the proposal of a new FPN module moving the top-k operation from post-processing phase to the detection phase and the introduction of an attention loss during the training, promoting accurate detections for low-confidence detections. The proposed modules are evaluated on two state-of-the-art monocular 3D object detection pipelines showing improvements on the KITTI dataset on the class car.

Main Review:
The proposed approach introduces two new modules for monocular 3D object detection: the Lite-FPN and the attention loss. Lite-FPN is the first contribution of the paper and consists in a general FPN module that can be introduced in existing works. **[1]** Although it is clearly reported in the method section that Light-FPN moves the top-K operations from post-processing to detection, the novelty of the proposed module is limited being more a different procedure than a new FPN architecture. The second contribution is the introduction of an attention loss promoting low confidence predictions to have an higher detection accuracy. Although novel, also the introduction of a new loss is a minor improvement. **[2]** Other weakness of the paper regards the

writing quality, the clarity in the explanation of the main contributions and, the overall organisation of the method section. The writing quality and the clarity in the explanations are largely below the acceptance threshold for the NeurIPS conference. Until the method section, it is not clear whether authors propose a new detection architecture or, as it is, improvements in some modules of detection pipelines. Also the experimental section lacks important aspects. First, to evaluate their proposals, authors report experiments on the validation set of the benchmark KITTI for the class car by using the AP metric. Unfortunately this is no longer the official detection metric in favour of the official AP metric. Therefore, the results in Table 1 and Table 2 should be updated being not completely reliable. Second, authors propose an improved version of SMOKE and compare it with state-of-the-art baselines showing better performance in Table 5. Although here the official AP is used, **[3]** important recent baselines are missing in the comparison such as M.Ding et. al "Learning depth-guided convolutions for monocular 3d object detection" and, Liu et. Al "Ground-aware monocular 3d object detection for autonomous driving" which, on the same KITTI benchmark, perform better than the proposed method. Finally, **[4]** experiments are performed only on the KITTI dataset and only for the car class. The addition of detection benchmarks such as nuScenes and the extension to all the benchmark classes are needed for a clear and fair comparison with existing methods.

Limitations And Societal Impact:

yes

Needs Ethics Review: No

Time Spent Reviewing: 1

Rating: 4: Ok but not good enough - rejection

Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

Code Of Conduct: While performing my duties as a reviewer (including writing reviews and participating in discussions), I have and will continue to abide by the NeurIPS code of conduct.

Official Review of Paper1226 by Reviewer P4Zx

NeurIPS 2021 Conference Paper1226 Reviewer P4Zx

29 Jun 2021 (modified: 17 Jul 2021)NeurIPS 2021 Conference Paper1226 Official ReviewReaders: Program Chairs, Paper1226 Senior Area Chairs, Paper1226 Area Chairs, Paper1226 Reviewers Submitted, Paper1226 Authors

Summary:

The authors propose two improvements for CenterNet-like models for 3D object detection from RGB images. First, they **propose to regress the 3D properties of the bounding boxes from a stack of features pooled at different resolutions from the backbone decoder rather than just from the last layer.** Second, they design an improved loss function that helps the training process to focus more on well localized boxes while optimizing the regression parameters of the 3D bounding boxes. In the experimental section, the two additions improve the performance of CenterNet and SMOKE on the KITTI dataset both in terms of quality and in terms of FPS.

Main Review:

Summary

The paper introduces two improvements for keypoint based 3D object detection networks and tests them on one datasets and two base algorithms: CenterNet and

SMOKE. In both experiments the additions proposed by the authors do bring some improvements over the baselines, but the final method still underperforms compared to other alternatives. However when taking the fps into account the methods provide a good tradeoff between accuracy and speed. The two contributions are a bit incremental with respect to other published works and I would have liked more experimental evidence across more classes and datasets which are not present in the current version of the paper. **For these reasons I’ m currently rating the paper 5, but I’ m happy to reconsider based on the feedback provided by the author during the rebuttal.** I will summarize below strengths and weaknesses of the paper as well as some questions and suggestions for the authors.

Strength

a) The two improvements introduced in this work are pretty simple to implement and general enough to be applied to a possibly huge range of 3D object detection models.

b) The novelties complicate the training phase of the model but have no side effect on the test time performance or inference time (improving both).

Weakness

a) Currently the experimental section of the paper is considering a single dataset (KITTI) and a single class (Cars). Considering that newer and bigger datasets for the task have been released (e.g., Waymo [I] or nuScenes [II]) testing on them would have made the claim of the paper more substantiated. Currently there is no experimental evidence that the proposed improvements generalize to different scenarios (which I do believe being the case). Moreover some of the competitors do test on all the KITTI classes, (e.g., the cited MonoDIS[31] or the non cited MonoPair[III]). In the current form there is no guarantee that the proposed solutions generalize besides the car class on the KITTI dataset.

I. Scalability in perception for autonomous driving: Waymo open dataset. Sun et Al.

II. nuScenes: A multimodal dataset for autonomous driving. Caesar et Al.

III. Monopair: Monocular 3d object detection using pairwise spatial relationships. Chen, Yongjian, et al.

b) The feature pooling schema reminds me of ROI-pooling in two stage object detection frameworks, but unfortunately there is no study on the effect of how many levels of feature representations to pool (i.e, 1, 2 or 3). For example it would have been interesting to have in Tab. 4 a line with only Attention Loss enabled as it corresponds to the proposed method pooling feature only from a single layer (the last one). Or even better a proper study showing a comparison between different methods polling from a different set of layers multi-scale features

c) The motivation behind the choices made when developing the improved loss function (l: 200-213) are a bit empirical and in general hard to follow. In the end, Eq. 5 gives more weight to confident prediction with a bad IOU. While I understand why giving more weight to elements based on their predicted confidence is valuable, it is not clear to me why the IOU3D part is needed in Eq. 5. If the boxes have a good IOU their regression loss `l_reg_i` will be low already, so it is not clear why it should be downweighted further. Moreover in the paper there is no study on the impact of `\beta` on the final performance which would help solve my doubts expressed above.

d) Some implementation details of the method are currently missing. For example: when pooling K locations during training/inference, are those obtained after a NMS stage based on max pooling as in the original CenterNet? If not, are the authors doing anything to avoid pooling overlapping locations nearby a local maximam? Moreover, what is the value of K used for the experiments?

e) There is a lack of discussion of limitations of the current method. The discussion about the limitations is limited to a single sentence in Section 5. When building on top of existing works (CenterNet/SMOKE) I would like to see both wins (e.g., the sample in Fig. 4) as well as Losses, if any (i.e., samples where the proposed model performs worse than the original CenterNet or SMOKE). This will help to expose possible biases induced by the new proposed training objective.

Questions

1. Why is the x N in eq. 5 needed? | In Eq. 2 L_reg is normalized by the number of keypoints while in Eq. 4 it is not. Wouldn't multiply by N the weights scale the overall loss by N as well?
2. Are gradients back propagated through the weight computation in Eq. 4 and 5? This will change the interpretation of the loss function as it will have a direct effect also on the keypoints prediction branch. **If so please state this in the final text.**
3. Is N in Eq. 1, 2 and 5 the number of keypoints per image or per batch of images? I'm assuming the former as the rest of the loss discussion seems to refer to a single image, but in the text N is presented as "the total number of keypoints per batch images".
4. **Can you clarify my doubts regarding weakness (d.)?**

Suggestions

- I would add to Table 3 also the results from the best performing model on the validation set: Ours-SMOKE with DLA-34
- Regarding Checklist 2.a and 2.b I don't think that this paper contains any theoretical result.

Limitations And Societal Impact:
Weakness (e) of my main review summarize my concern about the lack of discussion on the limitation of the current work.

Needs Ethics Review: No

Time Spent Reviewing: 4

Rating: 5: Marginally below the acceptance threshold

Confidence: 3: You are fairly confident in your assessment. It is possible that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.

Code Of Conduct: While performing my duties as a reviewer (including writing reviews and participating in discussions), I have and will continue to abide by the NeurIPS code of conduct.