View Reviews

Paper ID

10492

Paper Title

Self-supervised Learning of 3D Object Understanding by Data Association and Landmark Estimation for Image Sequence

Reviewer #1

Questions

1. [Summary] In 3-5 sentences, describe the key ideas, experiments, and their significance.

This paper utilizes video sequence to improve object pose estimation in monocular image. Three data association strategies are deployed to track multiple observatons of mutiple object in videos. The global landmark map can be obtained using multiple observations, and then object annotations can be derived. The experiments on KITTI dataset show promising results on objects' depth and orientation estimation.

2. [Strengths] What are the strengths of the paper? Clearly explain why these aspects of the paper are valuable.

1. The structure of this paper is well organized.
2. The experiments on KITTI dataset show the effectivness.

3. [Weaknesses] What are the weaknesses of the paper? Clearly explain why these aspects of the paper are weak. Please make the comments very concrete based on facts (e.g. list relevant citations if you feel the ideas are not novel). If applicable, please indicate key issues and questions which, if well addressed during the 1-page rebuttal, might influence you to change your rating.

1. There are many grammar mistakes in this paper.
2. The experiment uses only one dataset, which is not exhaustive enough.
3. The idea of this paper is easy for researchers in this field to come up with, so the novelty is insufficient.

4. . [Overall rating] Paper rating (pre-rebuttal)

Strong Reject

5. [Detailed comments] Additional comments regarding the paper (e.g. typos, any suggestions to make the submission stronger).

Introduction:
1. "sequencial dataset" should be "sequential data".
2. The last paragraph should describe the approach in this paper. The description here is not clear and gives no information about the method's novelty.

Related Work:
1. There are a lot of grammar problems.
-- In addition, encoders that can extract those deep features can effectively be trained through SGD can be freely fine-tuned with any constraint.
There is more than one verb in one sentence.
--Through these methods, it is achievable to perform more robust category classification or clustering for other domains, or to further improve the performance of determining objectness for multi-object scenes.
There are two "or"s in one sentence.
--The most effective in practical network learning is to directly learning the likelihood with the assigned labels for specific tasks.
There is no noun in this sentence. "to learning" is wrong.
There are also many other grammar problems in this section.
2. For the references cited, the author does not introduce these approaches in detail. At least this paper should give a one-sentence description of other approaches.

Approach:
1. There exists problems with sentence organization.
-- Then, the following limitations exist; first, localization may fail due to incorrect depth estimation. Next, pose estimation failure can be occurred due to the incorrect viewpoint prediction.
-- Therefore, even the observations are obtained from single network, observations can be corrected by taking average of the multiple observations.
This sentence should be re-organized to two sentences.
2. Why the author does not show 3D bounding box in the left column of Fig. 2?
3. The proposed method utilizes three methods for data association. The paper does not say whether these three methods are used simultaneously or if one fails and performs others. Do these three methods have different priority?
4. The equations 5-10 are simple geometric transformations without any novelty.

Experiments：
1. “detection annotaion” should be "detection annotation".
2. The experiments are only conducted on one dataset. The author shoud give performance on other datasets to demonstrate the generality of this method.
3. In 4.3, "for the object’s depth and orientation." should add "estimation".

6. [Reproducibility] Is the method described in this paper reproducible?
The paper includes information that would make it possible to reproduce the methods and experiments

7. [Confidence] Reviewer's confidence in his/her recommendation
Confident

10. Please provide an "Overall Rating", following the rebuttal and reviewers discussions.
Strong Reject. I will fight for rejecting this submission.

11. Justification of final rating. Describe the rationale for your final rating, including notes based on the rebuttal, discussion, and other reviews.
This paper is below the limit of the conference. I agreed with several other reviewers in rejecting it.

Reviewer #2

Questions

1. [Summary] In 3-5 sentences, describe the key ideas, experiments, and their significance.
The authors have presented a method to detect vehicles in the KITTY dataset that leverages multiple frames, and compare it to a single frame method.

2. [Strengths] What are the strengths of the paper? Clearly explain why these aspects of the paper are valuable.
The authors appear to be familiar with the concept of SLAM and object detection, and were able to implement an object detection model.

3. [Weaknesses] What are the weaknesses of the paper? Clearly explain why these aspects of the paper are weak. Please make the comments very concrete based on facts (e.g. list relevant citations if you feel the ideas are not novel). If applicable, please indicate key issues and questions which, if well addressed during the 1-page rebuttal, might influence you to change your rating.
This paper is written in a very confusing way that makes it difficult to follow, or even understand what are their goals. The most problem is that almost all sentences are so generic that it's hard to grasp what are they referring to. They aim to solve the problem of "3D Object Understanding", which may refer to an entire field. Just as an example, in fig.2 it says "For comparison, we show the green 2D bounding boxes obtained by other detection method.", but it never states what is this other detection method.

On top of this, using video to improve object detection, or using video to improve 3D mapping, are both tasks that have been performed extensively in literature, therefore the novelty is also low.

4. . [Overall rating] Paper rating (pre-rebuttal)
Strong Reject

5. [Detailed comments] Additional comments regarding the paper (e.g. typos, any suggestions to make the submission stronger).
Due to the lack of novelty, and the lack of definition of most concepts expressed in the paper, it is difficult to recommend it for publication. The idea might be valuable, but the authors need to be precise and not overstate their own work: descibe exactly what problem did you solve, e.g., localizing cars in the KITTY dataset, why is it meaningful, and how is it solved and evaluated.

6. [Reproducibility] Is the method described in this paper reproducible?
I don’t know / can’t tell

7. [Confidence] Reviewer's confidence in his/her recommendation
Confident

Reviewer #3

Questions

1. [Summary] In 3-5 sentences, describe the key ideas, experiments, and their significance.
The authors present a method to create annotations for object detection from a video sequence. Given a video and the known camera poses, a global map of static objects is built by SALM approach. In this way, objects information of the global map is used as the annotations. The proposed method is evaluated on KITTI dataset. Experimental results on depth and orientation are reported in the experiments.

2. [Strengths] What are the strengths of the paper? Clearly explain why these aspects of the paper are valuable.
In this paper, the authors use SLAM method to build a global map of static objects. By having the global map of static landmark (ex: car), annotations of vehicle can be obtained easily. Therefore, the proposed method provides a way to generate annotation of objects from a video.

3. [Weaknesses] What are the weaknesses of the paper? Clearly explain why these aspects of the paper are weak. Please make the comments very concrete based on facts (e.g. list relevant citations if you feel the ideas are not novel). If applicable, please indicate key issues and questions which, if well addressed during the 1-page rebuttal, might influence you to change your rating.
I have several questions about the method.

1. The global map built by SLAM method mainly contains static objects. Thus, it has limited benefit for application such as autonomous driving, since they are many dynamic objects in the scene. Also, in the experiments, dynamic objects are excluded from the video. Although the proposed method introduces a approach to generate object annotations, it seems that this method has limitation on generating annotations for dynamic objects.

2. Missing reference:
[ref-1] P. Li et al, Stereo Vision-based Semantic 3D Object and Ego-motion Tracking for Autonomous Driving, ECCV 2018
In [ref-1], the authors present a dynamic object bundle adjustment (BA) approach for 3D bbox detection. Although the method is about the stereo camera, it also tries to solve the training data problem.

4. . [Overall rating] Paper rating (pre-rebuttal)
Weak Reject

5. [Detailed comments] Additional comments regarding the paper (e.g. typos, any suggestions to make the submission stronger).

I have several comments about the presentation of the paper.

1. It would be better to provide the formulation of the problem. For example, it is not clear what is the global map in the paper. What is the known information (images in the video and camera poses)? and what is the desired output?

2. In the experiments, the evaluation metrics are depth and orientation of the objects. It is better to provide results with other 3D object detection metrics.

3. It is not clear how to get the dimension of 3D bbox in the map?

4. It would be better to simplify the presentation of the three methods for data association?

5. In Fig. 2, it would be great to give more details of the red and green box?

6. Please clarify the difference between upper bound and proposed method.

7. There are too many ‘however’ in the paper.

6. [Reproducibility] Is the method described in this paper reproducible?

The paper includes information that would make it possible to reproduce the methods and experiments

7. [Confidence] Reviewer's confidence in his/her recommendation

Confident

10. Please provide an "Overall Rating", following the rebuttal and reviewers discussions.

Reject. I vote and argue for rejecting this submission.

11. Justification of final rating. Describe the rationale for your final rating, including notes based on the rebuttal, discussion, and other reviews.

The authors didn't provide rebuttal response. I agree with other reviewers to reject the submission.

Reviewer #4

Questions

1. [Summary] In 3-5 sentences, describe the key ideas, experiments, and their significance.

In this paper, the authors propose a method to augment the training data for monocular camera image-based pose estimation. They use some data association techniques and additional global landmark map information to correct the predictions within a sequence. In the experiments, they choose the KITTI 3D object detection dataset as the basic training dataset with annotations and use the KITTI Odometry dataset as the targeted sequential dataset to generate annotations. Overall, the methods proposed in this paper is not comprehensive and convincing enough, and the related works for this task have not been well addressed by the authors.

2. [Strengths] What are the strengths of the paper? Clearly explain why these aspects of the paper are valuable.

The quality of the reported figures is good, but the descriptions below the figures lack clarity.

3. [Weaknesses] What are the weaknesses of the paper? Clearly explain why these aspects of the paper are weak. Please make the comments very concrete based on facts (e.g. list relevant citations if you feel the ideas are not novel). If applicable, please indicate key issues and questions which, if well addressed during the 1-page rebuttal, might influence you to change your rating.

This paper lacks sufficient language delicacy, and there are also some expressions in this paper that sound unnatural and wordy. For instance, “the need for training datasets for various environments persists, as the scope of the needs of autonomous vehicles or mobile drones are currently expanding globally”, “In addition, the performance of the visual perception should not be highly dependent on the environment for the adaptive applications”, “Due to the nature of deep learning based on statistics,”.

Some claims appear too repeatedly, such as “the performance is bound to the network itself”, “In other words, the obtained labeling result is bounded to the self-performance of the network”.
The related works are just listed together without elaborating the key ideas that inspire this paper, and di not quite contribute to the proposed method. For instance, “This series of processes is essentially used in most SLAM techniques using sequence input [32, 28, 4, 43, 48, 44, 10, 41, 42]”
Some related works even lack references, such as “By performing feature matching using SIFT or SURF, we can perform object tracking”, the baseline metho in Table 2.

4. . [Overall rating] Paper rating (pre-rebuttal)

Strong Reject

5. [Detailed comments] Additional comments regarding the paper (e.g. typos, any suggestions to make the submission stronger).

There are many typos and unnatural expressions in this paper. For instance, “combine these multiple detection”, “only camera poses at the time of observation are given”, “By further learning the network using this new dataset, the weak point of prediction for a single image of the network is compensated”, “performance improvement is confirmed in the pose estimation of the object, especially orientation estimation”.

6. [Reproducibility] Is the method described in this paper reproducible?

I don’t know / can’t tell

7. [Confidence] Reviewer's confidence in his/her recommendation

Very confident

10. Please provide an "Overall Rating", following the rebuttal and reviewers discussions.

Strong Reject. I will fight for rejecting this submission.

11. Justification of final rating. Describe the rationale for your final rating, including notes based on the rebuttal, discussion, and other reviews.

I stand with my initial rating on this paper since it requires too much wiriting to be polished and the idea proposed in this paper is confusing and not comprehensive.