View Reviews

Paper ID
53

Paper Title
Towards Implicit Representation of a Pose: a New Improved Pipeline of Augmented Autoencoder

Reviewer #1

Questions

2. Please assign a grade to the paper:
-1: Weak reject — A paper that doesn't have any technical issues, but you are unsure meets the bar for WiCV. Such a paper would likely be in the bottom 25-30% of papers

3. Is this abstract/ paper appropriate for WiCV?
Yes

4. Does the abstract/ paper adequately convey the material that will be presented?
Yes

5. Does the abstract/ paper describe work that is novel and/or interesting?
Yes

6. Write a review evaluating the paper in a minimum of 5 sentences. When evaluating the paper, take into account the following questions: 1. Is this abstract/paper appropriate for WiCV (Does it describe original research in Computer Vision or related fields?) 2. Does the abstract/paper describe work that is novel (to the best of your knowledge - you may not be able to adequately answer this if it is not your field) and/or interesting? 3. Does the abstract/paper adequately convey the material that will be presented?
- This paper presents a solution for 6DOF pose estimation, with an additional industrial use-case of 4 objects - a spark plug key, a nut, a nozzle, and a screw. The industrial use-case is interesting and practical as this may be used for pick and place of these objects in industry like in an automobile manufacturing plant employing manipulator robots.
- Keeping in mind the practicality of the problem, the design of the dataset to use highly cluttered backgrounds does not seem well-motivated. In such scenarios, it seems the background should also contain the same object. For example, there may be a tray full of screws from which a robot is expected to pick one at a time.
- Also, it seems that using a black background turned out to be better than using the cluttered backgrounds which indicates the difficulty of segmenting in such cases. But this aspect (w.r.t. segmentation difficulty) is not discussed.
- The results for the industrial use-case are only for the spark plug key object and other results are omitted despite there being enough space for an extra page of content. Even if spark plug key is the most difficult object class, all the results should be reported. Also results are not reported for prior methods on this new dataset, so one cannot assess whether this new dataset represents a bigger challenge than usual pose estimation benchmarks.
- The results on the other LineMod dataset also compare only with [25]. More prior works should be included in the comparisons since [25] is (at this time) over two years old. Moreover, on paperswithcode, I could find a LineMod benchmark - https://paperswithcode.com/sota/6d-pose-estimation-on-linemod where Augmented Autoencoder [25] is the second-last on the leaderboard of about 20 entries, so I believe there are stronger baselines available.
- Fig. 4 - There is a spike in the training losses towards the end which is not explained.

- Overall, this paper is an extension of [25] which replaces detection with segmentation which is non-trivial and interesting. However, the presentation of the paper is quite poor with limited results, apart from other issues described above. Hence, I give a weak reject rating.

Reviewer #2

Questions

2. Please assign a grade to the paper:
1: Weak accept — A paper that is in the top half of submissions

3. Is this abstract/ paper appropriate for WiCV?
Yes

4. Does the abstract/ paper adequately convey the material that will be presented?
Yes

5. Does the abstract/ paper describe work that is novel and/or interesting?
Yes

6. Write a review evaluating the paper in a minimum of 5 sentences. When evaluating the paper, take into account the following questions: 1. Is this abstract/paper appropriate for WiCV (Does it describe original research in Computer Vision or related fields?) 2. Does the abstract/paper describe work that is novel (to the best of your knowledge - you may not be able to adequately answer this if it is not your field) and/or interesting? 3. Does the abstract/paper adequately convey the material that will be presented?
Summary:
This paper proposes an improvement on the well-known implicit 6D pose estimation, namely Augmented Autoencoder. The paper argues that this approach is powerful since it does not require annotated datasets, but identifies some critical aspects of AAE: structured noise, such as adding background images instead of random noise, affects the implicit representations. It proposes using segmentation instead of detection. And compare to the original AAE using the LineMode Dataset.
In addition, it introduces a new synthetic industrial use-case that consists of 4 typical industrial objects and a geometric background. And test it with real images from said use-case.
Discussion of related work seems thorough although I am not familiar with some related work.
The paper was relatively easy to read, with some issues with the clarity of writing.

Strengths:
+ The paper mainly introduces two changes to AAE: Explore adding VOC images as background for LineMode Dataset (AAE va LessAAE which has black background) and using segmentation instead of detection. Both are justified theoretically.
+ The paper gives detail and constructs a synthetic dataset for an industrial use case.
+ For their results, quantitative and qualitative, it seems that this approach (using segmentation and VOC background images) works better for both datasets.
+ For metrics, they use eADD(-S), which is the error associated to the average distance to the correspondent model point and recall percentage. For related work, it looks like they use the same metrics.
+ After segmentation, they study the effects of using AAE and lessAAE, which seems to be novel.
+ Use Yolov7 for instance segmentation and a Less Augmented Autoencoder

Weaknesses:
- There is no mention of the release of the synthetic dataset. I assume that although the background in the synthetic dataset is not the same in different images, from Figure 2 it looks very similar. I suspect this might be one of the reasons why segmentation fails, explained in section 4.2 and shown in Figure 6.
- Some further discussion of the results is needed. For example, in Table 5, explain why there is a large difference between the Spark Plug Key and the AAR vs LessAAE. Or why is there a peak in the training losses for Figure 4.
-Figures can be improved:
-To improve readability, the size of the text should be increased in Figure 4, Table 1, Table 3, Table 4.
-To understand a pipeline, an overview figure of the method is needed (where we can see the images, CAD, and models with the outputs).
-Overall, the figures would be easier to understand if they were close to the explanation of said figures.

-Figure 4 can be deleted, as it does not contain much extra information and can be perfectly summed up with text.

- Some issues with the writing that can be easily modified:

-Line 76: “lead to breakthrough” should be “led to a breakthrough”.

-Line 127: explain what the concept of noise impedance is.

-Line 177-178: sentence “In this paper we move to 6D pose estimation as classification problem”. I suggest replacing this sentence by the definition of implicit models.

-L192: change met for found.

-L240: it is not clear what the values that define P are.

-L256: “we faced whit”. Delete

-L428: The Deep Image prior is not clear what it means. Define first and applications later.

-L491: ADD(-S) is this how the commonly name this metric? I think related work refers to it as ADD

-L484: remove italics for Less Augmented Autoencoder.

-L522: The dataset of LineMode is referred to as LM, but related work refers to it as LineMOD, so are there any differences?

-L550: the class Ape for LinneMOD has not been introduced. So the phrase is hard to understand.

-L556: needs rephrasing. “Recently have been proposed many accurate and efficient options”.

-L586: there is a reference to figure 4 that is not introduced. Missing as shown in figure 4.

-L616 & 618 two “in fact” in the same paragraph

Rating and Justification:

Overall, I think the weaknesses could be rectified for the camera-ready submission, especially so that people who are not in the field are able to understand.

The problem is motivated and the solution justified. Although I am not very familiar with the related work, it looks like the paper makes contributions that can help advance the field.