← Back to **Author Console** (/group?id=NeurIPS.cc/2023/Conference/Authors#your-submissions)

Training Dense Object Nets: A Novel Approach

(/pdf?id=lEnMPgQWiB)

Kanishk Navale (/profile?email=kanishk.navale%40sereact.ai), Ralf Gulde (/profile?id=~Ralf_Gulde1), Marc Tuscher (/profile?email=marc.tuscher%40sereact.ai), Oliver Riedel (/profile?email=oliver.riedel%40isw.uni-stuttgart.de)

09 May 2023 (modified: 15 Jul 2023) NeurIPS 2023 Conference Submission Conference, Area Chairs, Authors, Reviewers, Senior Area Chairs Revisions (/revisions?id=lEnMPgQWiB)

Keywords: Dense Object Nets, object generalization, computation costs, dense visual local descriptors, synthetic data generation, robot grasping

Abstract:
Our work proposes a novel framework that addresses the computational resource limitations associated with training Dense Object Nets (DON) while achieving robust and dense visual object descriptors. DON's descriptors are known for their robustness to viewpoint and configuration changes, but training them requires computationally expensive image pairs with correspondence mapping. This limitation hampers dimensionality and robustness, thereby restricting object generalization. To overcome this, we introduce a synthetic augmentation data generation procedure and a novel deep learning architecture that produces denser visual descriptors with reduced computational demands. Notably, our framework eliminates the need for image-pair correspondence mapping and showcases its application in a robot-grasping pipeline. Experimental results demonstrate that our approach yields descriptors as robust as those generated by DON.

Corresponding Author: kanishk.navale@sereact.ai
Reviewer Nomination: marc.tuscher@sereact.ai
Primary Area: Deep learning architectures
Claims: yes
Code Of Ethics: yes
Broader Impacts: yes
Limitations: yes
Theory: yes
Experiments: yes
Training Details: yes
Error Bars: yes
Compute: yes
Reproducibility: yes
Safeguards: n/a
Licenses: n/a
Assets: yes
Human Subjects: no
IRB Approvals: no
Submission Number: 5130

Filter by reply type
Filter by author
Search keywords...
Sort: Newest First

Everyone | Program Chairs | Area Chairs | Authors | Reviewer CgwT | Reviewers Submitted | Senior Area Chairs | Reviewer 6xja | Reviewer msA6 | Reviewer FtQk | Reviewer NQ4e | ✖

5 / 5 replies shown

Add: **Withdrawal** **Official Comment** **Author Rebuttal**

Official Review of Submission5130 by Reviewer CgwT

Official Review Reviewer CgwT 09 Jul 2023, 21:31 (modified: 01 Aug 2023, 21:17)

Program Chairs, Area Chairs, Authors, Reviewer CgwT, Reviewers Submitted, Senior Area Chairs

Summary:
This paper proposes to extend dense object nets (DON) [14] with a synthetic data generation and data augmentation pipeline for generating training data, and training on an auxiliary task of keypoint prediction based on [31]. The approach is evaluted on a custom dataset of cap images/models and compared with DON baseline in terms of accuracy and GPU memory consumption for training. Also some qualitative results of descriptor matchings in synthetic and real images are demonstrated and the outcome of two runs of robot experiments are reported.

Soundness: 1 poor
Presentation: 2 fair
Contribution: 1 poor
Strengths:

- The proposed approach for using keypoint detection as an auxiliary task for DON seems novel.

Weaknesses:

- The contributions of the approach are limited. Firstly, synthetic data generation for training is a standard technique and [20] already proposed the applied data augmentations. The second contribution of using an existing neural keypoint detection approach [31] as an auxiliary loss is incremental. The experimental results cannot convincingly demonstrate the benefit of this approach.
- The abstract uses terms such as dimensionality, robustness, and object generalization which are not properly defined. Also, the claims of robustness and generalization in relation to DON are not properly assessed in experiments and hence not justified.
- l. 41, what does generalization to a certain extent mean in a scientific sense?
- l. 49, what defines a "complete" dataset ?
- l. 96ff, the difference of the proposed method to [27] is not clear in this paragraph. Which reference is Adrian et al. ? Why does it matter here?
- The method is only demonstrated for caps. Does it generalize to other objects as well?
- Sec 3.1, why not render the images in the sampled view poses of the data augmentation, if a synthetic data generation pipeline is used.
- Fig. 2, there seems to be an extra branch of upsampling from the ResNet features to the dense object net descriptors. How are the dense object net descriptors trained if it is separate from the keypoint branch on which the loss functions are defined?
- l. 148, please refer to an equation in [21,31] to define f_E
- Sec. 3.3 please refer equations in [31] to define the loss function components. What is the effect of the loss components? please summarize.
- eq 5, define I_cam
- l. 162, what defines a "smoother gradient" ? smoother than what?
- eq 7, how is T_pred defined/determined?
- eq 7, T^dagger is simply T^-1
- eq 8, why is the result of this minimization problem an expectation value ? The distribution over u,v in I_d does not need to be a Gaussian.
- Sec. 4.2 what is the issue with DON's scaling properties? What is the difference in the network architectures if one just considers the part that maps images to dense object descriptors?
- Table 1, so why not use GPUs or GPU systems with more memory for training if DON has better performance? Is there a sweet spot of descriptor dimensionality after which the accuracy decreases? What is the memory footprint for inference/testing of both networks?

- For the real-world experiments, which data is actually used for training the framework? It seems the proposed synthetic data generation does not work (l. 228) ?
- Sec. 4.4 is hard to comprehend, for instance:
  - how is the descriptor picked for the grasping experiment?
  - what are "position and semantic object location offsets"?
  - l. 238 what does it mean to deactivate spatial probabilities?
  - l. 240 why do keypoints need to be projected into the image to determine the 6D pose of the object in the camer frame ?
  - l. 241 what defines the "aligned grasp" ?

Overall, the contribution of the paper is limited and the experiments are not convincing.

Further comments:

- add space between the images in Figs 4, 5.

Questions:

- see sec. weaknesses.

Limitations:
The paper does not provide an adequate discussion of limitations and assumptions.

Flag For Ethics Review: No ethics review needed.
Rating: 2: Strong Reject: For instance, a paper with major technical flaws, and/or poor evaluation, limited impact, poor reproducibility and mostly unaddressed ethical considerations.
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: Yes

Add: **Official Comment** **Rebuttal**

Official Review of Submission5130 by Reviewer 6xja

Official Review Reviewer 6xja 08 Jul 2023, 19:04 (modified: 01 Aug 2023, 21:17)
Program Chairs, Area Chairs, Authors, Reviewer 6xja, Reviewers Submitted, Senior Area Chairs

Summary:
The paper proposes an improvement to the Dense Object Network approach of [14] so as to decrease the training memory consumption when using more keypoints and also introducing a synthetic data generation procedure to reduce the need for image pairs with correspondence ground truth. The show that their proposed approach does not increase the memory consumption during training as the number of descriptor grows - and this allows them to get more accurate results.

Soundness: 2 fair
Presentation: 2 fair
Contribution: 1 poor
Strengths:
Downstream applications of dense descriptors to robotics are important and should be presented in NeurIPS. The proposed improvements to DON seem simple and practical.

Weaknesses:
Weakness 1: The main problem addressed by this contribution seems minor - there are multiple ways of reducing memory consumption (e.g. training on smaller batches, gradient checkpointing, distributing over multiple gpus, gradient accumulation etc). The authors do not make a convincing case that they have compared those options and they fail for their task. Also the synthetic dataset generation pipeline seems straightforward - the authors also do not examine what is the baseline performance that could be obtained by exploiting synthetic datasets that have been used to train dense descriptors that have been used for optical flow.

Weakness 2: From a computer vision standpoint it is hard to understand the merit of the present work when compared to the present state-of-the-art in self-supervised dense descriptor learning. There is no reference to the very rich ongoing work in this direction or comparison. Instead there is a very general introduction that is about robotics, LLMs, etc and the authors fail to clarify where they stand with respect to current works in self-supervised descriptor learning.

-Dense descriptors have been in use in computer since at least 2008: Fast Local Descriptor for Dense Matching, Tola et al, CVPR 2008 - a state of the art pre-deep learning can be found here: https://link.springer.com/book/10.1007/978-3-319-23048-1#toc (https://link.springer.com/book/10.1007/978-3-319-23048-1#toc)

There are multiple papers on understanding the use of self-supervised features as dense descriptors - citing a few here:

-DINOv2: Learning Robust Visual Features without Supervision, Oquab et al, 2023

- Emerging Properties in Self-Supervised Vision Transformers, (Dino V1), Caron et al 2021
- Deep ViT Features as Dense Visual Descriptors, Amir et al, 2021

One counter-argument could be that this is a robotics use case and not a computer vision one. Still, the robotics evaluation seems only qualitative. It is therefore hard to say whether this paper has moved the needle for the robotic application it aims to address. I do not think that in 2023 we can accept a paper that uses a standard vision architecture on the grounds of slightly improving a paper from 2018 in robotics.

Questions:
How would your metrics change if you used DinoV2 out of the box?

Limitations:
Yes

Flag For Ethics Review: No ethics review needed.
Rating: 3: Reject: For instance, a paper with technical flaws, weak evaluation, inadequate reproducibility and incompletely addressed ethical considerations.
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: Yes

Add: **Official Comment** **Rebuttal**

Official Review of Submission5130 by Reviewer msA6

Official Review Reviewer msA6 06 Jul 2023, 20:36 (modified: 01 Aug 2023, 21:17)
Program Chairs, Area Chairs, Authors, Reviewer msA6, Reviewers Submitted, Senior Area Chairs

Summary:
The paper presents an engineering effort i.e. mainly synthetic dataset creation and some loss function changes to an already established technique i.e. Dense Object Nets (DON) and claims to improve computational resources to train DON.

Soundness: 1 poor
Presentation: 1 poor
Contribution: 1 poor
Strengths:
The paper shows real robot experiments which are nice to have. The diagrams in the paper are good to supplement the text.

Weaknesses:

1. The paper claims to improve computational resources to train DON, which I think means using a lesser GPU memory, however the paper doesn't directly show comparison to DOn or similar approaches.
2. The paper only evaluates caps. Is there a reason for it? Does it generalize to other categories?
3. Limited technical contributions: The paper uses the exact same architecture and training methodology as DON and seemingly doing some data engineering (i.e. collecting some synthetic data) which the paper mentions a prior work doing that already. In essence, the experimental numbers the papers show are only their method and do not report DON or other similar work's computational numbers.
4. Is computational resources really an issue for DON? The paper shows 'max descriptor size' i.e. 512 can run at 7GB of memory which i think is reasonable enough given a laptop's GPU.
5. Overall, the flow of the paper looks more like a blog post than a technical paper. The authors should read high-quality papers and try to mimick them in their paper writing style to improve the flow. For instance, motivate the problem in the methodology section rather than jumping to what implementation or coding changes they did.

Questions:
Please see my remarks and questions above in the weakness section.

Limitations:
The paper only shows results on caps and I wonder if it generalizes to other more complicated objects (please see weakness section questions above)

Flag For Ethics Review: No ethics review needed.
Rating: 2: Strong Reject: For instance, a paper with major technical flaws, and/or poor evaluation, limited impact, poor reproducibility and mostly unaddressed ethical considerations.
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: Yes
First Time Reviewer: Yes

Add: **Official Comment** **Rebuttal**

Official Review of Submission5130 by Reviewer FtQk

Official Review Reviewer FtQk 06 Jul 2023, 19:02 (modified: 01 Aug 2023, 21:17)
Program Chairs, Area Chairs, Authors, Reviewer FtQk, Reviewers Submitted, Senior Area Chairs

Summary:
The work introduces an efficient framework for training Dense Object Nets (DON). The authors also introduce a novel synthetic data generation pipeline and discuss the use of their approach on a robotic grasping pipeline.

Soundness: 1 poor
Presentation: 1 poor
Contribution: 1 poor
Strengths:
The paper focuses on the critical challenges of current approaches to responsible and efficient use of computational resources.

Weaknesses:
I find the paper is not ready for publication for different reasons. Overall, the authors fail to convince on the real novelty and usefulness of their approach. The language needs strong revisions. Some parts are unclear (more details later in the questions) or too general to have a clear focus. The method is not clearly described and the procedure for collecting synthetic datasets is not fully convincing.

Questions:
I report here some questions that I hope may help to improve the quality of the paper:

- As far as I understand, the approach followed by the authors combines two already existing methods, i.e. the DON backbone trained on the KeypointNet task. I would state more clearly, while betterhighlighting any relevant difference
- The first part of the introduction is vague, ranging from AlphaGO, to ChatGPT to AlexNet. I'm not sure I can follow the flow of thoughts that lead the authors to the proposal
- "Existing representations may be unable to understand an object's geometrical and structural information, rendering them unsuitable for complex tasks" This is a strong statement, it would need appropriate citations. Indeed, current models for objects representations from visual data can describe them very accurately, so the invariances and the challenges of the considered application domain should be made more explicit
- "These dense visual object descriptors provide a generalized representation of objects to a certain extent. " Please clarify, this is important for highlighting the novelty of the proposed work
- At row 50 it is said that "the synthetic data generation pipeline does not rely on the noisy depth information produced by today's consumer-grade depth cameras" but indeed at row 134 depth information is actually employed for the data generation. Please clarify
- Apparently, the acquired dataset is rather "poor", with only one object. How this should guarantee the generalization of the approach to other objects?
- Row 134: "Using depth information, we project the computed correspondences to the camera frame..." Since the data are synthetically generated, you should have access to the full calibration parameters, why use the depth?

- I fail to get some of the essential elements of the method, I think that some more details justifying the role and objectives of each part would help clarify.
- Not always the paper is self-contained: for instance, when mentioning the loss functions (Sect. 3.3) it would be of help to briefly explain the role of each one of them
- Technical choices should be better justified. For instance: why res-net? or “We upsample the dense features from the identity layer (being identical to the last convolution layer in the backbone)...”. Why? Why not weight decay when training your method?
- Row 146: Who are the N key points mentioned here? The 24 involved in the matches? Why do you need the probabilities and the depth only for such points?
- Row 168: “...we extract dense visual object descriptors from the network and store one single descriptor of objects in a database manually for now” Unclear: how do you move from a set of visual descriptors to just one? Is this the right interpretation?
- Row 196: here you talk about 128 correspondences, while before they were 24. Please clarify
- Row 194 and Row 199: “...is computed with 256 image-pair correspondences for both models. The metrics mean and std. deviation is calculated from benchmarking three models trained for each descriptor dimension.” Please clarify who the models are
- On what datasets the results of Tables 1 and 2 have been computed?
- Row 221: what do you mean by multi-modal activations?
- Row 223: “As the synthetic data generation only needs mask and depth information, we could create a mask in no time.” Unclear, please clarify
- Figure 5: I’m not sure I understand. Why is the result consistent?
- There is no need to refer to the page other than a figure or a table (example: in Row 142)
- I'm not sure I understand the need of this new dataset (or better of this strategy to acquire the data). What are the benefits?
- For readability (and to favor the understanding of the connections with the related works) it would be beneficial to add in the introduction a quick idea of how the proposed approach works.
- Row 109 “...we do not use any loss functions...”, Row 117 “...Our approach encompasses... loss function modifications...”, These 2 statements seem partially in contrast, please clarify

Limitations:
Limitations are briefly discussed in the conclusions

Flag For Ethics Review: No ethics review needed.
Rating: 3: Reject: For instance, a paper with technical flaws, weak evaluation, inadequate reproducibility and incompletely addressed ethical considerations.
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: Yes

Add: **Official Comment** **Rebuttal**

Official Review of Submission5130 by Reviewer NQ4e

Official Review Reviewer NQ4e 27 Jun 2023, 11:26 (modified: 01 Aug 2023, 21:17)
Program Chairs, Area Chairs, Authors, Reviewer NQ4e, Reviewers Submitted, Senior Area Chairs

Summary:
In consideration of computationally expensive corresponding mapping in image pairs in training Dense Object Nets, the paper proposes a novel framework to address this issue and thus achieve robust and dense visual object descriptors. Specifically, the authors introduce a data generation procedure using synthetic augmentation and a learning-based framework to produce denser visual descriptors. The final result demonstrates the robustness of the generation of image descriptions.

Soundness: 2 fair
Presentation: 2 fair
Contribution: 2 fair
Strengths:

The proposed framework for training DON in a computationally efficient manner which aims to address the issue that is posed by the computationally intensive nature of DON. Moreover, the authors introduce a novel synthesis data generation procedure to generate a complete dataset that does not rely on the noisy depth information produced by cameras. The robotic grasp task has demonstrated the effectiveness of the proposed method.

Weaknesses:

1. More experiments should be explored. The proposed method is only evaluated on the robotic grasping application which is not enough to demonstrate its superior performance in addressing the computationally expensive corresponding mapping problem.
2. An ablation experiment is necessary to be implemented. For example, why the author chooses ResNet34 as the backbone rather than ResNet18 or ResNet 50.
3. The authors are obligated to show more details about Dense Object Network. In addition, are these training hyper-parameters cherry-picked?
4. The parameter settings are repetitive and confusing in Sec.4.1. Moreover, some typos should be corrected. For example, a peroid is necessary in Line 36 before “In recent work”. In otherwords should be rewritten as In other words in Line 144.
5. For statistical significance, the experiments should be conducted several times and the statistical significance of the results should be determined.

Questions:

See Weakness part

Limitations:

See Weakness part

Flag For Ethics Review: No ethics review needed.

Rating: 4: Borderline reject: Technically solid paper where reasons to reject, e.g., limited evaluation, outweigh reasons to accept, e.g., good evaluation. Please use sparingly.

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: Yes

First Time Reviewer: Yes

Add: **Official Comment** **Rebuttal**

About OpenReview (/about)
Hosting a Venue (/group?id=OpenReview.net/Support)
All Venues (/venues)
Sponsors (/sponsors)

Frequently Asked Questions (https://docs.openreview.net/getting-started/frequently-asked-questions)
Contact (/contact)
Feedback
Terms of Service (/legal/terms)
Privacy Policy (/legal/privacy)

OpenReview (/about) is a long-term project to advance science through improved peer review, with legal nonprofit status through Code for Science & Society (https://codeforscience.org/). We gratefully acknowledge the support of the OpenReview Sponsors (/sponsors).