Training Dense Object Nets: A Novel Approach

Anonymous Author(s)
Affiliation
Address
email

Abstract

We present a novel framework for mining dense visual object descriptors produced by Dense Object Nets (DON) without explicitly training DON. DON's dense visual object descriptors are robust to changes in viewpoint and configuration. However, training DON requires image pairs with correspondence mapping, which can be computationally expensive and limit the dimensionality and robustness of the descriptors, limiting object generalization. To overcome this, we propose a synthetic augmentation data generation procedure and a novel deep learning architecture that produces denser visual descriptors while consuming fewer computational resources. Furthermore, our framework does not require image-pair correspondence mapping and demonstrates its one of the applications as a robot-grasping pipeline. Experiments show that our approach produces descriptors as robust as DON.

1 Introduction

As of this writing, the ideal object representation for robot grasping and manipulation tasks is yet unknown. The existing representations may not be the best for tackling more complex tasks as they lack actual object information belonging to the same class and configuration (shape, color and size). In industrial robot-based automation, the objects are specifically coded for their visual features using 2D and 3D vision systems. The downside of this lies in the fact that the robot has to be taught to pick every other part with its visual representation. This process comes with the tedious schedule of teaching the robot to pick every part irrespective of the part's configuration, and viewpoint. The solution lies in using artificial intelligence (AI) equipped robots. A deep learning neural network (DNN) is based on artificial neurons capable to learning a task and is good as the task related data it is trained on. The data used to train DNN is often expensive as it requires engineered features that DNN can predict or regress. SIFT [1], SURF [2] and ORB [3] produce dense local descriptors of an object in an image and serve as target features to train DNN to yield object representation for robot grasping furthermore, these features computed by [1, 2, 3] come with its own inert limitations and cannot generalize objects well. Our interests of work is on reducing efforts to develop hand engineered features to train DNN and developing DNN that can generalize plathora of objects such that we spend less time teaching robot how to tend objects in realtime.

In 2018, Florence et al. [4] introduced a novel visual object representation to the robotics community, terming it "dense visual object descriptors". DON, an aritificial intelligence framework proposed by Florence et al. [4] produces dense visual object descriptors. In detail, the DON converts every pixel in the image ($I[u, v] \in \mathbb{R}^3$) to a higher dimensional embedding ($I_D[u, v] \in \mathbb{R}^D$) such that $D \in \mathbb{N}^+$ which are nothing but dense local descriptors of that pixel respective to the image. The dense visual object descriptor generalize an object up to a certain extent and have been recently applied to rope manipulation [5], block manipulation [6], robot control [7], fabric manipulation [8] and robot grasp pose estimation [9, 10]. Suwajanakorn et al. [11] propose self-supervised geometrically consistent keypoints, exploring the idea of optimizing a representation based on a sparse collection of keypoints or landmarks, but without access to keypoint annotations. The authors of [11] devise an end-to-end geometric reasoning framework first introduced by [12] to regresses a set of geometrically consistent keypoints coined as KeypointNet. This means that KeypointNet is capable of generalizing objects without the need of hand engineered features. Suwajanakorn et al. [11] show that using two unique objective loss functions, namely, a relative pose estimation loss and a multi-view consistency goal, uncovers the consistent keypoints across multiple views and object instances. Their affine translation-equivariant design may extend to previously unknown object instances trained on ShapeNet [13] dataset.

At first, we present modifications to the DNN inspired from [4] and [11] such that we seemlessly train and mine object representations composed of object generalizing dense local descriptors while training for KeypointNet task. Second, we develop synthetic dataset using [14] to train the DNN and prove that the mined dense local descriptors from our framework is as robust as dense visual object descriptors produced from DON while consuming less computation resources. Additionally, we demonstrate an self-supervised framework to train DON with semantically equivalent objects which is not previously demonstrated in [4, 15, 9, 10, 16, 17] to train DON.

References

[1] D. G. Lowe. “Object recognition from local scale-invariant features”. In: *Proceedings of the seventh IEEE international conference on computer vision*. Vol. 2. Ieee. 1999, pp. 1150–1157.

[2] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. “Speeded-up robust features (SURF)”. In: *Computer vision and image understanding* 110.3 (2008), pp. 346–359.

[3] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. “ORB: An efficient alternative to SIFT or SURF”. In: *2011 International conference on computer vision*. Ieee. 2011, pp. 2564–2571.

[4] P. R. Florence, L. Manuelli, and R. Tedrake. “Dense object nets: Learning dense visual object descriptors by and for robotic manipulation”. In: *arXiv preprint arXiv:1806.08756* (2018).

[5] P. Sundaresan, J. Grannen, B. Thananjeyan, A. Balakrishna, M. Laskey, K. Stone, J. E. Gonzalez, and K. Goldberg. “Learning Rope Manipulation Policies Using Dense Object Descriptors Trained on Synthetic Depth Data”. In: *CoRR* abs/2003.01835 (2020). arXiv: `2003.01835`.

[6] C.-Y. Chai, K.-F. Hsu, and S.-L. Tsao. “Multi-step Pick-and-Place Tasks Using Object-centric Dense Correspondences”. In: *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2019, pp. 4004–4011.

[7] P. Florence, L. Manuelli, and R. Tedrake. “Self-supervised correspondence in visuomotor policy learning”. In: *IEEE Robotics and Automation Letters* 5.2 (2019), pp. 492–499.

[8] A. Ganapathi et al. “Learning Dense Visual Correspondences in Simulation to Smooth and Fold Real Fabrics”. In: *2021 IEEE International Conference on Robotics and Automation (ICRA)*. 2021, pp. 11515–11522.

[9] A. Kupcsik, M. Spies, A. Klein, M. Todescato, N. Waniek, P. Schillinger, and M. Bürger. “Supervised Training of Dense Object Nets using Optimal Descriptors for Industrial Robotic Applications”. In: *arXiv preprint arXiv:2102.08096* (2021).

[10] D. B. Adrian, A. G. Kupcsik, M. Spies, and H. Neumann. “Efficient and Robust Training of Dense Object Nets for Multi-Object Robot Manipulation”. In: *2022 International Conference on Robotics and Automation (ICRA)*. IEEE. 2022, pp. 1562–1568.

[11] S. Suwajanakorn, N. Snavely, J. J. Tompson, and M. Norouzi. “Discovery of latent 3d keypoints via end-to-end geometric reasoning”. In: *Advances in neural information processing systems* 31 (2018).

[12] S. Levine, C. Finn, T. Darrell, and P. Abbeel. “End-to-end training of deep visuomotor policies”. In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 1334–1373.

[13] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, et al. “Shapenet: An information-rich 3d model repository”. In: *arXiv preprint arXiv:1512.03012* (2015).

[14] M. Denninger, M. Sundermeyer, D. Winkelbauer, Y. Zidan, D. Olefir, M. Elbadrawy, A. Lodhi, and H. Katam. “Blenderproc”. In: *arXiv preprint arXiv:1911.01911* (2019).

[15] P. R. Florence. “Dense visual learning for robot manipulation”. PhD thesis. Massachusetts Institute of Technology, 2020.

[16] D. Hadjivelichkov and D. Kanoulas. "Fully Self-Supervised Class Awareness in Dense Object Descriptors". In: *5th Annual Conference on Robot Learning*. 2021.

[17] L. Yen-Chen, P. Florence, J. T. Barron, T.-Y. Lin, A. Rodriguez, and P. Isola. *NeRF-Supervision: Learning Dense Object Descriptors from Neural Radiance Fields*. 2022.