HoughCL: Finding Better Positive Pairs in Dense Self-supervised Learning

Yunsung Lee1 2 3 Teakgyu Hong4 Han-Cheol Cho4 Junbum Cha4 Seungryong Kim2

Abstract

Recently, self-supervised methods show remarkable achievements in image-level representation learning. Nevertheless, their image-level self-supervisions lead the learned representation to sub-optimal for dense prediction tasks, such as object detection, instance segmentation, etc. To tackle this issue, several recent self-supervised learning methods have extended image-level single embedding to pixel-level dense embeddings. Unlike image-level representation learning, due to the spatial deformation of augmentation, it is difficult to sample pixel-level positive pairs. Previous studies have sampled pixel-level positive pairs using the winner-takes-all among similarity or thresholding warped distance between dense embeddings. However, these naive methods can be struggled by background clutter and outliers problems. In this paper, we introduce Hough Contrastive Learning (HoughCL), a Hough space based method that enforces geometric consistency between two dense features. HoughCL achieves robustness against background clutter and outliers. Furthermore, compared to baseline, our dense positive pairing method has no additional learnable parameters and has a small extra computation cost. Compared to previous works, our method shows better or comparable performance on dense prediction fine-tuning tasks.

1. Introduction

Recent self-supervised visual representation learning methods have made significant progress in image recognition since InfoNCE (Oord et al., 2018) based contrastive representation learning. Most of recent self-supervised visual representation learning (Chen et al., 2020a; He et al., 2020; Chen et al., 2020b; Caron et al., 2018; 2020; Grill et al., 2020; Chen & He, 2021; Ermolov et al., 2020; Zbontar et al., 2021; Bardes et al., 2021) have in common with maximizing the agreement between positive pairs of embedding vectors that are sampled from different views of the same image. However, most of these self-supervised methods only consider image-level embeddings lacking local information. It may be appropriate for image-level recognition tasks but can be sub-optimal for dense prediction tasks such as object detection or semantic segmentation.

Several recent works (Wang et al., 2021; Xie et al., 2021; Roh et al., 2021) have learned representations from pixel-level densely embedded vectors, and show improvements when transferring to downstream dense prediction tasks. In image-level self-supervised learning, positive pairs were easily assigned, because image-level features are invariant in data augmentation. However, since pixel-level features are variant in augmentation, it is difficult to assign pixel-level positive pairs. In DenseCL (Wang et al., 2021), they introduce a dense projection head that outputs dense feature vectors. To obtain positive pixel pairs, they simply calculate the cosine similarity between pixel vectors and choose the positive pair which has the highest similarity value. This simple winner-takes-all method suffers from background clutter and outliers. Meanwhile in PixPro (Xie et al., 2021), in addition to the dense projection head, an asymmetric network is introduced that computes its smoothed transform by propagating pixel-level features. Their assignments of dense positive pairs differ from DenseCL. Each point in a feature map is first warped to the original image space, and the distances between all pairs of points from the two feature maps are computed. By thresholding these distances, they assigned dense positive pairs. Though this method can simply find a positive pixel pair, there is a risk of semantically different pixels paired as positive because all pixels located close to each other are treated as positive. In addition, the threshold value is a hyper-parameter that requires a new manual setting.

In this paper, we introduce the pixel-level dense positive pairing method based on Hough geometric voting, inspired by the algorithm of Cho et al. (2015). Through weighted voting in Hough space, we can obtain geometrical consistent dense positive pairs. This geometric consistency gives a model robustness against background clutter and outliers. Furthermore, it does not require additional training parameters.
Thus, our method is generally applicable to self-supervised learning methods where matching of dense positive pairs exists. Compared to previous works, our method shows better or comparable performance on dense prediction fine-tuning tasks. In particular, experimental results of pre-training on Tiny ImageNet, a miniature of ImageNet, our method outperforms the DenseCL when transferring to downstream dense prediction tasks, including PASCAL VOC object detection (+3.0% AP), COCO object detection (+1.1% AP) and COCO instance segmentation (+0.9% AP).

2. Background

After self-supervised learning has become a new paradigm for pre-training in image-level recognition tasks, several recent self-supervised learning methods are designed for dense prediction tasks. DenseCL (Wang et al., 2021) introduces a dense projection head that outputs dense feature vectors, and it is a simple but effective way to learn a pixel-level dense representation. In this section, we briefly review our baseline method, DenseCL.

2.1. DenseCL: Dense Contrastive Learning

Compared to the existing paradigm, the core differences of DenseCL lie in the encoder and loss function. Given an input view, the dense feature maps are extracted by the backbone network, e.g., ResNet (He et al., 2016), and forwarded to the following projection head. The projection head consists of two sub-heads in parallel, which are global projection head and dense projection head, respectively. The global projection head can be any of the existing projection heads in He et al. (2020); Chen et al. (2020a;b), which takes the dense feature maps as input and generates a global feature vector for each view. In contrast, the dense projection head takes the same input but outputs dense feature vectors. The backbone and two parallel projection heads are end-to-end trained by optimizing a joint pairwise loss at the levels of both global features and local features.

In DenseCL, dense contrastive loss is the extension of the original contrastive loss function to a dense paradigm. \{\theta_0, \theta_1, \ldots\} is a set of encoded keys for each encoded query \(r\). \(r\) corresponds to one of the \(S_h \times S_w\) (for simpler illustration, they use \(S_h = S_w = S\)) feature vectors generated by the dense projection head. Each negative key \(t_{-}\) is the pooled feature vector of a view from a different image. The positive key \(t_{+}\) is assigned according to the extracted correspondence across views, which is one of the \(S^2\) feature vectors from another view of the same image. The dense contrastive loss is defined as:

\[
\mathcal{L}_r = \frac{1}{2} \sum_{i} - \log \left( \frac{\exp(\mathbf{r}^\top \mathbf{t}^\star_{+} / \tau)}{\sum_{e} \exp(\mathbf{r}^\top \mathbf{t}^\star_{e} / \tau)} \right),
\]

where \(\mathbf{r}^\star\) denotes the \(e\)th out of \(S^2\) encoded queries.

Overall, the total loss for DenseCL can be formulated as

\[
\mathcal{L} = (1 - \lambda) \mathcal{L}_q + \lambda \mathcal{L}_r,
\]

where \(\mathcal{L}_q\) is existing image-level contrastive loss (Oord et al., 2018). \(\lambda\) is set to 0.5 which is validated by experiments in Wang et al. (2021).

2.2. Dense Positive Pairs in DenseCL

In DenseCL, the dense correspondence between the two views of the same input image is the dense positive pairs \(t_{+}\) described in 2.1. For each view, the backbone network extracts feature maps \(\mathbf{F} \in \mathbb{R}^{H \times W \times K}\), and the dense projection head extracts dense feature vectors \(\Theta \in \mathbb{R}^{S_h \times S_w \times E}\). The correspondence between the dense feature vectors from the two views, \(\Theta_1\) and \(\Theta_2\), is made using the backbone feature maps \(\mathbf{F}_1\) and \(\mathbf{F}_2\). The \(\mathbf{F}_1\) and \(\mathbf{F}_2\) are first downsampled to have the spatial shape of \(S \times S\) by an adaptive average pooling, and then used to calculate the cosine similarity matrix \(\Delta \in \mathbb{R}^{S^2 \times S^2}\). The matching rule is winner-takes-all in feature vector similarity. The matching process can be formulated as \(c_i = \arg \max_j \text{sim}(\mathbf{f}_i, \mathbf{f}_j')\), where \(f_i\) is the \(i\)th feature vector of \(\mathbf{F}_1\), and \(f_j'\) is the \(j\)th of \(\mathbf{F}_2\), and \(\text{sim}(u, v)\) denotes the cosine similarity. It means that the positive pair of \(i\)th feature vector of \(\Theta_1\) is \(c_i\)th of \(\Theta_2\).

3. Hough Contrastive Learning

The Hough transform (Hough, 1962) is a classic method developed to identify primitive shapes in an image via geometric voting in a parameter space. In geometric matching, Cho et al. (2015) first extends it to the Probabilistic Hough Matching (PHM) algorithm for unsupervised object discovery. Recent semantic alignment and correspondence methods (Min et al., 2019; 2020; Liu et al., 2020; Min & Cho, 2021) employ Hough matching. Through Hough matching, these methods can form matches considering geometric consistency as well as appearance similarity.

As can be seen in 1, previous dense positive sample pairing methods, e.g., \(\text{argmax}\) in DenseCL (Wang et al., 2021) and thresholding warped distance in PixPro (Xie et al., 2021), may suffer from background clutter and outliers. These mismatches can give poor guidance information for the model to learn dense representations. To give the model robust guidance information against background clutter and outliers, we introduce geometric consistent dense positive pairing with PHM. The key idea of PHM is to re-weight appearance similarity by Hough space voting to enforce geometric consistency. By applying the PHM principle, we propose dense positive matching method that maintains more geometrical tendencies. In this paper, our method was applied to the baseline DenseCL, but it can be generally applied to self-supervised learning methods that use dense positive pairs.

In our context, let \(\mathcal{D} = (\mathcal{H}, \mathcal{H}')\) be two sets of dense pro-
HoughCL: Finding Better Positive Pairs in Dense Self-supervised Learning

4. Experiments

4.1. Pre-training Setup

To validate the performance of our method on various datasets, we conduct experiments on not only COCO and ImageNet, which are mainly used in the other methods, but also Tiny ImageNet, which is a relatively small dataset. COCO (Lin et al., 2014) consists of about 118K training images which containing common objects in complex everyday scenes. ImageNet (Deng et al., 2009) consists of about 1.28M training images in 1K image classes. Tiny ImageNet (Le & Yang, 2015) is a miniature of ImageNet. It consists of 100K training images of size 64x64 in 200 image classes.

The pre-training setup mostly follows DenseCL (Wang et al., 2021). A ResNet-50 (He et al., 2016) is adopted as a backbone. SGD optimizer is utilized and its weight decay and momentum are set to 0.0001 and 0.9, respectively. The initial learning rates are set to 0.5, 0.3, and 0.03 in Tiny ImageNet, COCO, and ImageNet, respectively and cosine annealing schedule is used. The batch size is set to 256, using 8 V100 GPUs. The number of training epochs are set to 200, 800, and 200 in Tiny ImageNet, COCO, and ImageNet, respectively.

4.2. Fine-tuning Setup

We evaluate the pre-trained model on three downstream dense prediction tasks: PASCAL VOC object detection (Everingham et al., 2010), and COCO object detection and instance segmentation (Lin et al., 2014). The fine-tuning setup follows DenseCL.
Table 1. Experimental results of PASCAL VOC object detection. A Faster R-CNN (C4-backbone) is fine-tuned on trainval07+12 set for 24K iterations and evaluated on test2007 set. The results are averaged over 2 independent trials. † indicates the scores are reported from (Wang et al., 2021)

| Dataset | Method   | AP   | AP50 | AP75 |
|---------|----------|------|------|------|
| Tiny    | random init. † | 32.8 | 59.0 | 31.6 |
| ImageNet| MoCo v2   | 47.6 | 75.3 | 51.2 |
|         | DenseCL   | 47.5 | 74.6 | 51.2 |
|         | HoughCL   | 50.5 | 76.9 | 55.0 |
| COCO    | MoCo v2†  | 54.7 | 81.0 | 60.6 |
|         | DenseCL†  | 56.7 | 81.7 | 63.0 |
|         | HoughCL   | 56.8 | 82.1 | 63.0 |
| ImageNet| super. IN† | 54.2 | 81.6 | 59.8 |
|         | MoCo v2†  | 57.0 | 82.2 | 63.4 |
|         | DenseCL†  | 58.7 | 82.8 | 65.2 |
|         | HoughCL   | 58.5 | 82.6 | 65.7 |

4.3. Results

Table 1 shows the experimental results of PASCAL VOC object detection. HoughCL outperforms the other methods in Tiny ImageNet and shows similar performance when pre-trained on COCO and ImageNet. In Tiny ImageNet, HoughCL achieves 3.0% AP and 3.8% AP75 improvements compared to DenseCL. This result indicates the efficiency of HoughCL by showing superior performance in relatively small scale dataset. In COCO and ImageNet, HoughCL shows similar AP scores compared to DenseCL, but it achieves slightly better AP75 scores in ImageNet.

Table 2. Experimental results of COCO object detection. A Mask R-CNN detector (FPN backbone) is fine-tuning on train2017 split with 1× schedule and evaluated on val2017 split. The results are averaged over 2 independent trials. † indicates the scores are reported from (Wang et al., 2021)

| Dataset | Method   | AP   | AP50 | AP75 |
|---------|----------|------|------|------|
| Tiny    | random init. † | 32.8 | 59.0 | 35.3 |
| ImageNet| MoCo v2   | 35.6 | 54.6 | 38.8 |
|         | DenseCL   | 35.4 | 54.0 | 38.6 |
|         | HoughCL   | 36.5 | 55.4 | 39.9 |
| COCO    | MoCo v2†  | 38.5 | 58.1 | 42.1 |
|         | DenseCL†  | 39.6 | 59.3 | 43.3 |
|         | HoughCL   | 39.5 | 59.3 | 43.1 |
| ImageNet| super. IN† | 39.7 | 59.5 | 43.3 |
|         | MoCo v2†  | 39.8 | 59.8 | 43.6 |
|         | DenseCL†  | 40.3 | 59.9 | 44.3 |
|         | HoughCL   | 40.0 | 59.9 | 43.6 |

Table 2 and Table 3 show the experimental results of COCO object detection and instance segmentation. Similar to the results of PASCAL VOC, HoughCL shows superior performances in Tiny ImageNet. It outperforms DenseCL by 1.1% AP and 1.3% AP75 in object detection and 0.9% AP and 1.0% AP75 in instance segmentation. In COCO and ImageNet, HoughCL achieves similar or slightly lower scores than other methods.

Overall, HoughCL shows superior performance in Tiny ImageNet, but similar performance in COCO and ImageNet. We think this is because HoughCL is not yet optimized for COCO and ImageNet.

Table 3. Experimental results of COCO instance segmentation. A Mask R-CNN detector (FPN backbone) is fine-tuning on train2017 split with 1× schedule and evaluated on val2017 split. The results are averaged over 2 independent trials. † indicates the scores are reported from (Wang et al., 2021)

| Dataset | Method   | AP   | AP50 | AP75 |
|---------|----------|------|------|------|
| Tiny    | random init. † | 29.9 | 47.9 | 32.0 |
| ImageNet| MoCo v2   | 32.5 | 51.8 | 34.9 |
|         | DenseCL   | 32.2 | 51.3 | 34.5 |
|         | HoughCL   | 33.1 | 52.6 | 35.5 |
| COCO    | MoCo v2†  | 34.8 | 55.3 | 37.3 |
|         | DenseCL†  | 35.7 | 56.5 | 38.4 |
|         | HoughCL   | 35.7 | 56.4 | 38.2 |
| ImageNet| super. IN† | 35.9 | 56.6 | 38.6 |
|         | MoCo v2†  | 36.1 | 56.9 | 38.7 |
|         | DenseCL†  | 36.4 | 57.0 | 39.2 |
|         | HoughCL   | 36.2 | 56.8 | 38.8 |

5. Conclusion

In this paper, we introduce a novel dense positive sample pairing method based on Hough geometric voting. Proposed method provides geometrically consistent dense positive pairs through weighted voting in Hough space. This geometric consistency gives a model robustness against background clutter and outliers. Experimental results show that HoughCL outperforms baselines especially in Tiny ImageNet, which consists of downscaled images from ImageNet. It empirically demonstrates HoughCL matches dense positive pair robustly in noisy setting, e.g., downsampling noise.

On the other hand, our method performs similar to the baselines on ImageNet and COCO datasets. Although this work improves robustness in dense representation learning, we believe that the geometrical consistency has the potential to improve the performance even on less noisy datasets, such as ImageNet and COCO. We will cover this topic in the future work.
References

Bardes, A., Ponce, J., and LeCun, Y. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. arXiv preprint arXiv:2105.04906, 2021.

Caron, M., Bojanowski, P., Joulin, A., and Douze, M. Deep clustering for unsupervised learning of visual features. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 132–149, 2018.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), Advances in Neural Information Processing Systems, volume 33, pp. 9912–9924, 2020.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In International conference on machine learning, pp. 1597–1607. PMLR, 2020a.

Chen, X. and He, K. Exploring simple siamese representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15750–15758, 2021.

Chen, X., Fan, H., Girshick, R., and He, K. Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297, 2020b.

Cho, M., Kwak, S., Schmid, C., and Ponce, J. Unsupervised object discovery and localization in the wild. Part-based matching with bottom-up region proposals. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1201–1210, 2015.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pp. 248–255. Ieee, 2009.

Ermolov, A., Siarohi, A., Sangineto, E., and Sebe, N. Whitening for self-supervised representation learning. arXiv preprint arXiv:2007.06346, 2020.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. The pascal visual object classes (voc) challenge. International journal of computer vision, 88 (2):303–338, 2010.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., Piot, B., kavukcuoglu, k., Munos, R., and Valko, M. Bootstrap your own latent - a new approach to self-supervised learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), Advances in Neural Information Processing Systems, volume 33, pp. 21271–21284, 2020.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778, 2016.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9729–9738, 2020.

Hough, P. V. Method and means for recognizing complex patterns, December 18 1962. US Patent 3,069,654.

Le, Y. and Yang, X. Tiny imagenet visual recognition challenge. CS 231N, 7:7, 2015.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In European conference on computer vision, pp. 740–755. Springer, 2014.

Liu, Y., Zhu, L., Yamada, M., and Yang, Y. Semantic correspondence as an optimal transport problem. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4463–4472, 2020.

Min, J. and Cho, M. Convolutional hough matching networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2940–2950, 2021.

Min, J., Lee, J., Ponce, J., and Cho, M. Hyperpixel flow: Semantic correspondence with multi-layer neural features. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3395–3404, 2019.

Min, J., Lee, J., Ponce, J., and Cho, M. Learning to compose hypercolumns for visual correspondence. In ECCV 2020-16th European Conference on Computer Vision. Springer, 2020.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748, 2018.

Roh, B., Shin, W., Kim, I., and Kim, S. Spatially consistent representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1144–1153, 2021.
Xie, Z., Lin, Y., Zhang, Z., Cao, Y., Lin, S., and Hu, H. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16684–16693, 2021.

Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. Barlow twins: Self-supervised learning via redundancy reduction. *arXiv preprint arXiv:2103.03230*, 2021.