Correlation between Alignment-Uniformity and Performance of Dense Contrastive Representations

Jong Hak Moon Student[1]
jhak.moon@kaist.ac.kr
Wonjae Kim Collaborator[2]
wonjae.kim@navercorp.com
Edward Choi Prof[1]
edwardchoi@kaist.ac.kr

[1] KAIST, Daejeon, South Korea
[2] Naver AI, Sungnam, South Korea

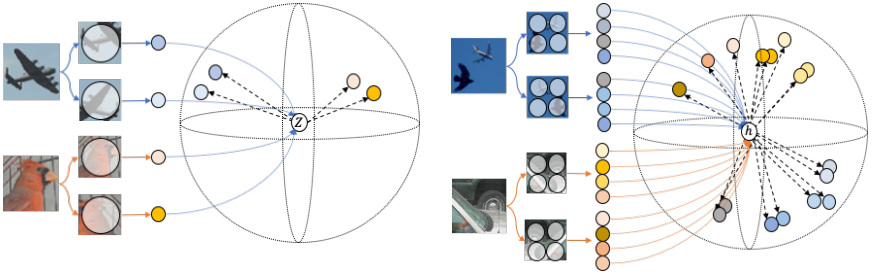
Abstract

Recently, dense contrastive learning has shown superior performance on dense prediction tasks compared to instance-level contrastive learning. Despite its supremacy, the properties of dense contrastive representations have not yet been carefully studied. Therefore, we analyze the theoretical ideas of dense contrastive learning using a standard CNN and straightforward feature matching scheme rather than propose a new complex method. Inspired by the analysis of the properties of instance-level contrastive representations through the lens of alignment and uniformity on the hypersphere, we employ and extend the same lens for the dense contrastive representations to analyze their underexplored properties. We discover the core principle in constructing a positive pair of dense features and empirically proved its validity. Also, we introduces a new scalar metric that summarizes the correlation between alignment-and-uniformity and downstream performance. Using this metric, we study various facets of densely learned contrastive representations such as how the correlation changes over single- and multi-object datasets or linear evaluation and dense prediction tasks.

1 Introduction

Instance-level CL (Contrastive Learning) with a single-object dataset (*e.g.* ImageNet [10]) [4, 5, 6, 12, 16, 29] has shown to be highly effective for learning visual representations in a self-supervised manner. To understand the semantic structures and behavior of this method, a few recent studies [7, 30] analyzed the latent space (*e.g.* unit hypersphere) from the perspective of uniformity and alignment (closeness). Intuitively, it is effective to analyze from these two perspectives, since features of all classes can be linearly separated from the rest of the feature space if they are sufficiently well clustered.

Although instance-level contrastive features have been successful in improving image classification performance, it has been observed that they do not enjoy the same transferability to dense prediction tasks (*e.g.* object detection tasks) [5, 6, 11, 17, 31, 32, 35, 36]. Since the receptive field of global averaged pooled features typically extends to the entire image,



(a) Instance-level CL (b) Dense CL

Figure 1: **Contrastive Representations on the hypersphere.** We demonstrate the difference in feature representation between instance- and dense CL on (a) single-object and (b) multi-object datasets. (a) represents an image as a single feature vector $\mathbf{z} \in \mathbf{R}^d$ containing global information, whereas (b) represents a set of vectors $\mathbf{h} \in \mathbf{R}^{d \times HW}$ exploited from a $H \times W$ feature map containing local feature information.

the pooled features are affected by background information, making it difficult to localize. To overcome this gap, recent studies [17, 31, 32, 35] have developed dense CL with multi-object datasets (*e.g.* MS-COCO [14]), using dense features to explicitly consider spatial information over regions and achieved comparable or better results compared to supervised ImageNet pre-training. Despite such initial success, these works beg an important yet unexplored question: "How different are the dense-level features compared to the instance-level features?" (Fig. 1) In this work, we investigate the dense feature representation in terms of alignment and uniformity inspired by the pioneering analyses of [7, 30]. We extend the conventional contrastive loss (InfoNCE [18]) to construct a more principled dense-level contrastive loss, and introduce a scalar metric to succinctly report the alignment-uniformity behavior of latent features. Based on extensive experiments and analysis using both single and multi-object pre-training datasets, and instance-level (*i.e.* linear evaluation) and dense downstream task (*i.e.* object detection), our findings and contributions can be summarized as follows:

- We empirically show that the alignment-uniformity property in dense features is correlated with both instance-level and dense-level downstream task performance.
- We find that, contrary to our belief, instance-level contrastive features pre-trained on multi-object dataset can perform well on object detection, and dense contrastive features pre-trained on single-object dataset can perform well on linear evaluation, both cases following the alignment-uniformity principle.
- We discover the core principle in constructing a positive pair of dense features and empirically proved its validity with a simple index-wise matching.

2 Related work

After the advent of SimCLR [6], unsupervised CL (contrastive learning) was explosively researched on the instance-level [4, 5, 12, 16, 29]. The core idea of this approach is sharing the InfoMax [15] principle under instantiation by maximizing mutual information between two transformed versions of the same image [2, 25, 33]. Recently Wang and Isola [30] empirically proved that a unit $l2$-norm constrained contrastive loss (InfoNCE [18]) can be decomposed into a metric of alignment ($l2$-distance) and uniformity (average pairwise Gaussian potential). Also, they proved that optimizing contrastive loss is equivalent to optimizing the

alignment between positive pairs and maintaining uniformity across all feature vectors in the hypersphere, and observed optimizing the alignment-uniformity properties is closely related to the downstream task performance such as linear evaluation. This hypersphere uniform distribution was generalized by Chen et al. [7] and extended to a wider set of prior distributions (*e.g.* uniform hypercube or Normal distribution). Our study is more related to Wang and Isola [30], and we extend this analysis to dense features that contain local spatial information.

Recently, He et al. [11], Sun et al. [23], Tan et al. [24] demonstrate a transfer learning gap between instance-level pre-training and dense prediction tasks such as object detection. In an effort to overcome this gap, several works [17, 31, 32, 35] generalized the instance discrimination from image-level to pixel-level to explore dense-level unsupervised CL and demonstrated improved downstream performance for dense prediction tasks. In contrast to numerous theoretical [1, 13, 20, 27, 28] and empirical analyses [7, 19, 22, 26, 30, 34, 37] to understand instance-level CL, no attempt has been made to understand dense CL. While there are many open questions, in this work we analyze how the pre-training impacts downstream tasks by extending the instance-level contrastive loss to the dense-level paradigm.

Additionally, unlike instance-level CL where positive pairs are easily constructed via augmentations, constructing positive dense feature pairs in dense CL is non-trivial. Each of the previous works devised its own strategy to solve this problem, such as calculating the cosine similarity between dense features [31], attention-based set-wise matching [32], and matching dense features with associated regions [17, 35]. In this work, we take a more straightforward approach and adopt an index-wise matching between dense features from two augmented views. In the experiments section, we compare this rather simple strategy with more sophisticated ones such as using cosine similarity or optimal transport, and report that our approach leads to comparable or better downstream performance. Furthermore, we analyze the effectiveness of the index-wise pairing strategy in terms of whether the pre-training dataset consists of single-object images or multi-object images.

3 Method

3.1 Preliminary: Instance-level Contrastive Loss

Instance-level CL can be seen as the lower bound of mutual information (MI) between a positive pair x and y [2, 18, 33]. Given $MI(x,y) = H(x) - H(x|y)$, the two right-hand side terms can be linked to the following two properties [7, 30]:

* Uniformity $H(x)$: Maximizing entropy leads to uniformly distributed latent vectors.
* Alignment $H(x|y)$: Minimizing conditional entropy given the positive pair of each item makes them be aligned in the latent space.

Note that the general form of contrastive loss is defined as follows,

$$L^{InsCont} = -\frac{1}{\mathcal{N}} \sum_{i,j \overset{\text{i.i.d}}{\sim} \mathcal{B}} \log \frac{e^{sim(\mathbf{z}_i,\mathbf{z}_j)/\lambda}}{\sum_{k \overset{\text{i.i.d}}{\sim} 2\mathcal{N}} \mathbb{1}_{[k \neq i]} e^{sim(\mathbf{z}_i,\mathbf{z}_k)/\lambda}}, \quad sim(\mathbf{x},\mathbf{y}) = \frac{\mathbf{x}\cdot\mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|} \tag{1}$$

where $\mathcal{N}$ denotes the number of randomly drawn instances, $\mathcal{B}$ the minibatch, $\mathbf{z}_i$ and $\mathbf{z}_j$ the positive pair of instance-level latent vectors projected into a hypersphere, λ the temperature, and $\mathbb{1}_{[k\neq i]\in 0,1}$ an indicator function. Eq. (1) can be rewritten as follows by applying

logarithmic rules:

$$L^{InsCont} = -\frac{1}{\mathcal{N}} \sum_{i,j \overset{\text{i.i.d}}{\sim} \mathcal{B}} (sim(\mathbf{z}_i, \mathbf{z}_j)/\lambda - \log(\sum_{k \overset{\text{i.i.d}}{\sim} 2\mathcal{N}} \mathbb{1}_{[k \neq i]} e^{sim(\mathbf{z}_i, \mathbf{z}_k)/\lambda})$$

$$= \underbrace{-\frac{1}{\mathcal{N}} \sum_{i,j \overset{\text{i.i.d}}{\sim} \mathcal{B}} sim(\mathbf{z}_i, \mathbf{z}_j)/\lambda}_{\text{alignment property}} + \underbrace{\frac{1}{\mathcal{N}} \sum_{i}^{\mathcal{N}} \log(\sum_{k \overset{\text{i.i.d}}{\sim} 2\mathcal{N}} \mathbb{1}_{[k \neq i]} e^{sim(\mathbf{z}_i, \mathbf{z}_k)/\lambda})}_{\text{distribution to be uniform property}}$$

where we confirm that the contrastive loss indeed consists of two objectives.

3.2 Dense Contrastive Loss

In order to analyze the behavior of dense features in CL, we first formalize the dense CL objective, a natural extension of instance-level CL to the dense-level.

Let f be a CNN encoder that transforms an input image x to dense feature vectors $\mathbf{h} = f(x) = \{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_{HW}\}$, $\mathbf{h}_i \in \mathbb{R}^d$, where HW is the spatial dimension size. Following the principle of *MI* maximization in Eq. (1), we assume that all $\mathbf{h}_i$'s in a single image are *i.i.d*. Although $\mathbf{h}_i$'s do share some global information, this assumption is based on the fact that the values of each $\mathbf{h}_i$ are not identical because each contains different spatial information. Also, this assumption is often implicitly seen in the previous dense CL studies to extract the corresponding feature. In particular, DenseCL[31] compares all individual cosine similarity scores of features and pulls the most similar pairs closer. Also, Setsim[32] matches the corresponding feature set by calculating the set similarity using the attention score of the individual features. Therefore, by following the implicit *i.i.d* assumption of the latest studies above, we perform index-wise feature matching by assuming *i.i.d* of the output feature to form positive and negative pairs. Dense contrastive loss can be defined as follows:

$$L^{DenseCont} = -\frac{1}{\mathcal{N}} \sum_{i,j \overset{\text{i.i.d}}{\sim} \mathcal{B}} \frac{1}{HW} \sum_{p}^{HW} \log \frac{e^{sim(\mathbf{h}_{(i,p)}, \mathbf{h}_{(j,p)})/\lambda}}{\sum_{k \overset{\text{i.i.d}}{\sim} 2\mathcal{N}} \sum_{q}^{HW} \mathbb{1}_{[k \neq i] \times [\frac{q \neq p}{k=j}]} e^{sim(\mathbf{h}_{(i,p)}, \mathbf{h}_{(k,q)})/\lambda}}, \quad sim(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|} \tag{2}$$

where $\mathbf{h}_{(i,p)}$ indicates p-th dense feature of the i-th sample, and $\mathbb{1}_{[k \neq i] \times [\frac{q \neq p}{k=j}] \in 0,1}$ an indicator function. Note that a positive pair of dense features in our formulation consists of two dense features from the same index (*i.e.* spatial position) of each augmented image pair (see the numerator of Eq. (2)). We discuss the strategy for choosing positive and negative dense pairs in further detail in Section 3.3. Eq. (2) can also be rewritten as follows by applying logarithmic rules:

$$L^{DenseCont} = -\frac{1}{\mathcal{N}} \sum_{i,j \overset{\text{i.i.d}}{\sim} \mathcal{B}} \frac{1}{HW} \sum_{p}^{HW} (sim(\mathbf{h}_{(i,p)}, \mathbf{h}_{(j,p)})/\lambda - \log \sum_{k \overset{\text{i.i.d}}{\sim} 2\mathcal{N}} \sum_{q}^{HW} \mathbb{1}_{[k \neq i] \times [\frac{q \neq p}{k=j}]} e^{sim(\mathbf{h}_{(i,p)}, \mathbf{h}_{(k,q)})/\lambda})$$

$$= \underbrace{-\frac{1}{\mathcal{N}} \sum_{i,j \overset{\text{i.i.d}}{\sim} \mathcal{B}} \frac{1}{HW} \sum_{p}^{HW} sim(\mathbf{h}_{(i,p)}, \mathbf{h}_{(j,p)})/\lambda}_{\text{alignment property}} + \underbrace{\frac{1}{\mathcal{N}} \sum_{i}^{\mathcal{N}} \frac{1}{HW} \sum_{p}^{HW} \log(\sum_{k \overset{\text{i.i.d}}{\sim} 2\mathcal{N}} \sum_{q}^{HW} \mathbb{1}_{[k \neq i] \times [\frac{q \neq p}{k=j}]} e^{sim(\mathbf{h}_{(i,p)}, \mathbf{h}_{(k,q)})/\lambda})}_{\text{distribution to be uniform property}} \tag{3}$$

where we again observe that dense CL consists of alignment and distribution objectives. Therefore, by optimizing Eq. (2), dense features will asymptotically achieve the alignment-uniformity properties, similar to the instance-level CL.

To control these properties more directly, we adopt the metrics proposed in Wang and Isola [30] and extend them to the dense-level. For the uniformity loss, we utilized a Gaussian potential kernel $G : \mathcal{S}^d \times \mathcal{S}^d \to \mathbb{R}_+$ [3, 9, 30] and the logarithm of the dense average pairwise Gaussian potential. Dense-level alignment-and-uniformity loss can be defined as:

$$L_a \triangleq -\frac{1}{\mathcal{N}} \sum_{i,j \overset{\text{i.i.d}}{\sim} \mathcal{B}} \frac{1}{HW} \sum_{p}^{HW} sim(\mathbf{h}_{(i,p)}, \mathbf{h}_{(j,p)}), \quad L_u \triangleq log \frac{1}{\mathcal{N}} \sum_{i,j \overset{\text{i.i.d}}{\sim} \mathcal{B}} \frac{1}{HW} \sum_{p}^{HW} G(\mathbf{h}_{(i,p)}, \mathbf{h}_{(j,p)})$$

where $G(\mathbf{x},\mathbf{y}) = e^{-\|\mathbf{x}-\mathbf{y}\|_2^2}$. denotes a pairwise Gaussian potential.

Perfect optimization of both properties is difficult to attain from a finite number of data points [30] but can be approximated when the data points (*e.g.* minibatch) are sufficiently large. Therefore, in addition to Eq. (2), we also use L_a and L_u as the objective functions of the pre-training phase and observe whether the two properties are correlated with the downstream tasks for a wide range of scenarios.

3.3 Dense Feature Matching

One issue in dense CL is finding the appropriate features to form positive pairs. The key to matching dense features is that positive pairs must share information (*i.e.* alignment), while negative pairs must repel each other (*i.e.* uniformity). Many studies provide complex strategies to pair strong positives and negative pairs to the anchor *e.g.* exploit geometrically identical features [17, 35], calculate attention score [32], or use momentum queue to enlarge the size of negative samples [31]. We address this issue with a spatially grounded dense feature matching (*i.e.* index-wise matching) based on the assumption from Section 3.2 that dense features of an instance and sampled data points are *i.i.d*. Our motivation for doing index-wise matching is to fairly compare the behavior of dense CL on multiple criteria as these tricks could yield various effects for each experiment.

Traditional CL[4, 5, 6, 12, 16, 29] can learn feature representations when the distance between positive samples is shorter than between negative samples. Also, this approach admits that negative samples contain noisy samples of the positive class, and these noises are negligible when the strong negative samples are large enough. In this context, our simple approach is also reasonable and effective in learning feature representation. For two dense feature sets $\mathbf{h}_1 = \{\mathbf{h}_{(1,1)},\dots,\mathbf{h}_{(1,HW)}\}$, $\mathbf{h}_{(1,i)} \in \mathbb{R}^d$ and $\mathbf{h}_2 = \{\mathbf{h}_{(2,1)},\dots,\mathbf{h}_{(2,HW)}\}$, $\mathbf{h}_{(2,i)} \in \mathbb{R}^d$ from two augmented images, positive pairs are formed by vectors of the same index in each set $pos = \{(\mathbf{h}_{(1,i)},\mathbf{h}_{(2,i)}),\dots,(\mathbf{h}_{(1,HW)},\mathbf{h}_{(2,HW)})\}$ and the vectors of different indices $neg = \tilde{\mathbf{h}}_2 = \{\mathbf{h}_{(2,j)},\dots,\mathbf{h}_{(2,HW)}\}, where\, j \neq i$ are formed as negative pairs including other dense feature vectors from different data points in $\mathcal{B}$. Therefore, our matching strategy forms a soft positive pair while forming many strong negative pairs ($\approx$ 12.5k dense features of other images; features from different data points) and some noisy negative pairs (different indices from the same data point). Such noisy pairs in negative pairs can be ignored given a large number of strong negative pairs. Although some negative pairs could share information (*e.g.* $\mathbf{h}_{(1,i)}$ and $\mathbf{h}_{(2,i+1)}$), asymptotically all negative pairs should follow a uniform distribution. Surprisingly, this simple matching strategy showed successful performance in all our experiments, suggesting that our *i.i.d* assumption was not unreasonable. We further investigate more sophisticated matching strategies that do not make such assumptions: dense feature matching based on cosine similarity [31], and set-wise matching based on earth mover distance [32]. We report in the supplementary that both strategies show either similar or inferior performance to the simple index-wise matching.

4 Experiments

Our experiments primarily focus on the correlation analysis between feature representations after pre-training and the performance of downstream tasks: linear evaluation as the instance-level task and object detection as the dense-level task. We pose three questions regarding dense features: 1) How does the alignment-uniformity property of dense contrast learning correlate with the performance of object detection and linear evaluation? 2) How different

is the behavior of dense feature representations on single or multi-object datasets? 3) How effective is the index-wise matching strategy in terms of different augmentation techniques?

In this section, we first describe experimental setup and how to quantify the correlation between alignment-uniformity property and downstream task performance. Then the following three subsections will address each of the three questions above.

4.1 Experimental Setup

Pre-training. We conduct pre-training experiments on two datasets: STL-10 [8] single-object dataset (∼103k images from the training and unlabeled sets) and MS COCO [14] multi-object dataset(∼118k images from the training set). We closely follow the hyper-parameters and data augmentation rules from the official implementation of Wang and Isola [30] for STL-10 and DenseCL [31] for COCO. We use Resnet18 as the backbone and extract the dense features from the penultimate layer (*i.e.* before the global average pooling layer). Then, these dense features are projected to two different sub-head blocks depending on the training scheme (instance-versus-dense). We train 200 STL-10 pre-trained models and 120 COCO pre-train models for 200 epochs with instance- and dense-level CL. Each model is optimized with a differently weighted combination of L_a and L_u, or various values of the temperature τ of $L_{InfoNCE}$. Please refer to the supplementary for further details.

Instance-level Evaluation. To evaluate the instance-level linear separation ability, we employ the STL-10 linear evaluation. We freeze the pre-trained weights and fine-tune only one additional linear classification layer for 100 epochs, strictly following the settings of Wang and Isola [30]. We use these results as a reference to correlate the instance-level alignment-uniformity properties using the global average pooled feature for each instance.

Dense-level Evaluation. When evaluating dense features, we follow the standard object detection protocol using the Faster R-CNN [21] detector (R18-C4 backbone) on the PASCAL VOC trainval 07+12 set and testing on the VOC test 2007 set. Optimization takes a total of 24k iterations. The learning rate is initialized to 0.02 and decayed to be 10 times smaller after 18k and 22k iterations. We use average precision (AP) as an evaluation metric and analyze the correlation by measuring the alignment-uniformity properties of dense features.

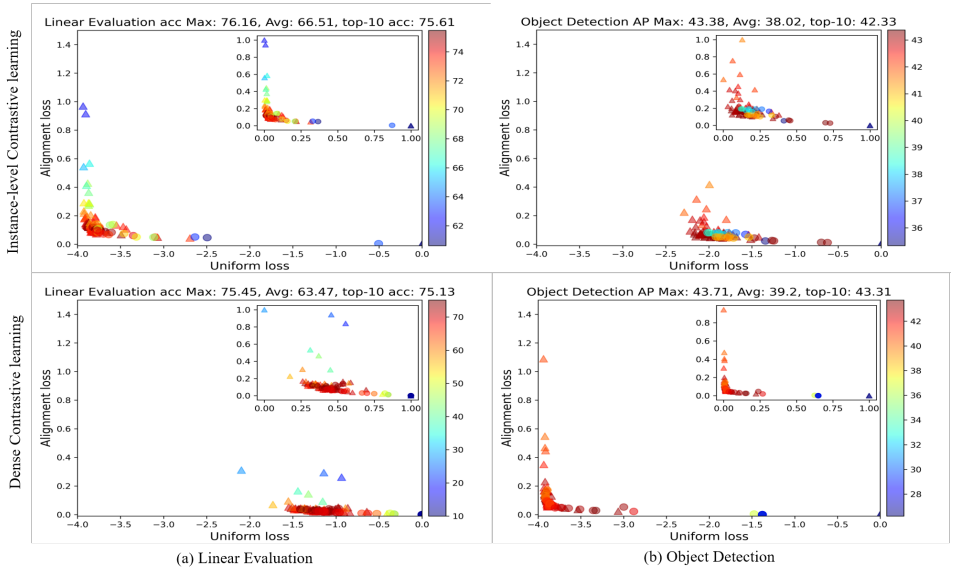
Quantifying Correlation. We quantify the strength of the correlation between alignment-uniformity properties and downstream task performance by utilizing the scalar-valued Kendall's τ, which is a rank-based correlation metric. Given $\mathcal{N}$ pre-trained models, the two losses (L_a, L_u), and the downstream task performance P_{task} are reordered with min-max normalization across $\mathcal{N}$ models as $r(L_a)$, $r(L_u)$, and $r(P_{task})$. Kendall's τ correlation metric is

$$\tau = \frac{P-Q}{\sqrt{(P+Q+T)(P+Q+U)}}$$

where, P and Q are the numbers of ordered and disordered pairs in $\{r(L_{a_i})+r(L_{u_i}), r(P_i)\}$, $i \in \mathcal{N}$. T and U are the numbers of ties in $\{r(L_{a_i})+r(L_{u_i})\}$ and $r(P_i)$, respectively. The correlation value varies between -1 and +1, with a value close to 0 indicating a weak correlation. Note that a negative correlation between the losses ($\{r(L_{a_i})+r(L_{u_i})\}$) and downstream task performance (P_{task}) indicate that alignment-uniformity are desirable properties, and contrastive pre-training is useful.

4.2 Results of Pre-training on Single-object Dataset

Instance-level Evaluation. Wang and Isola [30] demonstrated that the linear evaluation performance increased with the tendency to optimize alignment-uniformity. Inspired by its findings, we investigate the performance of linear evaluation and alignment-uniformity properties on the STL-10 testset using a global average pooling feature. As shown in Fig. 2



(a) Linear Evaluation (b) Object Detection

Figure 2: We show the alignment-uniformity property and downstream task performance for each 100 STL10 pre-trained models using instance- or dense-level features. All pre-trained models perform linear evaluation and object detection, then mark each point with color to show the performance. X and Y axes represent uniformity and alignment with a fixed scale. The symbol △ and ○ denotes $L_{InfoNCE}$ and L_a & L_u, respectively. We also show normalized L_a & L_u values in the upper right corner. Note that we examine the alignment-uniformity properties using the features depending on the evaluation aspect (instance vs dense) regardless of the pre-training scheme.

(a), the overall trend showed that the linear evaluation performance improved for the optimized alignment-uniformity property in both instance-level and dense CL, and all experiments showed a negative correlation (negative value of τ in Table 1). Also, instance-level and dense CL results achieved similar performance with a maximum accuracy of 76.16 and 75.45. These results show that dense contrast learning pre-trained on a single-object dataset has the ability to linearly separate by capturing the global information. We further investigate the behavior in the object detection task.

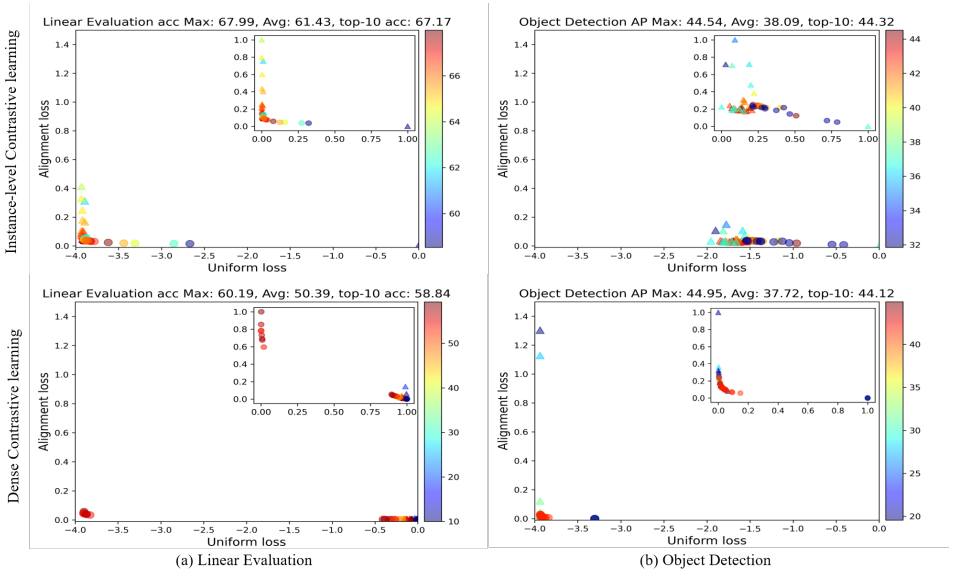
Dense-level Evaluation. To investigate the dense-level evaluation, we analyze the correlation between the alignment-uniformity of dense features on the STL-10 testset and VOC object detection performance. In this experiment, we can observe that the overall trend of the object detection performance is also correlated with the alignment-uniformity property in both instance-level and dense CL (Fig. 2 (b)) . Similar performance was achieved with a maximum AP of 43.38 and 43.71 in both instance-level and dense CL. The instance-level and dense CL using a single object showed a negative correlation between the alignment-uniformity and object detection ability with negative τ (Table 1). However, similar trends and performance may have been reached between instance level and dense contrast learning due to the inherent object-centric bias of the STL10 dataset. Still, the gap between the two pre-training schemes remains unknown. Therefore, we perform pre-training on a more complex setup involving multiple objects with the COCO dataset to ensure whether the cor-

Table 1: Single-object dataset results of instance and dense-level evaluation. We show the results for two different training scheme($L_{InfoNCE}$ and L_a & L_u) in a total of 200 experiments. L_a & L_u indicates loss of alignment and uniformity.

Pretraining	Loss	Instance-level Evaluation					Dense-level Evaluation				
		linear evaluation(Acc)				correlation	object detection(AP)				correlation
		exp	max	Avg	top10	τ	exp	max	Avg	top10	τ
Instance	L_a & L_u	70	76.16	64.39	75.56	-0.50	70	40.37	37.21	40.14	-0.31
	$L_{InfoNCE}$	30	75.47	71.99	74.97	-0.07	30	43.38	40.17	42.33	-0.41
	total	100	76.16	66.51	75.61	-0.45	100	43.38	38.02	42.33	-0.41
Dense	L_a & L_u	70	75.45	64.61	75.01	-0.19	70	43.44	38.99	43.19	-0.22
	$L_{InfoNCE}$	30	75.12	60.85	74.18	-0.01	30	43.71	39.63	42.80	-0.54
	total	100	75.45	63.47	75.13	-0.32	100	43.71	39.2	43.31	-0.12
Random init		1	28.04	-	-	1	31.93	-	-		

relation results of the STL10 pre-training are preserved.

4.3 Results of Pre-training on Multi-object Dataset.



(a) Linear Evaluation (b) Object Detection

Figure 3: We show the alignment-uniformity property and downstream task performance for each 60 COCO pre-trained models using instance- or dense-level features. Each point is marked with color to show its performance and uniformity and alignment properties are represented in X and Y axes with a fixed scale. The symbol △ and ○ denotes $L_{InfoNCE}$ and L_a & L_u, respectively.

Instance-level Evaluation. We conduct instance-level evaluations on COCO pre-trained models. The alignment-uniformity properties were measured using the global average pooled feature on the COCO testset while performing linear evaluation using the STL10 dataset. As shown in Fig. 3 (a), the trends of instance-level CL showed strong negative correlations with τ of -0.67 (Table 2). However, for the pre-training scheme with Dense CL, the results showed an irregular pattern depending on the uniformity, showing a weak correlation of -0.01 tau.

Also, COCO pre-training showed inferior to STL pre-training in linear evaluation with maximum accuracy of 67.99 and 60.19 in instance-level and dense CL. We perform an object detection task to investigate whether such a performance gap occurs in dense prediction tasks.

Dense-level Evaluation. To evaluate the dense features on COCO pre-trained model, we analyze the correlation between alignment-uniformity of dense features on COCO testset and VOC object detection performance. As seen from Fig. 3 (b), all experiments showed high performance as the alignment-uniformity metric decreased. Also, the instance-level and dense CL showed high performance with maximum AP of 44.54 and 44.95 and τ of -0.21 and -0.13 Table 2. From these results, pre-training schemes with instance-level or dense-contrast learning using multiple objects perform well in dense prediction tasks despite the complexity of rich semantic information.

Table 2: Multi-object dataset results for instance and dense-level evaluation.

Pretraining	Loss	Instance-level Evaluation					Dense-level Evaluation				
		linear evaluation(Acc)				correlation	object detection(AP)				correlation
		exp	max	Avg	top10	τ	exp	max	Avg	top10	τ
Instance	L_a & L_u	40	67.58	59.48	66.75	-0.54	40	44.54	38.27	43.94	-0.23
	$L_{InfoNCE}$	20	67.99	65.05	66.63	-0.67	20	44.51	37.77	42.83	-0.03
	total	60	67.99	61.43	67.17	-0.67	60	44.54	38.09	44.32	-0.13
Dense	L_a & L_u	40	60.19	53.41	58.64	-0.21	40	44.71	36.99	42.90	-0.41
	$L_{InfoNCE}$	20	59.29	46.36	57.30	-0.1	20	44.95	38.69	42.89	-0.54
	total	60	60.19	50.39	58.84	-0.01	60	44.95	37.72	44.12	-0.21
Random init		1	28.04	-	-	-	1	31.93	-	-	-

4.4 Confusing positive samples in Dense CL

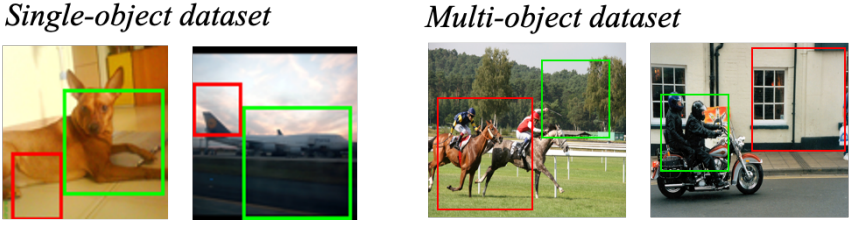


Figure 4: Confusing positive samples. The distances between the positive and negative pairs are similar.

Our assumption of feature matching by the index for positive pairs is that all features are *i.i.d.*, but two views from the same image should contain shared information. Single-object datasets, such as STL-10, are discriminated inter-class and object-centered. Due to the innate bias in these data sets, the mutual information in positive pairs (two random views in the same image) naturally shares similar information. However, in more complex setups with multiple objects, such as COCO, there is less chance of sharing semantically identical information even in positive pairs. To further investigate these biases in the data set, we analyze using non-overlapping image settings for confusing positive samples on STL10 and COCO datasets.

Table 3: Dense contrastive learning using not-obvious positive samples.

Pretraining	Instance-level Evaluation					Dense-level Evaluation				
	linear evaluation(Acc)				correlation	object detection(AP)				correlation
	exp	max	Avg	top10	τ	exp	max	Avg	top10	τ
Single-object	46	69.94	57.60	68.74	-0.35	46	43.06	40.01	42.78	-0.65
Multi-object	12	54.89	40.30	44.80	-0.15	12	32.49	32.03	32.12	0.03
Random init	1	28.04	-	-	-	1	31.93	-	-	-

In Table 3, the STL10 pre-training results of the linear evaluation and object detection achieved high performance on a single object dataset and showed a strong negative correlation. However, pre-training with confusing positive samples on multi-object datasets showed inferior results in linear evaluation and object detection tasks. In particular, object detection showed similar performance with random initialization result (maximum AP of 31.93) in achieving the maximum AP of 32.49 in the object detection task. It showed a positive correlation with alignment-uniformity (0.03τ). Therefore, the positive pairing method plays a crucial role in dense contrast learning so that positive pairs can share mutually agreeable information in multi-object datasets. Detailed setup and further experiments are shown in supplementary.

5 Conclusion

In this work, we mainly analyze the theoretical ideas of dense CL using a standard CNN and straightforward feature matching scheme rather than propose a new complex method. By extending existing instance-level CL analysis methods to dense-level, we observe the correlation between alignment-uniformity property of dense features and downstream tasks with newly proposed scalar metrics (linear evaluation and object detection). Also, we discover the core principle in constructing a positive pair of dense features and empirically proved its validity with a simple index-wise matching. In extensive experiments, we find that, regardless of pre-training schemes (instance-level or dense CL), pre-training on single object datasets showed the ability to linearly separate by capturing the global information and perform well on object detection tasks on multiple object datasets. Furthermore, our work can be potentially used to compare the performance of different CL schemes by evaluating alignment-uniformity properties of instance- and dense-level features before performing downstream tasks. The novelty of our work lies in carefully designed experiments and evaluation metric, allowing a reliable conversion from the “expected” to “confirmed”. We believe that the researchers can now safely rely on our findings and move on to developing more principled CL methods in the future, while treating our methods as a minimum baseline.

Acknowledgements

This work was supported by the KAIST-NAVER Hyper-Creative AI Center and the Institute of Information & Communications Technology Planning & Evaluation (IITP) grants (No.2019-0-00075 Artificial Intelligence Graduate School Program(KAIST) and No.2022-0-009840101003), and National Research Foundation of Korea (NRF) grant (NRF-2020H1D3A2A03100945) funded by the Korea government (MSIT).

References

[1] Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*, 2019.

[2] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *Advances in neural information processing systems*, 32, 2019.

[3] Sergiy V Borodachov, Douglas P Hardin, and Edward B Saff. *Discrete energy on rectifiable sets*. Springer, 2019.

[4] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pages 132–149, 2018.

[5] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020.

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[7] Ting Chen, Calvin Luo, and Lala Li. Intriguing properties of contrastive losses. *Advances in Neural Information Processing Systems*, 34:11834–11845, 2021.

[8] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings, 2011.

[9] Henry Cohn and Abhinav Kumar. Universally optimal distribution of points on spheres. *Journal of the American Mathematical Society*, 20(1):99–148, 2007.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[11] Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4918–4927, 2019.

[12] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[13] Jason D Lee, Qi Lei, Nikunj Saunshi, and Jiacheng Zhuo. Predicting what you already know helps: Provable self-supervised learning. *Advances in Neural Information Processing Systems*, 34:309–323, 2021.

[14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[15] Ralph Linsker. Self-organization in a perceptual network. *Computer*, 21(3):105–117, 1988.

[16] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6707–6717, 2020.

[17] Pedro O O Pinheiro, Amjad Almahairi, Ryan Benmalek, Florian Golemo, and Aaron C Courville. Unsupervised learning of dense visual representations. *Advances in Neural Information Processing Systems*, 33:4489–4500, 2020.

[18] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[19] Senthil Purushwalkam and Abhinav Gupta. Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases. *Advances in Neural Information Processing Systems*, 33:3407–3418, 2020.

[20] Senthil Purushwalkam and Abhinav Gupta. Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases. *Advances in Neural Information Processing Systems*, 33:3407–3418, 2020.

[21] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[22] Joshua Robinson, Li Sun, Ke Yu, Kayhan Batmanghelich, Stefanie Jegelka, and Suvrit Sra. Can contrastive learning avoid shortcut solutions? *Advances in neural information processing systems*, 34:4974–4986, 2021.

[23] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019.

[24] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10781–10790, 2020.

[25] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European conference on computer vision*, pages 776–794. Springer, 2020.

[26] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in Neural Information Processing Systems*, 33:6827–6839, 2020.

[27] Yuandong Tian, Lantao Yu, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning with dual deep networks. *arXiv preprint arXiv:2010.00578*, 2020.

[28] Christopher Tosh, Akshay Krishnamurthy, and Daniel Hsu. Contrastive learning, multi-view redundancy, and linear models. In *Algorithmic Learning Theory*, pages 1179–1206. PMLR, 2021.

[29] Trieu H Trinh, Minh-Thang Luong, and Quoc V Le. Selfie: Self-supervised pretraining for image embedding. *arXiv preprint arXiv:1906.02940*, 2019.

[30] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020.

[31] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3024–3033, 2021.

[32] Zhaoqing Wang, Qiang Li, Guoxin Zhang, Pengfei Wan, Wen Zheng, Nannan Wang, Mingming Gong, and Tongliang Liu. Exploring set similarity for dense self-supervised representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16590–16599, 2022.

[33] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018.

[34] Tete Xiao, Xiaolong Wang, Alexei A Efros, and Trevor Darrell. What should not be contrastive in contrastive learning. *arXiv preprint arXiv:2008.05659*, 2020.

[35] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16684–16693, 2021.

[36] Ceyuan Yang, Zhirong Wu, Bolei Zhou, and Stephen Lin. Instance localization for self-supervised detection pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3987–3996, 2021.

[37] Nanxuan Zhao, Zhirong Wu, Rynson WH Lau, and Stephen Lin. What makes instance discrimination good for transfer learning? *arXiv preprint arXiv:2006.06606*, 2020.