Mutual Conditional Probability for
Self-Supervised Learning

Takumi Kobayashi\textsuperscript{1,2}
takumi.kobayashi@aist.go.jp

\textsuperscript{1}National Institute of Advanced Industrial Science and Technology
Tsukuba, Japan
\textsuperscript{2}University of Tsukuba
Tsukuba, Japan

Abstract

Deep neural networks produce effective image feature representation through a supervised learning with plenty of annotation. To mitigate the label-hungry issue, self-supervised learning (SSL) works well for training the deep models without manual supervision. In SSL, pair-wise matching is widely applied to multi-view images via data augmentation techniques such as in a contrastive learning. In this work, we focus on a probabilistic distribution of the multi-view samples to effectively exploit the relationship among them which the pair-wise approaches hardly take into account. It leads to a novel loss based on mutual conditional probability through connecting SSL with mode seeking on the distribution. The method also exhibits connection to the other SSL methods. In the experiments on ImageNet classification tasks, the proposed method produces favorable performance in the framework of SSL.

1 Introduction

The last decade has witnessed great success of deep neural networks in various computer vision fields. For the remarkable performance, however, the deep models demand large amount of data equipped with sufficient annotations, which activates recent research effort toward semi-/un-supervised learning. In particular, a self-supervised approach attracts keen attention to bridge a performance gap between unsupervised and supervised learning \cite{8,14}.

The self-supervised learning (SSL) can be characterized by pretext tasks inducing losses which trigger the end-to-end learning of the deep models. In recent years, a contrastive loss \cite{15} is one of the popular losses to learn image feature representation \cite{5,7,8,17,24}. It is built upon pair-wise comparison between augmented images, i.e., multiple views\textsuperscript{1} of an image, through data augmentation techniques. The pair-wise losses can also be effectively applied to matching asymmetric encoder architectures \cite{6,14} and whitening \cite{12,19,36}.

The pair-wise approach minimizes a discrepancy between two views by enforcing even irrelevant view pairs to share the same feature representation as long as they are sampled from an identical source image via random data augmentation, e.g., image cropping. The degree of transformation is a key factor for learning effective feature representations \cite{31}.
and random augmentation process likely produces a inconsistent pair of views connected to different image contents (Fig. 1), while SSL has paid less attention to the contents of views.

In this paper, we introduce a probabilistic framework for multiple view samples. In contrast to the pair-wise approaches, multiple view samples are modeled by means of kernel density estimation (KDE) to cope with a multi-modal probabilistic distribution in which inconsistent view samples would form different modes (Fig. 1). Thereby, a mode in the distribution indicates robust representation of image content, leading to representation learning in SSL. In the probabilistic framework, we build a novel SSL loss based on mutual conditional probability on the distribution. The mutual conditional probability contributes to compact and discriminative feature representation which are favorable for SSL. The proposed method can also be viewed as a unified approach of the other SSL methods in the probabilistic framework.

1.1 Related works

Image-based pretext tasks. Instead of manual supervision, pretext tasks are constructed to provide annotation-free criteria or losses for guiding models to capture effective characteristics of image contents. Those tasks are, for example, to solve Jigsaw puzzle, estimate image rotation, predict spatial order, transform feature maps and fill in missing image region (inpainting). In the pretext tasks, losses are formulated as in supervised learning for classification and regression, according to the manually designed criterion.

Feature-based pretext tasks. There are also approaches to go into the details of feature distribution. On the assumption that a training dataset contains no duplicated image, instance discrimination imposes sample-wise classification tasks. Such a fine-grained task is relaxed via clustering to form classification over clusters, and the cluster assignment is improved so as to be compatible with mini-batch training.

Pair-wise matching. Without pseudo labeling such as via cluster assignment, pair-wise matching is effectively applied to a Siamese framework with data augmentation techniques. Data augmentation derives two view images from a source image as a positive pair to be contrasted with negative pairs drawn from different instances in a contrastive loss. One can see some architectural advances of the approach as well as asymmetric encoding architectures to remarkably exclude the negative pairs in a matching loss; it is analyzed such as through a stop-gradient technique. The pair-wise approaches, how-
ever, uniformly process multiple views without delving deep into the relationships among views, while we leverage a probabilistic approach to exploit the content-based relationships. The video SSL method [27] designs a loss function in a probabilistic manner. The method describes view sample distribution by means of parametric distribution in contrast to our kernel-density estimation, and it combines an instance discrimination and uncertainty minimization in a rather heuristic way while we unify the discriminative and generative models in a theoretical way via mutual conditional probability.

2  Method

To cope with multiple view samples drawn from an image, we model them in a probabilistic manner similarly to unsupervised clustering [9]. Then, mutual conditional probability is introduced to enhance the compactness and separability of the distribution, both of which contribute to view-invariance and sample discrimination in self-supervised learning (SSL).

2.1  Probabilistic models

2.1.1  Generative probability distribution: \( p(z|I) \)

Suppose we have \( M \) view samples \( \{x_i\}_{i=1}^M \), each of which is a \( d \)-dimensional feature vector extracted from one view of an input image \( I \). The probability density function over those samples is approximated by means of a kernel density estimation (KDE) [32] as

\[
p(z|I) = \frac{1}{MC_\sigma} \sum_{i=1}^M \exp\left(-\frac{1}{2\sigma^2} \|z - x_i\|^2\right),
\]

which contains a bandwidth parameter \( \sigma \) and a normalization constant \( C_\sigma \). We use the probe point \( z \in \mathbb{R}^d \) to explore the feature space of \( x_i \), which will be instantiated in Sec. 2.2 for SSL. The modes, local maxima points, of \( p(z|I) \) are obtained through considering the log-probability gradient of

\[
\sigma^2 \nabla \log p(z|I) = \frac{\sum_{i=1}^M \exp\left(-\frac{1}{2\sigma^2} \|z - x_i\|^2\right)x_i}{\sum_{i=1}^M \exp\left(-\frac{1}{2\sigma^2} \|z - x_i\|^2\right)} - z = \text{MS}_{p(z|I)}(z),
\]

which is a mean-shift (MS) vector [9] pointing to the local maxima around \( z \). The mean-shift vector normalized by the probability \( p(z|I) \) effectively seeks local modes [9], by iteratively updating the probe point \( z \) starting from \( x_i \) as

\[
z_i^0 = x_i, \quad z_i^{t+1} = z_i^t + \text{MS}_{p(z|I)}(z_i^t) = z_i^t + \sigma^2 \nabla \log p(z_i^t|I).
\]

Convergent point \( z_i^{\infty} \) indicates the local mode to which the sample \( x_i \) belongs. We can find modes \( \{z_i^{\infty}\}_{i=1}^N \) in an unsupervised manner without assigning sub-clusters to samples nor predefining number of modes. It is noteworthy that the mean-shift mode seeking is equivalent to minimizing the objective loss \( \ell_{\text{gen}}(z) = -\sigma^2 \log p(z|I) \Rightarrow \nabla_z \ell_{\text{gen}} = -\text{MS}_{p(z|I)}(z) \).

2.1.2  Discriminative probability distribution: \( p(I|z) \)

The view-sample distribution (1) on an image \( I \) is modeled in a generative way via KDE, lacking discrimination among images. In SSL, a view sample is usually associated with
TAKUMI KOBAYASHI: MUTUAL CONDITIONAL PROBABILITY FOR SSL

Figure 3: Comparison of three probabilistic models. $\mathcal{I}$ and $\hat{\mathcal{I}}$ are target and other images, respectively, and $\Rightarrow$ indicates increase of probability while $\Leftarrow$ means decrease.

one of $B$ images $\{I_b\}_{b=1}^B$ to construct a set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ with a label $y_i \in \{I_1, \cdots, I_B\}$ indicating an image index; $|\mathcal{D}| = N = MB$. Similarly to the procedure in Sec. 2.1.1, we can pursue discriminative points as follows. From the Bayesian viewpoint, the discriminative point is defined to maximize the posterior, $\text{argmax}_z p(I|z)$, of

$$p(I|z) = \frac{p(z|I)p(I|z)}{p(z)} \propto \sum_{i|y_i=I} \exp(-\frac{1}{\sigma^2} \|z-x_i\|^2_2) \sum_{i=1}^N \exp(-\frac{1}{\sigma^2} \|z-x_i\|^2_2),$$

where we assume uniform prior $p(I)$. It leads to the following loss and gradient,

$$\ell_{\text{disc}}(z) = -\sigma^2 \log p(I|z) \Rightarrow \nabla_z \ell_{\text{disc}} = -MS_{p(z|I)}(z) + MS_{p(z)}(z).$$

The gradient is based on difference between two mean-shift vectors of $p(z|I)$ and $p(z)$. The discriminative point is obtained by updating the probe point $z_i$ to minimize the loss (5),

$$z_i^0 = x_i, \quad z_i^{t+1} = z_i^t + \sigma^2 \nabla_z \log p(y_i|z_i^t),$$

where the class category is the one assigned to $x_i$, i.e., $I = y_i$ in (5). Compared to (3), it renders discriminative representation by considering difference from distributions of the other images, in a similar fashion to contrastive learning [15].

2.1.3 Mutual conditional probability: $p(z|I)p(I|z)$

From the perspective of minimizing losses, the generative probability $p(x|I)$ on an image $I$ effectively measures compactness of the distribution while being agnostic on separability; as shown in Fig. 3, the generative loss $\ell_{\text{gen}} = -\sigma^2 \log p(z|I)$ is decreased as the distribution on $I$ becomes compact, in disregard of the other distributions (images). On the other hand, the discriminative probability $p(I|z)$ cares about separability among distributions while paying less attention to the compactness of the distribution; it is sensitive to overlap between distributions as shown in Fig. 3 no matter how they are compact. In the literature of feature representation learning, those compactness and separability are related to robustness and discrimination of features, both of which are important to improve feature representation.

In this work, we incorporate those two factors into a loss by means of mutual conditional probability [16, 17] in our probabilistic framework (Sec. 2.1.1 & 2.1.2). The mutual conditional probability is defined as $p(z|I)p(I|z)$ which is the combination of generative and discriminative probabilities, thereby effectively unifying the factors of compactness and separability.
which are respectively addressed in those models as described above. In other words, it measures strength of coupling two components of an image $I$ and view sample $z$; obviously, it is maximized if there is one-to-one correspondence between $I$ and $z$. The mutual conditional probability has been applied in a text mining literature [13] to measure distinctiveness of words. Our probabilistic framework in Sec. 2.1.1&2.1.2 enables us to introduce the model into SSL by formulating the proposed loss as

$$
\ell(z) = -\sigma^2 \log[p(z|I)p(I|z)] = \ell_{gen}(z) + \ell_{disc}(z),
$$

(7)

$$
\Rightarrow \quad z_i^0 = x_i, \quad z_i^{t+1} = z_i^t + \sigma^2 \nabla_z \log[p(z_i^t|y_i)p(y_i|z_i^t)].
$$

(8)

Fig. 3 summarizes characteristics of the above-mentioned three probabilistic models. While the generative and discriminative probabilities lack sensitivity to either one of two factors, respectively, the proposed model favorably recognizes the compact and separable distribution; the bandwidth parameter $\sigma$ controls those compactness and separability, as discussed in Sec. 2.3. Thus, our loss contributes to enhancing feature representation of SSL in terms of view invariance as well as image-wise discriminativity, which are related to invariance against image perturbation and discrimination power for image class categories.

### 2.2 Self-Supervised Learning

We then connect the probabilistic model (Sec. 2.1.3) to self-supervised learning (SSL); the architectural overview is shown in Fig. 2. In a modern SSL framework [5, 7, 8, 12, 14], an image is viewed in multiple ways via data augmentation processes to produce multiple view samples on which neural networks are effectively trained in a contrastive [5] or matching [14] way without external supervision. Following the successful encoding architecture [13], each view sample is embedded in $d$-dimensional feature space via $x = \phi \circ \phi(I) \in \mathbb{R}^d$ with a projection head $\phi$ and a backbone feature extractor $\phi$, such as CNN [13], to be trained in SSL.

In this work, we apply the approach in Sec. 2.1 to probabilistically analyze the multi-view samples. Through $M$ random transformations $\{\tau_i\}_{i=1}^M$, a source image $I$ generates a set of $M$ view samples $\{x_i = \phi \circ \phi(\tau_i(I))\}_{i=1}^M$ on which the probabilistic distribution is constructed. We draw $B$ image instances $\{I_b\}_{b=1}^B$ usually packed in a mini-batch $B$ to provide $N = BM$ samples in total. The gradient-based update related to mean shift finds a local mode $z^*$ on $I$ to achieve robust representation against the transformations of $\{\tau_i\}_{i=1}^M$, which converge to the mode. Such a robust representation describes characteristics shared among those view images, which would be connected to an object shown in an image $I$. Our loss (7) also endows the robust representation with discriminativity among images $\{I_b\}_{b=1}^B$. For effectively learning representation, we tailor the formulation (Sec. 2.1.3) toward SSL as follows.

In mean shift procedure [13], the probe point $z_i$ is associated with the sample $x_i$ only through the initialization $z_i^0 = x_i$ in (8). To efficiently update representation $\phi \circ \phi$ in SSL, we associate $x_i$ with the probe point $z_i$ more directly by means of non-linear projection $\psi$ as $z_i = \psi(x_i) = \psi \circ \phi(\tau_i(I))$ in accordance with a prediction head [13]. Thereby, the iterative update in (8) is transformed into the update of the projection $\psi$ in an end-to-end learning which also improves $x_i$ via updating $\phi$ and $\psi$. The rapid update of the view sample $x_i$, however, is inconsistent with a stationarity assumption of the distribution (1) on $I$. To mitigate it, we apply the momentum models $\bar{\phi}$ and $\bar{\phi}$, exponential moving average [13] of $\phi$ and $\phi$ as in [13], to gradually update samples while stopping back-propagation on them [13]; so generated view samples are denoted as $\{\bar{x}_i = \bar{\phi} \circ \bar{\phi}(\tau_i(I))\}_{i=1}^M$. For stable learning on an image $I$, the probe $z_i$ explores over $M − 1$ view samples of $\{\bar{x}_j\}_{j \neq i}$ excluding
where the right-hand-side indicates a straightforward probability for improving SSL. For the case of multiple views related to contrastive loss of MoCo-v3 \cite{moco}. It is noteworthy that the two distinctive losses are unified in our probabilistic framework by means of mutual conditional probability in Sec. 3.1.

As a compromise between these extreme cases, we set the temperature \( T \) to provide feature embedding \( \mathbf{z}_{bi} = \psi \circ \phi (\mathbf{t}_{bi}(\mathbf{I}_b)) \) and \( \bar{\mathbf{x}}_{bi} = \bar{\phi} \circ \bar{\phi} (\mathbf{t}_{bi}(\mathbf{I}_b)) \).

### 2.3 Discussion

#### Connection to other SSL losses.

In case of two views, \( M = 2 \), our losses are reduced to

\[
\ell_{gen}(\mathbf{z}_b) = -T \log \exp \left( \frac{\mathbf{z}_{b1}^\top \bar{\mathbf{x}}_{b2}}{T} \right) = -\mathbf{z}_{b1}^\top \bar{\mathbf{x}}_{b2}, \quad \ell_{disc}(\mathbf{z}_b) \propto -\log \frac{\exp(\mathbf{z}_{b1}^\top \bar{\mathbf{x}}_{b2}/T)}{\sum_{c=1}^B \sum_{j=1, j \neq i}^M \exp(\mathbf{z}_{bi}^\top \bar{\mathbf{x}}_{cj}/T)},
\]

which are based on comparison between \( \mathbf{z}_b = \psi \circ \phi (\mathbf{t}_{b1}(\mathbf{I}_b)) \) and \( \bar{\mathbf{x}}_b = \bar{\phi} \circ \bar{\phi} (\mathbf{t}_{b1}(\mathbf{I}_b)) \).

The generative loss \( \ell_{gen} \) is equivalent to BYOL \cite{BYOL}, while the discriminative one \( \ell_{disc} \) is related to contrastive loss of MoCo-v3 \cite{moco}. It is noteworthy that the two distinctive losses of \( \ell_{gen} \) and \( \ell_{disc} \) are unified in our probabilistic framework by means of mutual conditional probability for improving SSL. For the case of multiple views \( M > 2 \), in the proposed multi-view loss (9), those pair-wise losses (11) are not straightforwardly summed up but merged in the probabilistic way via KDE (1).

**Temperature \( T \).** The temperature controls the bandwidth in KDE (1). The higher temperature \( T \) blurs the distribution into uniform one which takes into account even irrelevant (noisy) view samples. Actually, as \( T \to \infty \), the gradient of the generative loss is written by

\[
\frac{\partial \ell_{gen}(\mathbf{z}_b)}{\partial \mathbf{z}_b} = -\sum_{j \neq i} \exp(\mathbf{z}_{b1}^\top \bar{\mathbf{x}}_{bj}/T) \frac{\bar{\mathbf{x}}_b}{\sum_{j \neq i} \exp(\mathbf{z}_{b1}^\top \bar{\mathbf{x}}_{bj}/T)} \to_{T \to \infty} -\frac{1}{M-1} \sum_{j \neq i} \bar{\mathbf{x}}_b = \frac{\partial}{\partial \mathbf{z}_b} \mathbb{E}_{j \neq i} [-\mathbf{z}_{b1}^\top \bar{\mathbf{x}}_{bj}],
\]

where the right-hand-side indicates a straightforward summation of pair-wise BYOL loss \cite{BYOL}, \( \ell_{BYOL}(\mathbf{z}_b) = \mathbb{E}_{j \neq i} \frac{1}{2} \| \mathbf{z}_b - \bar{\mathbf{x}}_b \|_2^2 = 1 + \mathbb{E}_{j \neq i} [-\mathbf{z}_{b1}^\top \bar{\mathbf{x}}_{bj}] \) which is actually employed as a mean-shift SSL \cite{BYOL}.

On the other hand, the lower temperature \( T \) produces sparse distribution where the loss (9) makes sense only on the nearest-neighbor sample to the probe \( \mathbf{z} \), excluding all the other samples which contain both irrelevant and relevant ones;

\[
\frac{\partial \ell_{gen}(\mathbf{z}_b)}{\partial \mathbf{z}_b} = -\sum_{j \neq i} \exp(\mathbf{z}_{b1}^\top \bar{\mathbf{x}}_{bj}/T) \bar{\mathbf{x}}_b \to_{T \to 0} -\arg \max_{\{\mathbf{x}_b\}_{j \neq i}} \mathbf{z}_{b1}^\top \bar{\mathbf{x}}_{bj}.
\]

As a compromise between these extreme cases, we set the temperature \( T = 1 \) showing effectiveness in Sec. 3.1.
3 Experimental Results

We apply the proposed method to train CNNs on image recognition tasks in the SSL framework; the experimental settings are detailed in the supplementary material. We first analyze the performance of the method in Sec. 3.1, which is followed by comparison experiments with the other methods in Sec. 3.2.

3.1 Ablation study

We analyze the method on ImageNet-100 [33] dataset, a random 100-class subset of ImageNet [10]. We follow the setting of [12] which applies ResNet-18 [16] as a backbone feature extractor \( \phi \) with a projection \( \varphi \) and a probe head \( \psi \) both of which are implemented as two-layer MLPs to produce 128-dimensional features. For measuring performance, we apply both \( k \)-NN classification with \( k = 5 \) and linear classification to the \textit{frozen} backbone feature extractor \( \phi \) by using annotated training set [12].

3.1.1 Multiple views

Processing multiple views contains a critical issue regarding computation cost. \( M \) view samples per image demand \( M \)-forward&backward via \( \psi \circ \phi \circ \phi \) for \( z \) and \( M \)-forward paths via \( \tilde{\phi} \circ \tilde{\phi} \) for \( \tilde{x} \); in total, it requires \( 2M \)-forward and \( M \)-backward computations per image. For example, the standard SSL procedure based on pair-wise losses [8, 14] \((M = 2)\) consumes 4-forward and 2-backward computation on a \( 224 \times 224 \) patch image via random cropping. To reduce computation cost for many views \( M > 2 \), we apply the following three approaches.

\textbf{Smaller crop size [3].} It is straightforward to reduce image crop size of \( 224 \times 224 \) into, e.g., \( 160 \times 160 \) and \( 128 \times 128 \) which reduce computation costs by \( \frac{1}{2} \) and \( \frac{1}{3} \), respectively\(^2\).

\textbf{Memory bank [22].} We can draw samples related to \( \tilde{x} \) by means of \( k \)-NN in a memory bank with a small extra computation cost [22]. The memory bank pools momentum samples \( \tilde{x} \) of various views and images to provide \( M = 1 + k \) views in a \textit{pseudo} manner.

\textbf{Momentum-free representation [6].} The projection samples \( x = \varphi \circ \phi (\tau(I)) \) are available as view samples [6] without momentum encoder \( \tilde{\phi} \circ \tilde{\phi} \). They are given as by-products in a forward path toward \( z \) \textit{for free}, though breaking stationarity of distributions (Sec. 2.2).

These various approaches to construct view samples are compared in our SSL framework (Sec. 2.2) as shown in Table 1. They are designed based on the same computation budget as the standard \( M = 2 \)-view approach [8, 12] which processes 2 image patches of \( 224 \times 224 \) pixels; we denote it by \( 2 \cdot (224 \times 224) \). For reference, we also show in Table 1 the result of \( 4 \cdot (224 \times 224) \), though doubling the computation budget.

Small-sized cropping is superior to the other approaches, demonstrating that real view images with momentum encoding are more effective than the other pseudo ones using memory bank and the momentum-free encoding. Samples in a memory bank are less contributive to enhance diversity of views, and momentum-free samples break the stationarity assumption of distribution. There is a trade-off between image path size and number of views; larger number of views are sampled by cropping smaller-sized image patches. Plenty of views contribute to embedding various image characteristics into the distribution (1) while

\(^2\)The numbers of pixels are reduced by \( \frac{160^2}{256^2} = \frac{25}{49} \approx \frac{1}{2} \cdot \frac{128^2}{256^2} = \frac{16}{49} \approx \frac{1}{3} \).
Table 1: Performance comparison (acc, %) on various approaches to produce multiple views.

<table>
<thead>
<tr>
<th>views</th>
<th>linear</th>
<th>k-NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 \cdot (224 \times 224)</td>
<td>77.94</td>
<td>68.90</td>
</tr>
<tr>
<td>4 \cdot (160 \times 160)</td>
<td>81.64</td>
<td>74.06</td>
</tr>
<tr>
<td>6 \cdot (128 \times 128)</td>
<td>81.30</td>
<td>73.82</td>
</tr>
<tr>
<td>11 \cdot (96 \times 96)</td>
<td>79.20</td>
<td>71.54</td>
</tr>
</tbody>
</table>

The larger-sized image patch enriches the feature representation of a view sample. The approach of \( M = 4 \) to crop \( 4 \cdot (160 \times 160) \) patches well balances those two factors; the image size of \( 96 \times 96 \) is too small to extract meaningful features and \( M = 2 \) views in \( 2 \cdot (224 \times 224) \) fails to cover various image parts. Therefore, we apply \( M = 4 \) views cropping \( 160 \times 160 \) patches, exhibiting superior performance even to the costly approach of \( 4 \cdot (224 \times 224) \). The larger view patch of \( 224 \times 224 \) contains various image characteristics but is likely overlapped with the other views degrading diversity (variance) of view distribution. The approach of \( 4 \cdot (160 \times 160) \) provides favorably diverse view distribution with effective feature representation \( x \).

3.1.2 Probabilistic models with temperature

We compare the probabilistic models in Sec. 2.1. There are three types of models, generative probability (GENPRO) \( p(z|I) \), discriminative probability (DISCPRO) \( p(I|z) \) and mutual conditional probability (MUCONPRO) \( p(I|z)p(z|I) \) which provide the proposed SSL loss (10) as described in Sec. 2.2. Table 2 shows performance comparison of those models with various temperatures \( T \).

At the low temperature \( T = 0 \), the losses focus on the nearest neighbor samples over view samples in (13), failing to extract relationships among views; especially, it collapses the discriminative model DISCPRO. On the other hand, the high temperature \( T = \infty \) blurs the distribution over views to pay equal attention even to irrelevant views in (12). Thus, the temperature \( T = 1 \) produces favorable performance and the proposed MUCONPRO further boosts the performance by effectively unifying the generative and discriminative models.
Table 2: Performance comparison (acc. %) of various probabilistic models (Sec. 2.1) with various temperatures on ImageNet-100 using 4 views of 160×160 image patches (Table 1).

<table>
<thead>
<tr>
<th>Model</th>
<th>$T = 0$</th>
<th>$T = 1$</th>
<th>$T = \infty$</th>
</tr>
</thead>
<tbody>
<tr>
<td>GENPRO $p(z</td>
<td>I)$</td>
<td>79.24</td>
<td>81.38</td>
</tr>
<tr>
<td>DISCRO $p(I</td>
<td>z)$</td>
<td>17.86</td>
<td>81.56</td>
</tr>
<tr>
<td>MUCONPRO $p(z</td>
<td>I)p(I</td>
<td>z)$</td>
<td>78.38</td>
</tr>
</tbody>
</table>

(a) Linear evaluation
(b) k-NN evaluation

Table 3: Top-1 accuracy (linear eval.) on ImageNet with 100-epoch training.

<table>
<thead>
<tr>
<th>Model</th>
<th>Acc. (%)</th>
<th>Model</th>
<th>Acc. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MoCo-v2 [7]</td>
<td>67.4</td>
<td>MS-SSL [22]</td>
<td>72.4</td>
</tr>
<tr>
<td>MoCo-v3 [8]</td>
<td>68.9</td>
<td>GENPRO $p(z</td>
<td>I)$</td>
</tr>
<tr>
<td>SimSiam [6]</td>
<td>68.1</td>
<td>DISCRO $p(I</td>
<td>z)$</td>
</tr>
<tr>
<td>W-MSE [12]</td>
<td>69.4</td>
<td>MUCONPRO $p(z</td>
<td>I)p(I</td>
</tr>
</tbody>
</table>

As discussed with Fig. 3, it incorporates compactness and discriminativity, contributing to robustness for the higher temperatures $T \geq 1$. It should be noted that GENPRO with $T = \infty$ corresponds to the recent SSL approach [22] which simply sums up pair-wise BYOL loss.

### 3.2 Performance comparison

**ImageNet.** We then evaluate performance on the large-scale ImageNet dataset which is a standard benchmark dataset for 1K-category classification. We perform SSL on the training set to train backbone $\phi$ of ResNet-50 [16] over 100 training epochs following the protocol [6]. For evaluation, a linear classifier is trained on top of the frozen backbone on the annotated training set to measure classification accuracy on the validation set.

Performance comparison is shown in Table 3 including our three methods of GENPRO, DISCRO, MUCONPRO. In comparison to Table 2 on ImageNet-100, GENPRO is inferior to the other two methods in this ImageNet-1K which is 10-times larger than ImageNet-100. Through increasing diversity of training image samples, the discrimination part in DISCRO and MUCONPRO works better by contrasting various images to learn discriminative features, compared to the generative one (GENPRO) which only considers view distribution within an image.

The proposed MUCONPRO also outperforms the other methods on the ImageNet dataset, approaching even to the supervised one which is also trained over 100 epochs. It produces superior performance to the baselines of BYOL [6] and MoCoV3 [8] and even to their multi-view extension, GENPRO and DISCRO, in our framework (Sec. 2.1). The method is also competitive with SwAV which is based on clustering over multiple views of smaller-sized image patches [3]. These results demonstrate that the proposed method effectively leverages our probabilistic model to extract image characteristics shared among multiple view samples in the SSL framework. In this ImageNet task, our method produces favorable performance simply by 100-epoch training with a small batch size of 256 only on 4 NVIDIA V100 GPUs; it implies the feasibility of SSL even on limited computational resources without high-performance GPU clusters.
Table 4: Classification accuracies (%) by linear evaluation on various tasks.

<table>
<thead>
<tr>
<th>labeled sample</th>
<th>BYOL</th>
<th>MoCov3</th>
<th>W-MSE</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUN-397 [35]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10%</td>
<td>36.8</td>
<td>37.1</td>
<td>39.8</td>
<td><strong>40.7</strong></td>
</tr>
<tr>
<td>20%</td>
<td>43.1</td>
<td>43.4</td>
<td>45.2</td>
<td><strong>45.8</strong></td>
</tr>
<tr>
<td>50%</td>
<td>49.5</td>
<td>49.8</td>
<td>50.3</td>
<td><strong>52.0</strong></td>
</tr>
<tr>
<td>100%</td>
<td>52.9</td>
<td>53.3</td>
<td>53.4</td>
<td><strong>54.7</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>class category</th>
<th>BYOL</th>
<th>MoCov3</th>
<th>W-MSE</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>ImageNet-LT [23]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Many</td>
<td>45.36</td>
<td>45.82</td>
<td>45.23</td>
<td><strong>49.17</strong></td>
</tr>
<tr>
<td>Medium</td>
<td>32.62</td>
<td>32.40</td>
<td>32.56</td>
<td><strong>34.81</strong></td>
</tr>
<tr>
<td>Few</td>
<td>16.57</td>
<td>16.81</td>
<td>18.01</td>
<td><strong>18.62</strong></td>
</tr>
<tr>
<td>All</td>
<td>35.34</td>
<td>35.45</td>
<td>35.46</td>
<td><strong>38.14</strong></td>
</tr>
</tbody>
</table>

Various types of recognition tasks. Finally, the methods are evaluated in SSL on the other types of recognition tasks, scene classification on SUN-397 [35] and imbalanced classification on ImageNet-LT [23]. We apply the same training protocols as in Sec. 3.1 by using ResNet-18. On SUN-397, performance is evaluated in a semi-supervised setting where we train backbone of ResNet-18 in SSL on whole training set without labels and then build a linear classifier on partial subsets of a labeled training set. We follow the protocol of imbalanced learning [20] to evaluate performance on ImageNet-LT. Table 4 shows performance results demonstrating that our method produces favorable performance even on these types of image recognition tasks other than the standard ImageNet.

4 Conclusion

We have proposed a novel SSL loss in a probabilistic manner. Multiple view samples are useful for feature representation learning, and a probability distribution is formulated on the multiple views to exploit relationships among them. In this probabilistic framework, we propose the loss based on mutual conditional probability for improving compactness and discriminativity of feature representation in SSL. The method exhibits connection to BYOL [14] and MoCo-v3 [8] as well as produces favorable empirical performance on various image recognition tasks in comparison to the other SSL methods.

References


