Mutual Conditional Probability for Self-Supervised Learning

Takumi Kobayashi[1,2]
takumi.kobayashi@aist.go.jp

[1] National Institute of Advanced Industrial Science and Technology
Tsukuba, Japan

[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 [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 [15] is one of the popular losses to learn image feature representation [5, 7, 8, 17, 24]. It is built upon pair-wise comparison between augmented images, i.e., multiple *views*[1] of an image, through data augmentation techniques. The pair-wise losses can also be effectively applied to matching asymmetric encoder architectures [6, 14] and whitening [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 [31]

[1] *View* denotes a degraded image transformed from a source image through a data augmentation process.

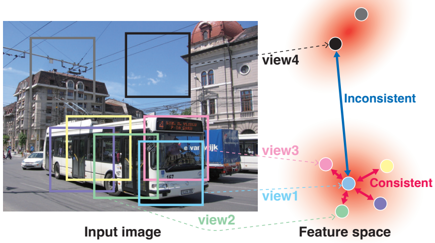


Figure 1: Multiple *views* produced by random data augmentation. They are likely to contain irrelevant pairs.

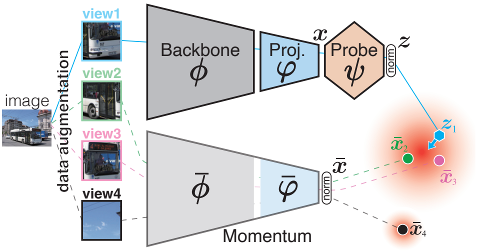


Figure 2: Our SSL architecture. Back-prop goes through the solid (cyan) line, while the dashed ones indicate stop-gradient [6].

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) [32] 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 [25], estimate image rotation [13], predict spatial order [18], transform feature maps [26] and fill in missing image region (inpainting) [28]. 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 [11, 34] imposes sample-wise classification tasks. Such a fine-grained task is relaxed via clustering to form classification over clusters [2, 37], and the cluster assignment is improved so as to be compatible with mini-batch training [3, 29].

Pair-wise matching. Without pseudo labeling such as via cluster assignment, pair-wise matching is effectively applied to a Siamese framework [15] 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 [15]. One can see some architectural advances of the approach in [5, 7, 8, 17] as well as asymmetric encoding architectures [14] to remarkably exclude the negative pairs in a matching loss; it is analyzed such as through a stop-gradient technique [6]. The pair-wise approaches, however, 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: $\mathtt{p}(z|\mathcal{I})$

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

$$\mathtt{p}(\boldsymbol{z}|\mathcal{I}) = \frac{1}{MC_\sigma}\sum_{i=1}^{M}\exp\Big(-\frac{1}{2\sigma^2}\|\boldsymbol{z}-\boldsymbol{x}_i\|_2^2\Big), \tag{1}$$

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

$$\sigma^2\nabla\log\mathtt{p}(\boldsymbol{z}|\mathcal{I}) = \frac{\sum_{i=1}^{M}\exp(-\frac{1}{2\sigma^2}\|\boldsymbol{z}-\boldsymbol{x}_i\|_2^2)\boldsymbol{x}_i}{\sum_{i=1}^{M}\exp(-\frac{1}{2\sigma^2}\|\boldsymbol{z}-\boldsymbol{x}_i\|_2^2)} - \boldsymbol{z} = \mathtt{MS}_{\mathtt{p}(\boldsymbol{z}|\mathcal{I})}(\boldsymbol{z}), \tag{2}$$

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

$$\boldsymbol{z}_i^0 = \boldsymbol{x}_i, \quad \boldsymbol{z}_i^{t+1} = \boldsymbol{z}_i^t + \mathtt{MS}_{\mathtt{p}(\boldsymbol{z}|\mathcal{I})}(\boldsymbol{z}_i^t) = \boldsymbol{z}_i^t + \sigma^2\nabla\log\mathtt{p}(\boldsymbol{z}_i^t|\mathcal{I}). \tag{3}$$

Convergent point $\boldsymbol{z}_i^\infty$ indicates the local mode to which the sample $\boldsymbol{x}_i$ belongs. We can find modes $\{\boldsymbol{z}_i^\infty\}_{i=1}^{N}$ 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 $\ell_{gen}(\boldsymbol{z}) = -\sigma^2\log\mathtt{p}(\boldsymbol{z}|\mathcal{I}) \Rightarrow \nabla_{\boldsymbol{z}}\ell_{gen} = -\mathtt{MS}_{\mathtt{p}(\boldsymbol{z}|\mathcal{I})}(\boldsymbol{z})$.

2.1.2 Discriminative probability distribution: $\mathtt{p}(\mathcal{I}|z)$

The view-sample distribution (1) on an image $\mathcal{I}$ is modeled in a generative way via KDE, lacking *discrimination* among images. In SSL, a view sample is usually associated with

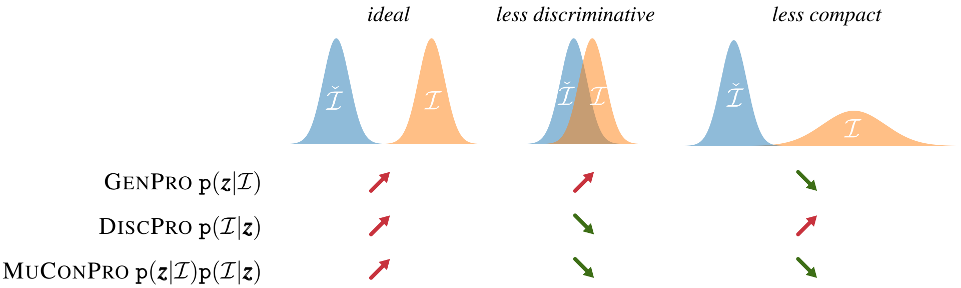

ideal
less discriminative
less compact
$\check{\mathcal{I}}$
$\mathcal{I}$
$\check{\mathcal{I}}$ $\mathcal{I}$
$\check{\mathcal{I}}$
$\mathcal{I}$
GenPro $\mathrm{p}(z|\mathcal{I})$ ↗ ↗ ↘
DiscPro $\mathrm{p}(\mathcal{I}|z)$ ↗ ↘ ↗
MuConPro $\mathrm{p}(z|\mathcal{I})\mathrm{p}(\mathcal{I}|z)$ ↗ ↘ ↘

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

one of B images $\{\mathcal{I}_b\}_{b=1}^B$ to construct a set $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ with a label $y_i \in \{\mathcal{I}_1, \cdots, \mathcal{I}_B\}$ 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, $\arg\max_{\boldsymbol{z}} \mathrm{p}(\mathcal{I}|\boldsymbol{z})$, of

$$\mathrm{p}(\mathcal{I}|\boldsymbol{z}) = \frac{\mathrm{p}(\boldsymbol{z}|\mathcal{I})}{\mathrm{p}(\boldsymbol{z})} \propto \frac{\sum_{i|y_i=\mathcal{I}} \exp(-\frac{1}{2\sigma^2}\|\boldsymbol{z}-\boldsymbol{x}_i\|_2^2)}{\sum_{i=1}^N \exp(-\frac{1}{2\sigma^2}\|\boldsymbol{z}-\boldsymbol{x}_i\|_2^2)}, \tag{4}$$

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

$$\ell_{disc}(\boldsymbol{z}) = -\sigma^2 \log \mathrm{p}(\mathcal{I}|\boldsymbol{z}) \quad \Rightarrow \quad \nabla_{\boldsymbol{z}} \ell_{disc} = -\mathtt{MS}_{\mathrm{p}(\boldsymbol{z}|\mathcal{I})}(\boldsymbol{z}) + \mathtt{MS}_{\mathrm{p}(\boldsymbol{z})}(\boldsymbol{z}). \tag{5}$$

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

$$\boldsymbol{z}_i^0 = \boldsymbol{x}_i, \quad \boldsymbol{z}_i^{t+1} = \boldsymbol{z}_i^t + \sigma^2 \nabla_{\boldsymbol{z}} \log \mathrm{p}(y_i|\boldsymbol{z}_i^t), \tag{6}$$

where the class category is the one assigned to $\boldsymbol{x}_i$, i.e., $\mathcal{I} = y_i$ 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: $\mathrm{p}(z|\mathcal{I})\mathrm{p}(\mathcal{I}|z)$

From the perspective of minimizing losses, the generative probability $\mathrm{p}(x|\mathcal{I})$ on an image $\mathcal{I}$ effectively measures compactness of the distribution while being agnostic on separability; as shown in Fig. 3, the generative loss $\ell_{gen} = -\sigma^2 \log \mathrm{p}(\boldsymbol{z}|\mathcal{I})$ is decreased as the distribution on $\mathcal{I}$ becomes compact, in disregard of the other distributions (images). On the other hand, the discriminative probability $\mathrm{p}(\mathcal{I}|\boldsymbol{z})$ 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* [1, 4] in our probabilistic framework (Sec. 2.1.1&2.1.2). The mutual conditional probability is defined as $\mathrm{p}(\boldsymbol{z}|\mathcal{I})\mathrm{p}(\mathcal{I}|\boldsymbol{z})$ 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 $\mathcal{I}$ and view sample $\boldsymbol{z}$; obviously, it is maximized if there is *one-to-one* correspondence between $\mathcal{I}$ and $\boldsymbol{z}$. The mutual conditional probability has been applied in a text mining literature [1, 4] 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

$$\ell(\boldsymbol{z}) = -\sigma^2 \log[\mathrm{p}(\boldsymbol{z}|\mathcal{I})\mathrm{p}(\mathcal{I}|\boldsymbol{z})] = \ell_{gen}(\boldsymbol{z}) + \ell_{disc}(\boldsymbol{z}), \tag{7}$$

$$\Rightarrow \quad \boldsymbol{z}_i^0 = \boldsymbol{x}_i, \quad \boldsymbol{z}_i^{t+1} = \boldsymbol{z}_i^t + \sigma^2 \nabla_{\boldsymbol{z}} \log[\mathrm{p}(\boldsymbol{z}_i^t|y_i)\mathrm{p}(y_i|\boldsymbol{z}_i^t)]. \tag{8}$$

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 σ 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 [5], each view sample is embedded in d-dimensional feature space via $\boldsymbol{x} = \varphi \circ \phi(\mathcal{I}) \in \mathbb{R}^d$ with a projection head φ and a backbone feature extractor ϕ, such as CNN [16], 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 $\mathcal{I}$ generates a set of M view samples $\{\boldsymbol{x}_i = \varphi \circ \phi(\tau_i(\mathcal{I}))\}_{i=1}^M$ on which the probabilistic distribution is constructed. We draw B image instances $\{\mathcal{I}_b\}_{b=1}^B$ usually packed in a mini-batch $\mathcal{B}$ to provide $N = BM$ samples in total. The gradient-based update related to mean shift finds a local mode $\boldsymbol{z}^*$ on $\mathcal{I}$ to achieve robust representation against the transformations of $\{\tau_i\}_{i|\boldsymbol{z}_i^\infty = \boldsymbol{z}^*}$ that 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 $\mathcal{I}$. Our loss (7) also endows the robust representation with discriminativity among images $\{\mathcal{I}_b\}_{b=1}^B$. For effectively learning representation, we tailor the formulation (Sec. 2.1.3) toward SSL as follows.

In mean shift procedure [9], the probe point $\boldsymbol{z}_i$ is associated with the sample $\boldsymbol{x}_i$ only through the initialization $\boldsymbol{z}_i^0 = \boldsymbol{x}_i$ in (8). To efficiently update representation $\varphi \circ \phi$ in SSL, we associate $\boldsymbol{x}_i$ with the probe point $\boldsymbol{z}_i$ more directly by means of non-linear projection ψ as $\boldsymbol{z}_i = \psi(\boldsymbol{x}_i) = \psi \circ \varphi \circ \phi(\tau_i(\mathcal{I}))$ in accordance with a prediction head [14]. Thereby, the iterative update in (8) is transformed into the update of the projection ψ in an end-to-end learning which also improves $\boldsymbol{x}_i$ via updating ϕ and φ. The rapid update of the view sample $\boldsymbol{x}_i$, however, is inconsistent with a stationarity assumption of the distribution (1) on $\mathcal{I}$. To mitigate it, we apply the momentum models $\bar{\phi}$ and $\bar{\varphi}$, exponential moving average [30] of ϕ and φ as in [7, 8, 14, 17], to gradually update samples while stopping back-propagation on them [6]; so generated view samples are denoted as $\{\bar{\boldsymbol{x}}_i = \bar{\varphi} \circ \bar{\phi}(\tau_i(\mathcal{I}))\}_{i=1}^M$. For stable learning on an image $\mathcal{I}$, the probe $\boldsymbol{z}_i$ explores over $M-1$ view samples of $\{\bar{\boldsymbol{x}}_j\}_{j\neq i}$ excluding

$\bar{\boldsymbol{x}}_i$ which is so close to $\boldsymbol{z}_i$ as to impede MS-based update by dominating the probability (1) for small σ. Toward effective SSL, both the view sample $\bar{\boldsymbol{x}}$ and the probe vector $\boldsymbol{z}$ are normalized by L_2 norm; such a normalization is compatible with mean shift in a vMF framework [21]. In summary, our method (7) produces the following SSL loss; for the probe $\boldsymbol{z}_{bi}$,

$$\ell_{SSL}(\boldsymbol{z}_{bi}) = -2T\log\sum_{j=1|j\neq i}^{M}\exp\left(\boldsymbol{z}_{bi}^{\top}\bar{\boldsymbol{x}}_{bj}/T\right) + T\log\sum_{c=1,j=1|c\neq b\vee j\neq i}^{B,M}\exp\left(\boldsymbol{z}_{bi}^{\top}\bar{\boldsymbol{x}}_{cj}/T\right), \tag{9}$$

$$\Rightarrow \min_{\psi,\phi,\varphi}\left[\frac{1}{BM}\sum_{b=1}^{B}\sum_{i=1}^{M}\ell_{SSL}\left(\frac{\psi\circ\varphi\circ\phi(\tau_{bi}(\mathcal{I}_b))}{\|\psi\circ\varphi\circ\phi(\tau_{bi}(\mathcal{I}_b))\|_2}\right)\right], \tag{10}$$

where the bandwidth parameter σ is re-parameterized into a conventional softmax temperature $T=\sigma^2$. The set of transformation $\{\tau_{bi}\}_{i=1}^{M}$ is randomly drawn by data augmentation for the image $\mathcal{I}_b$ to provide feature embedding $\boldsymbol{z}_{bi}=\psi\circ\varphi\circ\phi(\tau_{bi}(\mathcal{I}_b))$ and $\bar{x}_{bi}=\bar{\varphi}\circ\bar{\phi}(\tau_{bi}(\mathcal{I}_b))$.

2.3 Discussion

Connection to other SSL losses. In case of two views, $M=2$, our losses are reduced to

$$\ell_{gen}(\boldsymbol{z}_{b1}) = -T\log\exp\left(\frac{\boldsymbol{z}_{b1}^{\top}\bar{\boldsymbol{x}}_{b2}}{T}\right) = -\boldsymbol{z}_{b1}^{\top}\bar{\boldsymbol{x}}_{b2},\ \ \ell_{disc}(\boldsymbol{z}_{b1}) \propto -\log\frac{\exp(\boldsymbol{z}_{b1}^{\top}\bar{\boldsymbol{x}}_{b2}/T)}{\sum_{c\neq b\vee j\neq 1}^{B,2}\exp(\boldsymbol{z}_{b1}^{\top}\bar{\boldsymbol{x}}_{cj}/T)}, \tag{11}$$

which are based on comparison between $\boldsymbol{z}_{b1}=\psi\circ\varphi\circ\phi(\tau_{b1}(\mathcal{I}_b))$ and $\bar{x}_{b2}=\bar{\varphi}\circ\bar{\phi}(\tau_{b2}(\mathcal{I}_b))$. The generative loss ℓ_{gen} is equivalent to BYOL [14], while the discriminative one ℓ_{disc} is related to contrastive loss of MoCo-v3 [8]. It is noteworthy that the two distinctive losses of ℓ_{gen} and ℓ_{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

$$\frac{\partial\ell_{gen}(\boldsymbol{z}_{bi})}{\partial\boldsymbol{z}_{bi}} = -\frac{\sum_{j\neq i}\exp(\boldsymbol{z}_{bi}^{\top}\bar{\boldsymbol{x}}_{bj}/T)\bar{\boldsymbol{x}}_{bj}}{\sum_{j\neq i}\exp(\boldsymbol{z}_{bi}^{\top}\bar{\boldsymbol{x}}_{bj}/T)} \xrightarrow{T\to\infty} -\frac{1}{M-1}\sum_{j\neq i}\bar{\boldsymbol{x}}_{bj} = \frac{\partial}{\partial\boldsymbol{z}}\mathop{\mathbb{E}}_{j\neq i}[-\boldsymbol{z}_{bi}^{\top}\bar{\boldsymbol{x}}_{bj}], \tag{12}$$

where the right-hand-side indicates a straightforward *summation* of pair-wise BYOL loss [14], $\ell_{BYOL}(\boldsymbol{z}_{bi})=\mathbb{E}_{j\neq i}\frac{1}{2}\|\boldsymbol{z}_{bi}-\bar{\boldsymbol{x}}_{bj}\|_2^2=1+\mathbb{E}_{j\neq i}[-\boldsymbol{z}_{bi}^{\top}\bar{\boldsymbol{x}}_{bj}]$ which is actually employed as a mean-shift SSL [22].

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 $\boldsymbol{z}$, excluding all the other samples which contain both irrelevant and *relevant* ones;

$$\frac{\partial\ell_{gen}(\boldsymbol{z}_{bi})}{\partial\boldsymbol{z}_{bi}} = -\frac{\sum_{j\neq i}\exp(\boldsymbol{z}_{bi}^{\top}\bar{\boldsymbol{x}}_{bj}/T)\bar{\boldsymbol{x}}_{bj}}{\sum_{j\neq i}\exp(\boldsymbol{z}_{bi}^{\top}\bar{\boldsymbol{x}}_{bj}/T)} \xrightarrow{T\to 0} -\arg\max_{\{\bar{\boldsymbol{x}}_{bj}\}_{j\neq i}}\boldsymbol{z}_{bi}^{\top}\bar{\boldsymbol{x}}_{bj}. \tag{13}$$

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 ϕ with a projection φ and a probe head ψ 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 *frozen* backbone feature extractor ϕ 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 $\psi \circ \varphi \circ \phi$ for $\boldsymbol{z}$ and M-forward paths via $\bar{\varphi} \circ \bar{\phi}$ for $\bar{\boldsymbol{x}}$; 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 224×224 patch image via random cropping. To reduce computation cost for many views $M > 2$, we apply the following three approaches.

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

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

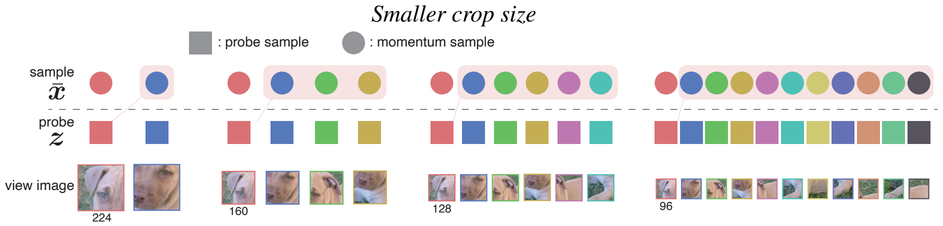
Momentum-free representation [6]. The projection samples $\boldsymbol{x} = \varphi \circ \phi(\tau(\mathcal{I}))$ are available as view samples [6] without momentum encoder $\bar{\varphi} \circ \bar{\phi}$. They are given as by-products in a forward path toward $\boldsymbol{z}$ *for free*, though breaking 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, 14] which processes 2 image patches of 224×224 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.

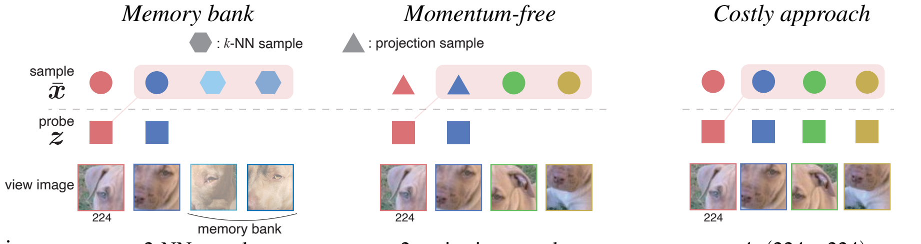
Small-sized cropping is superior to the other approaches, demonstrating that *real* view images with momentum encoding are more effective than the other pseudo ones using memory bank and the momentum-free encoding. Samples in a memory bank are less contributive to enhance diversity of views, and momentum-free 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 $\frac{160^2}{224^2} = \frac{25}{49} \sim \frac{1}{2}$, $\frac{128^2}{224^2} = \frac{16}{49} \sim \frac{1}{3}$.

Table 1: Performance comparison (acc, %) on various approaches to produce multiple views.



views	$2\cdot(224\times224)$	$4\cdot(160\times160)$	$6\cdot(128\times128)$	$11\cdot(96\times96)$
linear	77.94	**81.64**	81.30	79.20
k-NN	68.90	**74.06**	73.82	71.54



views	2-NN samples	2 projection samples	$4\cdot(224\times224)$
linear	78.04	78.68	81.12
k-NN	69.94	69.86	73.54

the larger-sized image patch enriches the feature representation of a view sample. The approach of $M=4$ to crop $4\cdot(160\times160)$ patches well balances those two factors; the image size of 96×96 is too small to extract meaningful features and $M=2$ views in $2\cdot(224\times224)$ fails to cover various image parts. Therefore, we apply $M=4$ views cropping 160×160 patches, exhibiting superior performance even to the costly approach of $4\cdot(224\times224)$. The larger view patch of 224×224 contains various image characteristics but is likely overlapped with the other views degrading diversity (variance) of view distribution. The approach of $4\cdot(160\times160)$ provides favorably diverse view distribution with effective feature representation $\boldsymbol{x}$.

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) $\mathrm{p}(\boldsymbol{z}|\mathcal{I})$, discriminative probability (DISCPRO) $\mathrm{p}(\mathcal{I}|\boldsymbol{z})$ and mutual conditional probability (MUCONPRO) $\mathrm{p}(\mathcal{I}|\boldsymbol{z})\mathrm{p}(\mathcal{I}|\boldsymbol{z})$ 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).

	(a) Linear evaluation			(b) k-NN evaluation		
Model	$T=0$	1	∞	$T=0$	1	∞
GenPro $\mathrm{p}(\boldsymbol{z}\|\mathcal{I})$	79.24	81.38	80.90	71.50	73.38	72.98
DiscPro $\mathrm{p}(\mathcal{I}\|\boldsymbol{z})$	17.86	81.56	81.08	16.58	73.46	73.38
MuConPro $\mathrm{p}(\boldsymbol{z}\|\mathcal{I})\mathrm{p}(\mathcal{I}\|\boldsymbol{z})$	78.38	**81.64**	81.50	70.16	**74.06**	74.00

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

Model	Acc. (%)	Model	Acc. (%)
SimCLR [5]	66.5	*Supervised [16]*	*76.2*
BYOL [14]	66.5	SwAV [3]	72.1
MoCo-v2 [7]	67.4	MS-SSL [22]	72.4
MoCo-v3 [8]	68.9	GenPro $\mathrm{p}(\boldsymbol{z}\|\mathcal{I})$	70.9
SimSiam [6]	68.1	DiscPro $\mathrm{p}(\mathcal{I}\|\boldsymbol{z})$	72.1
W-MSE [12]	69.4	MuConPro $\mathrm{p}(\boldsymbol{z}\|\mathcal{I})\mathrm{p}(\mathcal{I}\|\boldsymbol{z})$	**72.9**

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 ϕ 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, DiscPro, 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 discimination part in DiscPro 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 [14] and MoCoV3 [8] and even to their multi-view extension, GenPro and DiscPro, 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.

labeled sample	SUN-397 [35]				class category	ImageNet-LT [23]			
	BYOL	MoCov3	W-MSE	Ours		BYOL	MoCov3	W-MSE	Ours
10%	36.8	37.1	39.8	**40.7**	Many	45.36	45.82	45.23	**49.17**
20%	43.1	43.4	45.2	**45.8**	Medium	32.62	32.40	32.56	**34.81**
50%	49.5	49.8	50.3	**52.0**	Few	16.57	16.81	18.01	**18.62**
100%	52.9	53.3	53.4	**54.7**	All	35.34	35.45	35.46	**38.14**

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

[1] Ralf D. Brown. Automatically-extracted thesauri for crosslanguage ir: When better is worse. In *1st Workshop on Computational Terminology (Computerm)*, pages 15–21, 1998.

[2] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *ECCV*, 2018.

[3] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020.

[4] Libo Chen, Peter Fankhauser, Ulrich Thiel, and Thomas Kamps. Statistical relationship determination in automatic thesaurus construction. In *CIKM*, pages 267–268, 2005.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

[6] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, 2021.

[7] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv,* 2003.04297, 2020.

[8] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, 2021.

[9] Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[11] Alexey Dosovitskiy, Philipp Fischer, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(9):1734–1747, 2016.

[12] Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. Whitening for self-supervised representation learning. In *ICML*, 2021.

[13] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by pre- dicting image rotations. In *ICLR*, 2018.

[14] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *NeurIPS*, 2020.

[15] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, 2006.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and RossGirshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

[18] Olivier J. Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, S. M. Ali Eslami, and Aaron van den Oord. Data-efficient image recognition with contrastive predictive coding. In *ICML*, 2020.

[19] Tianyu Hua, Wenxiao Wang, Zihui Xue, Sucheng Ren, Yue Wang, and Hang Zhao. On feature decorrelation in self-supervised learning. In *ICCV*, 2021.

[20] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *ICLR*, 2020.

[21] Takumi Kobayashi and Nobuyuki Otsu. Von mises-fisher mean shift for clustering on a hypersphere. In *ICPR*, pages 2130–2133, 2010.

[22] Soroush Abbasi Koohpayegani, Ajinkya Tejankar, and Hamed Pirsiavash. Mean shift for self-supervised learning. In *ICCV*, pages 10326–10335, 2021.

[23] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *CVPR*, pages 2537–2546, 2019.

[24] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *CVPR*, pages 6707–6717, 2020.

[25] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, 2016.

[26] David Novotny, Samuel Albanie, Diane Larlus, and Andrea Vedaldi. Self-supervised learning of geometrically stable features through probabilistic introspection. In *CVPR*, 2018.

[27] Jungin Park, Jiyoung Lee, Ig-Jae Kim, and Kwanghoon Sohn. Probabilistic representations for video contrastive learning. In *CVPR*, pages 14711–14721, 2022.

[28] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. In *CVPR*, pages 2536–2544, 2016.

[29] Qi Qian, Yuanhong Xu, Juhua Hu, Hao Li, and Rong Jin. Unsupervised visual representation learning by online constrained k-means. *arXiv,* 2105.11527, 2021.

[30] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1195–1204, 2017.

[31] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning. In *NeurIPS*, 2020.

[32] M.P. Wand and M.C. Jones. *Kernel Smoothing*. Chapman and Hall, 1995.

[33] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *ICML*, 2020.

[34] Zhirong Wu, Yuanjun Xiong, X Yu Stella, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018.

[35] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010.

[36] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *ICML*, 2021.

[37] Chengxu Zhuang, Alex Lin Zhai, and Daniel Yamins. Local aggregation for unsupervised learning of visual embeddings. In *ICCV*, pages 6002–6012, 2019.