Representation Uncertainty in Self-Supervised Learning as Variational Inference

Hiroki Nakamura*
Panasonic Holdings Corp.
nakamura.hiroki003@jp.panasonic.com

Masashi Okada*
Panasonic Holdings Corp.
okada.masashi001@jp.panasonic.com

Tadahiro Taniguchi
Ritsumeikan University
taniguchi@em.ci.ritsumei.ac.jp

Abstract

In this study, a novel self-supervised learning (SSL) method is proposed, which considers SSL in terms of variational inference to learn not only representation but also representation uncertainties. SSL is a method of learning representations without labels by maximizing the similarity between image representations of different augmented views of an image. Meanwhile, variational autoencoder (VAE) is an unsupervised representation learning method that trains a probabilistic generative model with variational inference. Both VAE and SSL can learn representations without labels, but their relationship has not been investigated in the past. Herein, the theoretical relationship between SSL and variational inference has been clarified. Furthermore, a novel method, namely variational inference SimSiam (VI-SimSiam), has been proposed. VI-SimSiam can predict the representation uncertainty by interpreting SimSiam with variational inference and defining the latent space distribution. The present experiments qualitatively show that VI-SimSiam could learn uncertainty by comparing input images and predicted uncertainties. Additionally, we described a relationship between estimated uncertainty and classification accuracy.

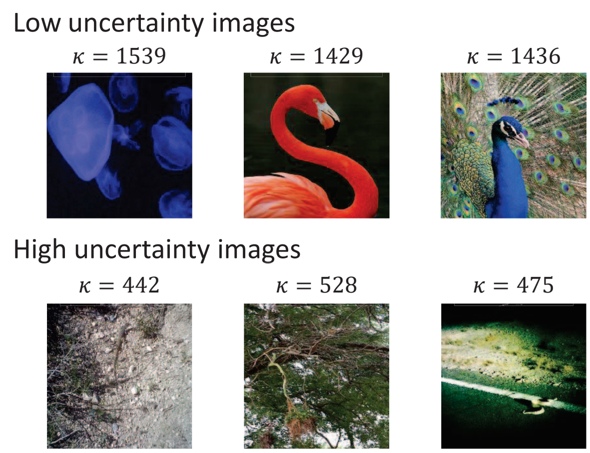

Figure 1: Input images and the predicted uncertainty parameter κ. Our method learns not only representations of images but also uncertainties in them without supervision. The lower the κ, the higher the uncertainty. The images with a low κ appear to have less salient features than those with a high κ.

1. Introduction

Self-supervised learning (SSL) is a framework for learning representations of data [6, 14, 20, 7, 5, 61, 50, 37, 28, 39]. This method enables training of high-performance models in downstream tasks (e.g., image classification and object detection) without substantial manually labeled data through pre-training the network to generate features. It can mitigate the annotation bottleneck, one of the crucial barriers to the practical application of deep learning. Some state-of-the-art SSL methods, such as SimSiam [6], SimCLR [5], and DINO [4], train image encoders by maximizing the similarity between representations of different augmented views of an image.

The probabilistic generative models with variational inference provide another approach for representation learning [27]. This approach learns latent representations in an unsupervised fashion by training inference and generative models (i.e., autoencoders) together. It can naturally incorporate representation uncertainty by formulating them as probabilistic distribution models (e.g., Gaussian). However, the pixel-wise objective for reconstruction is sensitive to rare samples [32] in such methods. Furthermore, this representation learning is recently found to be less competitive than the SSL methods on the benchmarking classification tasks [32, 37]. Although SSL and variational inference seem

*Equal contribution

highly related learning representations without supervision, their theoretical connection has not been fully explored.

In this study, we incorporate the variational inference concept to make the SSL uncertainty-aware and conduct a detailed representation uncertainty analysis. The contributions of this study are summarized as follows.

- We clarify the relationship between SSL (i.e., SimSiam, SimCLR, and DINO) and variational inference, generalizing the SSL methods as the variational inference of spherical or categorical latent variables (§4).
- We derive a novel SSL method called variational inference SimSiam (VI-SimSiam) by incorporating the above relationship. It learns to predict not only representations but also their uncertainty (§ 5).
- We demonstrate that VI-SimSiam successfully estimates uncertainty without labels while achieving competitive classification performance with SimSiam. We qualitatively evaluate the uncertainty estimation capability by comparing input images and the estimated uncertainty parameter κ, as shown in Fig. 1. In addition, we also describe that the predicted representation uncertainty κ is related to the accuracy of the classification task (§ 6).

A comparison of SimSiam and VI-SimSiam is illustrated in Fig. 2, where VI-SimSiam estimates the uncertainty by predicting latent distributions.

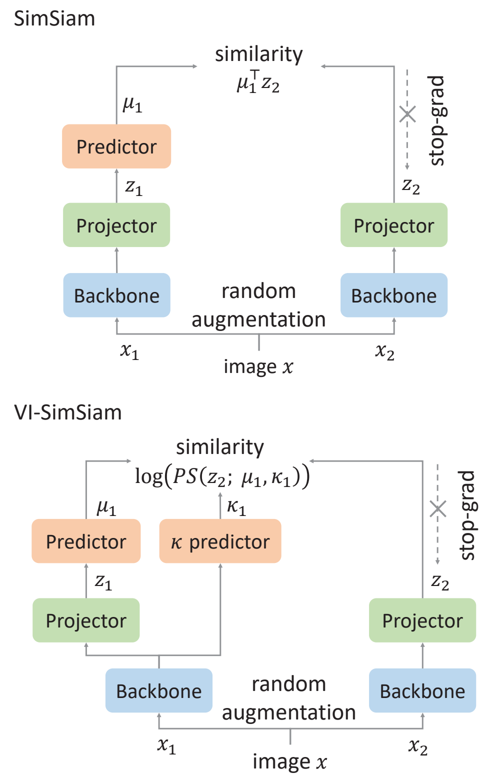

Figure 2: Schematics of VI-SimSiam and SimSiam [6]. These methods use the augmented images x_1 and x_2 as input images. In VI-SimSiam, we define the distribution of latent space that follows the Power Spherical distribution [10]. Unlike SimSiam, VI-SimSiam can predict the representation uncertainty of the image κ.

2. Related work

2.1. Self-supervised learning

SSL [5, 20, 6, 14, 4] has been demonstrated to have notable performance in many downstream tasks, such as classification and object detection. Contrastive SSL methods, including SimCLR[5] and MoCo [20], learn to increase the similarity of representation pairs augmented from an image (*positive pairs*) and to decrease the similarity of representation pairs augmented from different images (*negative pairs*). Conversely, non-contrastive SSL, including SimSiam [6], BYOL [14], and DINO [4], learn a model using only the positive pairs. Zbontar et al. [61] proposed another non-contrastive method using the redundancy-reduction principle of neuroscience. Furthermore, several studies have theoretically analyzed the SSL methods, such as Tian et al. [50] investigated the reasons behind the superior performance of the non-contrastive methods. Tao et al. [49] claimed that the (non-)contrastive SSL methods can be unified into one form using gradient analysis. Zhang [62] demonstrated a theoretical connection between masked autoencoder [19] and contrastive learning. However, these studies assumed a deterministic formulation without considering uncertainty in the representations.

2.2. Variational inference

In deep learning, variational inference is generally formulated using auto-encoding variational Bayes [27]. The variational inference estimates latent distributions, such as Gaussian [27], Gaussian mixture [51], and von Mises-Fischer (vMF) distribution for *hyperspherical* latent space [9]. Although Wang et al. [53] pointed out that SSL methods learn representations on the *hypersphere*, their relevance to the spherical variational inference [9] has yet to be investigated extensively.

A multimodal variational autoencoder [31, 55, 47, 48] is trained to infer latent variables from *multiple observations*[1] from different modalities. The latent variable distribution of multimodal variational inference is often assumed to be the *product of experts* or the *mixture of experts* of unimodal distributions [31, 55, 47]. Sutter et al. [48] also clarified the connection between them and generalized them as *mixture-of-products-of-experts* (MoPoE) VAE. Multimodal variational inference seems highly related to the SSL utiliz-

[1]We treat multiple observations as *multiview* and *multimodal*.

ing the *multiview inputs*. However, their relationship has been unclear.

2.3. Uncertainty-aware methods

Uncertainty-aware methods [12, 25, 26, 36, 54] have been proposed to solve the problem of learning hindered by data with high uncertainty. Kendall and Gal [25] proposed a method that estimated data uncertainty in regression and classification tasks by assuming that outputs follow a normal distribution. Scott et al. [46] proposed a stochastic spherical loss for classification tasks based on the von Mises–Fisher distributions. Additionally, Mohseni et al. [36] and Winkens et al. [54] presented methods for estimating uncertainty by combining SSL with a supervised classification task. Uncertainty-aware methods have been proposed for other tasks as well, such as human pose estimation [44, 41, 15], optical flow estimation [23], object detection [22, 17], and reinforcement learning [33, 40, 16].

Several studies have suggested methods to incorporate uncertainty in self-supervised learning of specific tasks [45, 42, 59, 52]. Poggi et al. [45] proposed an uncertainty-aware and self-supervised depth estimation. They considered the variance in depth estimated from multiple models as uncertainty. Wang et al. [52] demonstrated an uncertainty-aware SSL for three-dimensional object tracking, wherein the ratio of the distances between positive and negative pairs was treated as uncertainty. However, these methods hardly discussed representation uncertainty.

3. Preliminary

The formulations of the SSL methods and variational inference are briefly reviewed in this section.

3.1. Self-supervised learning methods

SimSiam SimSiam is a non-contrastive SSL method with an objective function that is defined as follows;

$$\mathcal{J}_{\text{SimSiam}} \coloneqq g_\theta(x_1)^\top f_\phi(x_2) + g_\theta(x_2)^\top f_\phi(x_1), \quad (1)$$

where x_1 and x_2 are two augmented views of a single image. The term g_θ and f_ϕ are encoders parameterized with θ and ϕ, respectively, and they map the image x to a spherical latent $z \in \mathbb{S}^{d-1}$. In the literature on non-contrastive SSL, the two encoders g_θ and f_ϕ are referred to as *online* and *target* networks, respectively. SimSiam defines the online network g_θ as $g_\theta = h_\theta \circ f_\theta$, where h_θ and f_θ are referred to as a *predictor network* and *projector network* [14, 6], respectively.

SimCLR The objective function of a contrastive SSL, such as SimCLR is generally described as follows;

$$\begin{aligned} \mathcal{J}_{\text{SimCLR}} &\coloneqq \mathcal{J}^{(1,2)}_{\text{SimCLR}} + \mathcal{J}^{(2,1)}_{\text{SimCLR}}, \\ \mathcal{J}^{(i,j)}_{\text{SimCLR}} &\coloneqq \log \frac{\exp(g_\theta(x_i)^\top f_\phi(x_j))}{\sum_{x' \in \mathcal{B}} \exp(g_\theta(x')^\top f_\phi(x_j))}. \end{aligned} \quad (2)$$

where $g_\theta(x), f_\phi(x) \in \mathbb{S}^{d-1}$, and $\mathcal{B}$ denotes a minibatch.

DINO DINO is another non-contrastive SSL with an objective and a latent space (i.e., categorical latent) different from those of SimSiam. It is described as;

$$\mathcal{J}_{\text{DINO}} \coloneqq -\mathcal{H}(P_{1,\phi}, P_{2,\theta}) - \mathcal{H}(P_{2,\phi}, P_{1,\theta}), \quad (3)$$

$$P_{i,\phi} \coloneqq \text{softmax}\left((f_\phi(x_i) - c)/\tau_\phi\right), \quad (4)$$

$$P_{j,\theta} \coloneqq \text{softmax}\left(g_\theta(x_j)/\tau_\theta\right), \quad (5)$$

where $\mathcal{H}(P, Q)(= \sum P \log Q)$ denotes cross entropy between two probabilities, $g_\theta(x), f_\phi(x) \in \mathbb{R}^d$, and (τ, c) are the parameters for $sharpening$ and $centering$ operations discussed later in § 4.3.

3.2. Multimodal generative model and inference

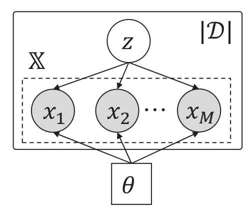

Figure 3: Graphical model.

Fig. 3 shows a graphical model for multimodal generative models, where $\mathcal{D}$ indicates a dataset, $\mathbb{X} = \{x_i\}_{i=1}^M$ is a set of multimodal observations x_i, z is a latent variable of the observations, θ is a deterministic parameter of the generative model $p(\mathbb{X}|z, \theta)$, and M is the number of modalities corresponding to the data augmentation types in this paper[1]. In the SSL context, $\mathbb{X}$ can be regarded as augmented images from stochastic generative processes.

The objective is to find a parameter θ^* that maximizes marginal observation likelihood;

$$\theta^* = \underset{\theta}{\text{argmax}}\, \mathcal{J} = \underset{\theta}{\text{argmax}}\, \mathbb{E}_{p(z)}[\log p(\mathbb{X}|z, \theta)]. \quad (6)$$

Since the marginalization $\mathbb{E}_{p(z)}[\cdot]$ is intractable, we can instead maximize the evidence lower bound (ELBO);

$$\begin{aligned} &\mathbb{E}_{p(z)}[\log p(\mathbb{X}|z, \theta)] \geq \mathcal{J}_{\text{ELBO}} \coloneqq \\ &= \mathbb{E}_{q(z|\mathbb{X},\phi)}\left[\log p(\mathbb{X}|z, \theta)\right] - D_{\text{KL}}[q(z|\mathbb{X}, \phi)|p(z)], \end{aligned} \quad (7)$$

where $q(z|\mathbb{X}, \phi)$ is a variational inference model parameterized with ϕ. By optimizing $\mathcal{J}_{\text{ELBO}}$ with respect to both θ and ϕ, $q(z|\mathbb{X}, \phi)$ approaches the posterior $p(z|\mathbb{X}, \theta)$ as the following relation holds;

$$\mathcal{J} - \mathcal{J}_{\text{ELBO}} = D_{\text{KL}}[q(z|\mathbb{X}, \phi)|p(z|\mathbb{X}, \theta)] \geq 0, \quad (8)$$

where $D_{\mathrm{KL}}[\cdot]$ is the Kullback-Leibler divergence. Notably, the posterior varies during the optimization process since it depends on the parameterized generative model; i.e., $p(z|\mathbb{X},\theta) \propto p(\mathbb{X}|z,\theta)p(z)$.

4. Self-supervised learning as inference

This section shows a connection between SSL and multimodal variational inference. Usually, a generative model $p(\mathbb{X}|z,\theta)$ and variational inference model $q(z|\mathbb{X},\phi)$ are realized with deep neural networks to solve the problem of Eq. (6), and they are trained via ELBO optimization. Instead, let us consider directly realizing the posterior $p(z|\mathbb{X},\theta)$ as deep neural networks. For this purpose, we remove the generative model term in Eq. (7) by applying Bayes' theorem;

$$p(\mathbb{X}|z,\theta) = p(z|\mathbb{X},\theta)p(\mathbb{X}|\theta)/\mathbb{E}_{p(\mathbb{X}|\theta)}[p(z|\mathbb{X},\theta)]. \quad (9)$$

Since $p(\mathbb{X}|\theta)$ is intractable, we approximate it with the empirical data distribution $p_{\mathcal{D}}(\mathbb{X})$. Substituting Eq. (9) into Eq. (7) and applying the approximation yields a new objective;

$$\begin{aligned} \mathcal{J}_{\mathrm{SSL}} &\coloneqq \mathcal{J}_{\mathrm{align}} + \mathcal{J}_{\mathrm{uniform}} + \mathcal{J}_{\mathrm{KL}} + p_{\mathcal{D}}(\mathbb{X}), \\ &\stackrel{+}{=} \mathcal{J}_{\mathrm{align}} + \mathcal{J}_{\mathrm{uniform}} + \mathcal{J}_{\mathrm{KL}}, \end{aligned} \quad (10)$$

where,

$$\mathcal{J}_{\mathrm{align}} \coloneqq \mathbb{E}_{q(z|\mathbb{X},\phi)}\left[\log p(z|\mathbb{X},\theta)\right], \quad (11)$$

$$\mathcal{J}_{\mathrm{uniform}} \coloneqq \mathbb{E}_{q(z|\mathbb{X},\phi)}\left[-\log p_{\mathcal{D}}(z|\theta)\right], \quad (12)$$

$$\mathcal{J}_{\mathrm{KL}} \coloneqq -D_{\mathrm{KL}}[q(z|\mathbb{X},\phi)|p(z)], \quad (13)$$

$$p_{\mathcal{D}}(z|\theta) \coloneqq \mathbb{E}_{p_{\mathcal{D}}(\mathbb{X})}[p(z|\mathbb{X},\theta)]. \quad (14)$$

Furthermore, let $p(z|\mathbb{X},\theta)$ and $q(z|\mathbb{X},\phi)$ respectively be Product-of-Experts (PoE) and Mixture-of-Experts (MoE) of the single modal inference models;

$$p(z|\mathbb{X},\theta) = \eta_\theta \prod_{j=1}^{M} p(z|x_j,\theta), \quad (15)$$

$$q(z|\mathbb{X},\phi) \coloneqq \frac{1}{M}\sum_{i=1}^{M} q(z|x_i,\phi), \quad (16)$$

where $\eta_\theta^{-1} \coloneqq \int \prod p(z|x_j,\theta)dz$ is the renormalization term. Then, we can rewrite $\mathcal{J}_{\mathrm{align}}$ as a form that encourages aligning latent variables from different modals;

$$\mathcal{J}_{\mathrm{align}} = \sum_{i,j} \mathbb{E}_{q(z|x_i,\phi)}\left[\log p(z|x_j,\theta)\right] + M\log\eta_\theta, \quad (17)$$

Eq. (15) is from Prop. 4.1 described below. Eq. (16) is the definition theoretically validated in [47, 48]. Practically, we can ignore unimodal comparisons (i.e., $i = j$) since they provide less effective information.

Proposition 4.1. Let $p(z)$ be a non-informative prior. The multi-modal posterior $p(z|\mathbb{X},\theta)$ takes the form of PoE of the single-modal posteriors $p(z|x_j,\theta)$.

Proof. See Appx. A.1.

We claim that Eq. (10) generalizes the objectives in Eqs. (1), (2) and (3) as summarized in Table 1. In the rest of this section, we describe how to recover the objectives in the table from Eq. (10). In the derivations, the term $M\log\eta_\theta$ is ignored by approximating it as a constant (denoted as $\stackrel{+}{\simeq}$).

4.1. SimSiam as inference

First, we discuss the relationship between SimSiam and the following definition involving the hyperspherical space $\mathbb{S}^{d-1}$.

Definition 4.2.

$$p(z) \coloneqq \mathcal{U}(\mathbb{S}^{d-1}), \quad (18)$$

$$q(z|\mathbb{X},\phi) \coloneqq \frac{1}{M}\sum_{i=1}^{M}\delta(z - f_\phi(x_i)), \quad (19)$$

$$p(z|\mathbb{X},\theta) \coloneqq \eta_\theta \prod_{j=1}^{M} \mathrm{vMF}(z;\mu = g_\theta(x_j),\kappa), \quad (20)$$

$$\mathrm{vMF}(z;\mu,\kappa) \coloneqq C_{\mathrm{vMF}}(\kappa)\exp(\kappa\mu^\top z), \quad (21)$$

where $g_\theta(x), f_\phi(x) \in \mathbb{S}^{d-1}$, and $\mathcal{U}(\mathbb{S}^{d-1})$ is the uniform distribution on the hypersphere. The term $\mathrm{vMF}(z;\mu,\kappa)$ is the von-Mises-Fisher distribution [35] parameterized with the mean direction $\mu \in \mathbb{S}^{d-1}$ and concentration (inversed variance) $\kappa \in \mathbb{R}^+$. The term $C_{\mathrm{vMF}}(\kappa)$ is a normalization constant defined with the modified Bessel function. In §4, κ and $C_{\mathrm{vMF}}(\kappa)$ are defined to be constants.

Relation to $\mathcal{J}_{\mathrm{align}}$

Claim 4.3. Under Def. 4.2, we can recover $\mathcal{J}_{\mathrm{SimSiam}}$ in Eq. (1) from $\mathcal{J}_{\mathrm{align}}$;

$$\mathcal{J}_{\mathrm{align}} \stackrel{+}{\simeq} \sum_{i,j} g_\theta^\top(x_j)f_\phi(x_i), \quad (22)$$

Proof. See Appx. A.2.

Relation to $\mathcal{J}_{\mathrm{uniform}}$

Claim 4.4. The presence of the predictor h_θ implicitly maximizes $\mathcal{J}_{\mathrm{uniform}}$.

Proof. Here, we borrow theoretical findings from the DirectPred literature [50]. In the literature, considering that $g_\theta = h_\theta \circ f_\theta$, the predictor h_θ, defined as the following linear model, can lead to successful convergence;

$$h_\theta(z) \coloneqq (U\Lambda^{\frac{1}{2}}U^\top)z, \quad (23)$$

Table 1: The summary of SSL methods interpreted as variational inference.

	SimSiam [6] (§4.1)	SimCLR [5] (§4.2)	DINO [49] (§4.3)	VI-SimSiam (ours) (§5)
$p(z)$	$\mathcal{U}(\mathbb{S}^{d-1})$		$\mathcal{U}(\Delta^{d-1})$	$\mathcal{U}(\mathbb{S}^{d-1})$
$q(z\|x_i, \phi)$	Deterministic		Categorical	Deterministic
$p(z\|x_j, \theta)$	von-Mises-Fisher		Categorical	Power Spherical
$\mathcal{J}_{\text{align}}$	$z_j^\top z_i$	$\log \frac{\exp(z_j^\top z_i)}{\sum_z \exp(z^\top z_i)}$	$\sum P_j \cdot \log P_i$	$\log C_{\text{PS}}(\kappa_j) + \kappa_j \operatorname{log1p}(z_j^\top z_i)$
$\mathcal{J}_{\text{uniform}}$	DirectPred		Centering	DirectPred
$\mathcal{J}_{\text{KL}}$	Const.		Sharpening	Const.
Uncertainty aware	No (uncertainty parameter κ is *fixed*)		Yes (but has not been discussed)	Yes (uncertainty parameter κ is *estimated*)

where U and Λ are the eigenvectors and eigenvalues of the covariance matrix of random variables z from $f_\theta(x)$; i.e.,

$$\mathbb{E}_{p_{\mathcal{D}}(x)}[f_\theta(x) f_\theta^\top(x)] = U \Lambda U^\top. \tag{24}$$

With Eq. (23), the cosine similarity can be rewritten as[2];

$$g_\theta^\top(x_j) f_\phi(x_i) = (U^\top z_j)^\top \Lambda^{\frac{1}{2}} (U^\top z_i), \tag{25}$$

where $z_j = f_\theta(x_j)$ and $z_i = f_\phi(x_i)$ are projected by $U^\top$ (i.e., projection matrix of PCA) so that each dimension is as independent as possible. In addition, the eigenvalues Λ converge to the identity matrix, i.e., the variances of z are constant. These behaviors encourage the marginalized z-distribution to be *isotropic*, i.e., $\mathcal{U}(\mathbb{S}^{d-1})$. Consequently, $\mathcal{J}_{\text{uniform}}$ is maximized since it takes the maximum iff $p_{\mathcal{D}}(z|\theta) = \mathcal{U}(\mathbb{S}^{d-1})$.

Relation to $\mathcal{J}_{\text{KL}}$ Since $q(z|\mathbb{X}, \phi)$ is deterministic, $\mathcal{J}_{\text{KL}}$ is constant.

4.2. SimCLR as inference

Further, we derive the contrastive learning objective with Def. 4.2. $\mathcal{J}_{\text{KL}}$ is constant, as in the previous section, and hence, is disregarded in the current section.

Relation to $\mathcal{J}_{\text{align}}$ and $\mathcal{J}_{\text{uniform}}$ Converse to the previous non-contrastive case, we consider explicitly optimizing $\mathcal{J}_{\text{uniform}}$, leading to the following claim;

Claim 4.5. Under Def. 4.2, we can recover $\mathcal{J}_{\text{SimCLR}}$ in Eq. (2) from $\mathcal{J}_{\text{align}}$ and $\mathcal{J}_{\text{uniform}}$;

$$\mathcal{J}_{\text{align}} + \mathcal{J}_{\text{uniform}} \overset{+}{\simeq} \sum_{i,j} \mathbb{E}_{q(z|x_i,\phi)} \left[\log \frac{\exp(g_\theta(x_i)^\top f_\phi(x_j))}{\sum_{x' \sim \mathcal{B}} \exp(g_\theta(x')^\top f_\phi(x_j))} \right]. \tag{26}$$

Proof. See Appx. A.3.

[2] $g_\theta^\top(x_j) f_\phi(x_i) = ((U\Lambda^{\frac{1}{2}}U^\top) z_j)^\top z_i = (U^\top z_j)^\top \Lambda^{\frac{1}{2}} U^\top (UU^\top z_i)$, where $UU^\top = I$.

Note that as per Ref. [53], Eq. (2) can be decomposed into two objectives similar to $\mathcal{J}_{\text{align}}$ and $\mathcal{J}_{\text{uniform}}$. In contrast, our derivation here aimed to derive Eq. (2) in a general form from variational inference.

4.3. DINO as inference

Finally, we derive the objective of DINO with the following definitions, where z is defined in the simplex Δ^{d-1},

Definition 4.6.

$$p(z) := \mathcal{U}(\Delta^{d-1}), \tag{27}$$

$$q(z|\mathbb{X}, \phi) := \frac{1}{M} \sum_{i=1}^{M} \mathrm{Cat}\left(z; P_{i,\phi}\right), \tag{28}$$

$$p(z|\mathbb{X}, \theta) := \eta_\theta \prod_{j=1}^{M} \mathrm{Cat}\left(z; P_{j,\theta}\right), \tag{29}$$

where the definitions of $P_{i,\phi}$ and $P_{j,\theta}$ follow Eqs. (4) and (5), and Cat means categorical distribution.

Relation to $\mathcal{J}_{\text{align}}$

Claim 4.7. Under Def. 4.6, we can recover $\mathcal{J}_{\text{DINO}}$ in Eq. (3) from $\mathcal{J}_{\text{align}}$;

$$\mathcal{J}_{\text{align}} \overset{+}{\simeq} \sum_{i,j} -\mathcal{H}(P_{i,\phi}, P_{j,\theta}), \tag{30}$$

Proof. Substituting Eqs. (28) and (29) into Eq. (11) derives Eq. (30).

Relation to $\mathcal{J}_{\text{uniform}}$

Claim 4.8. *Centering* in Eq. (4) optimizes $\mathcal{J}_{\text{uniform}}$.

Proof. The centering parameter c in Eq. (4) is determined and updated with an exponential moving average of batch mean of f_ϕ; $c \leftarrow mc + (1-m)\mathbb{E}_{\mathcal{B}}[f_\phi(x)]$ [4], where m is a rate parameter. This batch-norm-like centering encourages the marginalized distribution $q(z|\phi) = \mathbb{E}_{p_{\mathcal{D}}(x)}[q(z|x, \phi)]$

to be $\mathcal{U}(\Delta^{d-1})$[3] As a result, $p_{\mathcal{D}}(z|\theta)$ will also be uniform while optimizing $\mathcal{J}_{\text{align}}$, thus maximizing $\mathcal{J}_{\text{uniform}}$.

Relation to $\mathcal{J}_{\text{KL}}$

Claim 4.9. *Sharpening* in Eq. (4) regularizes $\mathcal{J}_{\text{KL}}$.

Proof. Since $p(z)$ is defined to be uniform, $\mathcal{J}_{\text{KL}}$ acts as an entropy regularizer to the variational distribution $q(z|\mathbb{X}, \phi)$. With a carefully designed temperature τ_ϕ, sharpening directly regularizes the entropy for learning success; e.g., τ_ϕ must be within 0.02 and 0.06 [4].

4.4. Discussion

Relation to another SSL method Moreover, we can integrate Barlow Twins [61] into this *SSL as inference* framework. Barlow Twins considers the cross-correlation matrix $\mathbb{E}_{\mathcal{B}}[g_\theta(x_i) f_\phi(x_j)^\mathsf{T}]$ and optimizes its diagonal and non-diagonal elements to be ones and zeros, corresponding to the optimization of $\mathcal{J}_{\text{align}}$ and $\mathcal{J}_{\text{uniform}}$, respectively.

Limitations In the derivation, we made some assumptions; $p(\mathbb{X}|\theta) \simeq p_{\mathcal{D}}(\mathbb{X})$ in Eq. (9), $q(z|\mathbb{X}, \phi)$ is MoE (Eq. (16)), and the renormalization term η_θ can be ignored in Eq. (17). The $\mathcal{J}_{\text{SSL}}$ derived under these assumptions is not guaranteed to be the lower bound of $\mathcal{J}$ in Eq. (6). However, the previous success of the SSL method empirically indicates that the assumptions are valid and the optimization progressed while keeping $\mathcal{J} \geq \mathcal{J}_{\text{SSL}}$. We hypothesize that to update the target network f_θ such as stop gradient and exponential moving average (EMA) [14, 6, 4, 18], the heuristics may contribute to the above optimization behavior. However, further theoretical analysis is necessary for future work. Notably, the stop gradient and EMA can be interpreted as the EM algorithm in variational inference [1]. It is also an attractive research direction to extend SSL methods by relaxing the assumption of $q(z|\mathbb{X}, \phi)$; e.g., changing its form from MoE to MoPoE [48], and replacing the deterministic δ in Eq. (19) with probability distributions.

Representation Uncertainty As shown in Table 1. DINO inherently can estimate uncertainty because it fully estimates categorical parameters. However, its ability has hardly been discussed. Conversely, although SimSiam and SimCLR assumed hyperspherical distributions, the uncertainty parameter κ is fixed, thus missing the ability for uncertainty estimation. To bridge the gap of uncertainty awareness, we propose a new uncertainty-aware method VI-SimSiam by extending SimSiam in §5. The capability of uncertainty estimation of VI-SimSiam and DINO is subsequently evaluated in §6.

[3]In contrast, batch-norm makes the batch distribution $\mathcal{N}(0, 1)$ [24].

5. Variational inference SimSiam

This section introduces a novel uncertainty-aware SSL method by extending SimSiam based on the *SSL as inference* principle. The previous derivation assumes that κ is constant. We relax this assumption and allowed κ to be estimated by a newly introduced encoder $u_\theta : x \to \kappa$. Additionally, we replace the vMF distribution with another spherical distribution, i.e., Power Spherical (PS) distribution [10]. This is because estimating gradients through the modified Bessel function in $C_{\text{vMF}}(\kappa)$ is computationally expensive and unstable [10]. The modified posterior is defined as:

$$p(z|\mathbb{X}, \theta) \coloneqq \eta_\theta \prod_{j=1}^{M} \text{PS}(z; \mu = g_\theta(x_j), \kappa = u_\theta(x_j)), \quad (31)$$

$$\text{PS}(z; \mu, \kappa) \coloneqq C_{\text{PS}}(\kappa)(1 + \mu^\mathsf{T} z)^\kappa. \quad (32)$$

The normalization constant of the PS distribution $C_{\text{PS}}(\kappa)$ is defined with the beta distribution. It can be efficiently computed using general deep-learning frameworks. Substituting Eqs. (19) and (31) into Eq. (11) yields:

$$\mathcal{J}_{\text{align}} \overset{+}{\simeq} \sum_{i,j} \left(\log C_{\text{PS}}\left(u_\theta(x_j)\right) + u_\theta(x_j) \,\text{log1p}\left(g_\theta(x_j)^\mathsf{T} f_\phi(x_i)\right)\right), \quad (33)$$

where $\text{log1p}(x) \coloneqq \log(1 + x)$. Similar to Eqs. (1) and (22), this loss maximizes the cosine similarity of features from different models, but the similarity term is weighted by κ.

A novel method referred to as variational inference SimSiam (VI-SimSiam) that optimizes Eq. (33) is proposed. The architecture and pseudo code are shown in Fig. 2 and Alg. 1, respectively, where the number of modalities is $M = 2$ and the single-modal comparisons are ignored.

6. Experiments

We evaluate VI-SimSiam for two aspects-performance and method of uncertainty prediction. First, we compare the performance of VI-SimSiam and SimSiam. We perform a linear evaluation of these methods on ImageNet100 [11][4]. Second, we investigate representation uncertainty. We qualitatively evaluate representation uncertainty by comparing input images and the predicted uncertainty parameter κ. Then, we examine the relationship between uncertainty and classification accuracy. We also study how DINO predicts representation uncertainty.

6.1. Linear evaluation

We conduct self-supervised pretraining with ImageNet100 dataset without labels to learn image representations using SimSiam and VI-SimSiam at 50, 100, 200, and

[4]ImageNet100 is a 100-category subset of ImageNet [11].

Algorithm 1 Pseudocode of VI-SimSiam, PyTorch-like

```
# backbone: ResNet-backbone
# f: Projector
# h, u: Predictor for mu and kappa

# Definition of power spherical dist.
class PSd(Distribution):
  def _init_(self, mu, kappa):
     ...

for x in loader: # Load a minibatch x
  x1, x2 = aug(x), aug(x) # Apply augmentation
  y1, y2 = backbone(x1), backbone(x2)
  z1, z2 = f(y1), f(y2) # Project

  mu1, mu2 = h(z1), h(z2) # Predict mu
  kappa1, kappa2 = u(y1), u(y2) # Predict kappa
  p1, p2 = PSd(mu1, kappa1), PSd(mu2, kappa2)

  # Stop gradient & compute loss
  z1, z2 = z1.detach(), z2.detach()
  L = - p1.log_prob(z2) - p2.log_prob(z1)
  L.mean().backward() # Back-prop.
  update(backbone, f, h, u) # SGD update
```

Table 2: Top-1 accuracies of linear evaluation on ImageNet100. Each setting is repeated in triplicate to compute mean and standard derivation. We report the result as "mean $\pm$ std."

Method	50 epochs	100 epochs	200 epochs	500 epochs
SimSiam [6]	41.07$\pm$0.99	63.86$\pm$1.48	**78.49**$\pm$0.89	**81.61**$\pm$0.23
VI-SimSiam	**63.89**$\pm$2.74	**73.78**$\pm$0.58	76.31$\pm$0.32	77.49$\pm$0.42

500 epochs. Then, we trained a linear classifier on frozen representations on the training set of the dataset with the labels. Finally, we evaluated it in the test set of the dataset. The implementation details are reported in Appx. B. We used Top-1 accuracy as an evaluation metric.

Table 2 shows Top-1 accuracy on the validation split of ImageNet100. VI-SimSiam achieves a competitive result to SimSiam in several epochs. VI-SimSiam significantly outperforms SimSiam, especially when the number of epochs is less, e.g., 50 and 100.

6.2. Qualitative analysis of uncertainty

We evaluate uncertainty estimation qualitatively by comparing an input image to the predicted concentration κ, a parameter related to the uncertainty. We use images from the validation split of ImageNet100. We use the model pretrained with ImageNet100 on 100 epochs. Fig. 1 shows the images for which κ is in the top 1% and those for which κ is in the bottom 1%. When κ is low, i.e., the uncertainty of the latent variable is high, there are few noticeable features in the input images. This result shows that our method can estimate high uncertainty for semantically uncertain images. Additional image examples estimated to have high or low uncertainty are shown in Appx. L.

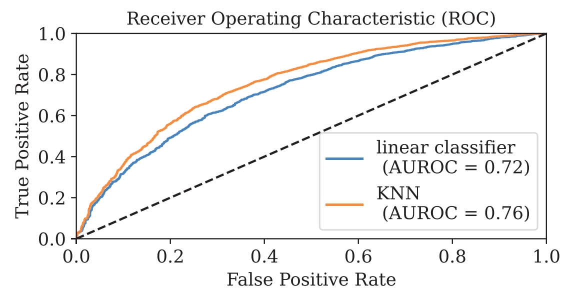

Figure 4: ROC and AUROC when the κ predicted by VI-SimSiam is used as the confidence score.

6.3. Quantitative analysis of uncertainty

We investigate the effects of uncertainty on a classification task with ImageNet100 dataset. We evaluate the relationship between uncertainty and Top-1 accuracy using the AUROC (Area Under the Receiver Operating Characteristics Curve) metrics following as per Ref. [13]. Fig. 4 shows ROC and AUROC when the κ predicted by VI-SimSiam is used as the confidence score. We use the VI-SimSiam-trained 500 epochs on ImageNet100. We use two classifiers – a linear classifier and a k-nearest-neighbor (KNN) classifier [56]. The AUROC score of 0.72 by the linear classifier is greater than the chance rate of 0.5 $(p < 0.01)$[5]. Furthermore, the AUROC score of KNN is 0.76. The score of the linear classifier is lower than that of the KNN, possibly due to the presence of epistemic uncertainty in the linear classifier itself. These results suggest we can estimate the difficulty of classifying an image based on uncertainty-related parameters without training the classification model or having the correct labels.

6.4. Relationship between DINO and uncertainty

We consider that DINO can estimate uncertainty because it fully estimates distributions, similar to VI-SimSiam. In this section, we discuss how DINO expresses uncertainty. We assume that the entropy of the latent variable, related to the variability of the distribution such as κ, has a relationship with uncertainty.

To discuss this, we evaluate the relationship between its entropy and Top-1 accuracy using the AUROC metrics. We use the negative entropy of representation as the confidence score. The implementation details are reported in Appx. B. We use ImageNet100 dataset, and the number of epochs is set to 200. The results are shown in Fig 5. The AUROC score of 0.67 by linear classification is greater than the chance rate of 0.5 $(p < 0.01)$.

Fig. 6 shows the scatter plot of κ by VI-SimSiam and the entropy of representation by DINO. Their correlation

[5] How to calculate the p-value is mentioned in Appx. J.

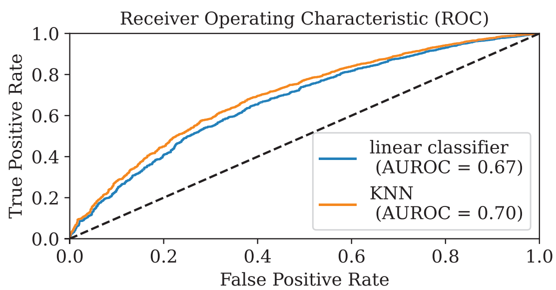

Figure 5: ROC and AUROC when the negative entropy predicted by DINO is used as the confidence score.

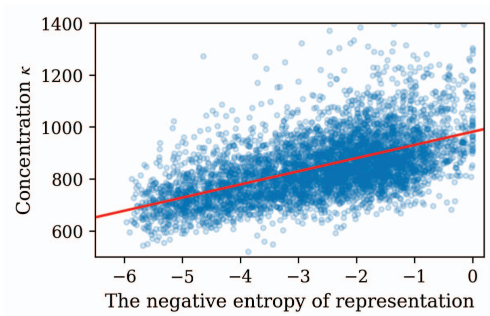

Figure 6: Scatter plot of κ of VI-SimSiam and entropy of DINO. The solid red line is the linear regression line. Their correlation coefficient is 0.5459.

Table 3: Mean and standard derivation of κ for each augmentation. The term "normal" means without augmentations.

normal	blur	color jitter	grayscale	random crop
$852.63\pm_{143.69}$	$836.05\pm_{147.96}$	$843.94\pm_{144.59}$	$846.14\pm_{140.53}$	$812.26\pm_{159.05}$

coefficient is 0.5459. The results show that entropy is highly related to κ ($p < 0.01$). Therefore, the entropy of representation by DINO seems to be related to uncertainty.

6.5. Relation to image augmentation

We investigate how image augmentations affect uncertainty. We use the validation dataset of ImageNet100. We also prepare image augmentations for *blur*, *color jitter*, *grayscale*, and *random crop*. The random crop scale for *random crop* is set from 0.05 to 1.0. We perform five augmentations for each image and calculate the average concentration κ estimated by the pre-trained model for each augmentation. The results are shown in Table 3. The variance of the κ of the *random crop* is greater, and its mean value is less than those of other augmentations. Fig. 7 shows examples of images that applied *random crop*. This result shows κ of images excluding important objects by cropping is low. Therefore, *random crop* is considered to have a more significant impact on κ than the other augmentations.

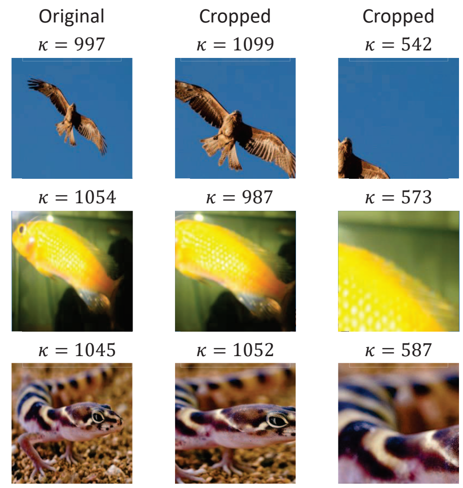

Figure 7: A set of cropped images and κ's; the images on the left are original. The images in the center and on the right are cropped images. This result shows that the κ values of images excluding important objects by cropping are low.

6.6. Limitation

Performance of Linear evaluation. In Sec 6.1, VI-SimSiam underperforms at 200 and 500 epochs, when compared with SimSiam. When the cosine similarity is low during training, SimSiam only learns to increase the cosine similarity, while VI-SimSiam learns to also decrease κ. It is assumed that VI-SimSiam underperforms at greater epochs because it is not trained to increase cosine similarity for inputs with representations that are difficult to predict. We compare the cosine similarity of SimSiam and VI-SimSiam in Appx. K. Our method would be more effective if a solution to this issue is proposed.

Utilization of Representation Uncertainty. Our future work involves proposing novel applications using representation uncertainty. One example of such use of uncertainty is in the selection of data, the prototype results of which are reported in Appx. E.

7. Conclusion

In this work, we clarify the theoretical relationship between variational inference and SSL. Additionally, we propose variational inference SimSiam (VI-SimSiam), which could model the latent variable's uncertainty. We investigate the estimated uncertainty parameter κ from various perspectives. We derive the relationship between κ and input images and between κ and classification accuracy. We also experimentally demonstrate that uncertainty could be estimated even when the latent variable follows the categorical distribution.

References

[1] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006. 6

[2] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *ECCV*, 2014. 13

[3] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *NeurIPS*, 2020. 11

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 1, 2, 5, 6

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 1, 2, 5

[6] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, 2021. 1, 2, 3, 5, 6, 7, 13

[7] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, 2021. 1

[8] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *CVPR*, 2014. 13

[9] Tim R Davidson, Luca Falorsi, Nicola De Cao, Thomas Kipf, and Jakub M Tomczak. Hyperspherical variational auto-encoders. In *UAI*, 2018. 2

[10] Nicola De Cao and Wilker Aziz. The power spherical distrbution. *ICML, INNF+*, 2020. 2, 6

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 6

[12] Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural safety*, 2009. 3

[13] Yukun Ding et al. Revisiting the evaluation of uncertainty estimation and its application to explore model complexity-uncertainty trade-off. In *CVPRW*, 2020. 7

[14] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *NeurIPS*, 2020. 1, 2, 3, 6

[15] Nitesh B Gundavarapu, Divyansh Srivastava, Rahul Mitra, Abhishek Sharma, and Arjun Jain. Structured aleatoric uncertainty in human pose estimation. In *CVPRW*, 2019. 3

[16] Wei Guo, Marc Brittain, and Peng Wei. Safety enhancement for deep reinforcement learning in autonomous separation assurance. In *ITSC*, 2021. 3

[17] David Hall, Feras Dayoub, John Skinner, Haoyang Zhang, Dimity Miller, Peter Corke, Gustavo Carneiro, Anelia Angelova, and Niko Sünderhauf. Probabilistic object detection: Definition and evaluation. In *WACV*, 2020. 3

[18] Manu Srinath Halvagal, Axel Laborieux, and Friedemann Zenke. Predictor networks and stop-grads provide implicit variance regularization in byol/simsiam. *arXiv:2212.04858*, 2022. 6

[19] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022. 2

[20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 1, 2

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 11

[22] Yihui He, Chenchen Zhu, Jianren Wang, Marios Savvides, and Xiangyu Zhang. Bounding box regression with uncertainty for accurate object detection. In *CVPR*, 2019. 3

[23] Eddy Ilg, Ozgun Cicek, Silvio Galesso, Aaron Klein, Osama Makansi, Frank Hutter, and Thomas Brox. Uncertainty estimates and multi-hypotheses networks for optical flow. In *ECCV*, 2018. 3

[24] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, pages 448–456, 2015. 6

[25] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *NeurIPS*, 2017. 3

[26] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *CVPR*, 2018. 3

[27] Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1, 2

[28] Nikos Komodakis and Spyros Gidaris. Unsupervised representation learning by predicting image rotations. In *ICLR*, 2018. 1

[29] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCVW*, 2013. 13

[30] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 13

[31] Richard Kurle, Stephan Günnemann, and Patrick Van der Smagt. Multi-source neural variational inference. In *AAAI*, 2019. 2

[32] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *TKDE*, 2021. 1

[33] Björn Lütjens, Michael Everett, and Jonathan P How. Safe reinforcement learning with model uncertainty estimates. In *ICRA*, 2019. 3

[34] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 13

[35] Kanti V Mardia, Peter E Jupp, and KV Mardia. *Directional statistics*, volume 2. Wiley Online Library, 2000. 4

[36] Sina Mohseni, Mandar Pitale, JBS Yadawa, and Zhangyang Wang. Self-supervised learning for generalizable out-of-distribution detection. In *AAAI*, 2020. 3

[37] Alejandro Newell and Jia Deng. How useful is self-supervised pretraining for visual tasks? In *CVPR*, 2020. 1

[38] M-E Nilsback and Andrew Zisserman. A visual vocabulary for flower classification. In *CVPR*, 2006. 13

[39] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, 2016. 1

[40] Masashi Okada, Norio Kosaka, and Tadahiro Taniguchi. Planet of the bayesians: Reconsidering and improving deep planning network by incorporating bayesian inference. In *IROS*, 2020. 3

[41] Masashi Okada, Shinji Takenaka, and Tadahiro Taniguchi. Multi-person pose tracking using sequential monte carlo with probabilistic neural pose predictor. In *ICRA*, 2020. 3

[42] Tongyao Pang, Yuhui Quan, and Hui Ji. Self-supervised bayesian deep learning for image recovery with applications to compressive sensing. In *ECCV*, 2020. 3

[43] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, 2012. 13

[44] Ilia Petrov, Vlad Shakhuro, and Anton Konushin. Deep probabilistic human pose estimation. *IET Computer Vision*, 2018. 3

[45] Matteo Poggi, Filippo Aleotti, Fabio Tosi, and Stefano Mattoccia. On the uncertainty of self-supervised monocular depth estimation. In *CVPR*, 2020. 3

[46] Tyler R Scott, Andrew C Gallagher, and Michael C Mozer. von mises-fisher loss: An exploration of embedding geometries for supervised learning. In *ICCV*, 2021. 3

[47] Yuge Shi, Brooks Paige, Philip Torr, et al. Variational mixture-of-experts autoencoders for multi-modal deep generative models. *NeurIPS*, 2019. 2, 4

[48] Thomas M. Sutter, Imant Daunhawer, and Julia E Vogt. Generalized multimodal ELBO. In *ICLR*, 2021. 2, 4, 6

[49] Chenxin Tao, Honghui Wang, Xizhou Zhu, Jiahua Dong, Shiji Song, Gao Huang, and Jifeng Dai. Exploring the equivalence of siamese self-supervised learning via a unified gradient framework. In *CVPR*, 2022. 2, 5

[50] Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In *ICML*, 2021. 1, 2, 4, 11

[51] Jakub Tomczak and Max Welling. Vae with a vampprior. In *AISTATS*, 2018. 2

[52] Jianren Wang, Siddharth Ancha, Yi-Ting Chen, and David Held. Uncertainty-aware self-supervised 3d data association. In *IROS*, 2020. 3

[53] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *ICML*, 2020. 2, 5

[54] Jim Winkens, Rudy Bunel, Abhijit Guha Roy, Robert Stanforth, Vivek Natarajan, Joseph R Ledsam, Patricia MacWilliams, Pushmeet Kohli, Alan Karthikesalingam, Simon Kohl, et al. Contrastive training for improved out-of-distribution detection. *arXiv preprint arXiv:2007.05566*, 2020. 3

[55] Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. *NeurIPS*, 2018. 2

[56] Zhirong Wu, Yuanjun Xiong, X Yu Stella, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018. 7

[57] Jianxiong Xiao, Krista A Ehinger, James Hays, Antonio Torralba, and Aude Oliva. Sun database: Exploring a large collection of scene categories. *IJCV*, 2016. 13

[58] 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. 13

[59] Hongbin Xu, Zhipeng Zhou, Yali Wang, Wenxiong Kang, Baigui Sun, Hao Li, and Yu Qiao. Digging into uncertainty in self-supervised multi-view stereo. In *ICCV*, 2021. 3

[60] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017. 11

[61] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *ICML*, 2021. 1, 2, 6

[62] Qi Zhang, Yifei Wang, and Yisen Wang. How mask matters: Towards theoretical understandings of masked autoencoders. In *NeurIPS*, 2022. 2