Can Generative Models Improve Self-Supervised Representation Learning?

Sana Ayromlou[1], **Vahid Reza Khazaie**[1], **Fereshteh Forghani**[2*], **Arash Afkanpour**[1]
[1]Vector Institute
[2]York University
sana.ayromlou@vectorinstitute.ai
vahidreza.khazaie@vectorinstitute.ai
forghani@yorku.ca
arash.afkanpour@vectorinstitute.ai

Abstract

The rapid advancement in self-supervised representation learning has highlighted its potential to leverage unlabeled data for learning rich visual representations. However, the existing techniques, particularly those employing different augmentations of the same image, often rely on a limited set of simple transformations that cannot fully capture variations in the real world. This constrains the diversity and quality of samples, which leads to sub-optimal representations. In this paper, we introduce a framework that enriches the self-supervised learning (SSL) paradigm by utilizing generative models to produce semantically consistent image augmentations. By directly conditioning generative models on a source image, our method enables the generation of diverse augmentations while maintaining the semantics of the source image, thus offering a richer set of data for SSL. Our extensive experimental results on various joint-embedding SSL techniques demonstrate that our framework significantly enhances the quality of learned visual representations by up to 10% Top-1 accuracy in downstream tasks. This research demonstrates that incorporating generative models into the joint-embedding SSL workflow opens new avenues for exploring the potential of synthetic data. This development paves the way for more robust and versatile representation learning techniques.

Introduction

Self-supervised learning (SSL) is a machine learning approach where methods leverage large amounts of unlabeled data for representation learning. In the absence of any prior knowledge, most SSL methods assume that each data point is semantically different from other examples in a dataset. While this assumption can help devise a self-supervised representation learning algorithm, it is not sufficient to guarantee the generalization of representations to new tasks. As a result, *joint-embedding* SSL methods (e.g., Chen et al. (2020); He et al. (2020); Grill et al. (2020)) utilize augmentation techniques to generate two or more views of the same image and map them to nearby points in the representation space. These augmentations typically consist of a series of transformations, such as random crop, color jitter, or Gaussian blur, as shown in Fig. 1(a), which are used to define *positive labels* for the SSL task. These augmentations determine pixel-space changes that should result in similar embedding-space representations. Since these transformations specify what representations learn, a natural question is whether additional transformations improve the generalization and robustness of representations.

While some previous works have studied the impact of augmentations in representation learning (e.g., Chen et al. (2020), Caron et al. (2020)), until recently, this area has remained largely under-explored. More recently, new image transformations have been introduced in the context of self-supervised representation learning. For example, Mansfield et al. (2023) introduced random field transformations by applying local affine and color transformations whose parameters are specified by Gaussian random fields (Adler, Taylor et al. 2007). Rojas-Gomez et al. (2023) proposed style transfer as a form of transformation for SSL. While these works show that new transformations can improve representation learning, their transformations remain limited to certain predefined forms.

An important feature of SSL transformations is that while they transform an image in pixel space, they do not significantly modify its semantics. Guided by this view, we propose the notion of a *generalized transformation*, which satisfies these conditions: (1) the transformed image should be semantically similar to the original image and (2) it should not make modifications that render the image unrealistic, which often happens by some of the standard transformations (e.g., color jitter). The question arises: given a source image, how can one generate transformed images that satisfy these conditions? For this task, we leverage instance-conditioned generative models. In particular, we study latent diffusion models (Rombach et al. 2022) and conditional Generative Adversarial Networks (GAN) (Casanova et al. 2021) that generate an output that is semantically similar to a source image. Our extensive empirical study in in-distribution and out-of-distribution cases demonstrates the effectiveness of our method for representation learning. The code to reproduce our empirical study is available at https://github.com/VectorInstitute/GenerativeSSL. The key contributions of our work are as follows:

- We propose to leverage instance-conditioned generative models for generating semantically similar augmentations in joint-embedding SSL, advancing beyond the standard

*Work done during an internship at Vector Institute.

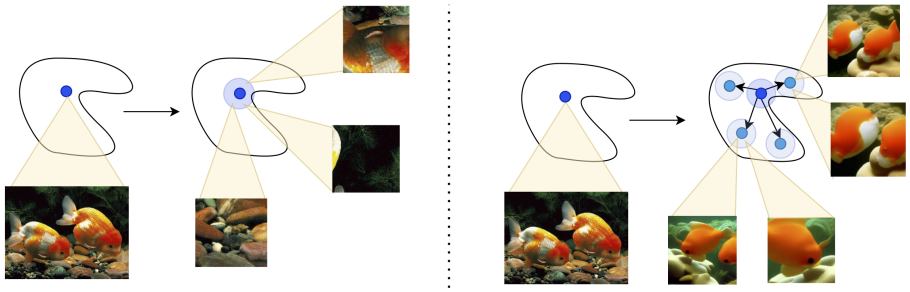

Figure 1: Generative augmentations produce a more diverse set of images with similar semantics. **a)** The standard SSL augmentations offer limited diversity for effective representation learning. **b)** By generating instance-conditioned samples, and then applying the standard augmentations on top, we add more diversity in training data, leading to better representations.

augmentation techniques.

- We perform a rigorous and extensive empirical study of the instance-conditioned generative augmentation, demonstrating its effectiveness for learning representations with high generalization across in-distribution and out-of-distribution tasks.
- By relying on instance-conditioned image generation, we eliminate the need for text-based image generation, as suggested in Tian et al. (2024). This enables the application of our method to datasets without a textual description, resulting in a more versatile and effective strategy for SSL training.

Background

Self-Supervised Learning

Self-supervised learning (SSL) leverages unlabeled data to create rich and generalizable representations. This approach offers numerous advantages due to its ability to train on extensive unlabeled data (Balestriero et al. 2023). A line of SSL methods, known as *joint-embedding* methods, encourages two views of the same image, formed by augmentations such as cropping or color jitter, to be mapped to similar representations. This approach enables encoders to learn to differentiate the representation of each image, leading to decent linear separability.

These methods, based on their objective functions or architectures, can be categorized as follows (Balestriero et al. 2023): 1) *Contrastive learning* methods employ the InfoNCE contrastive loss (Oord, Li, and Vinyals 2018), which in an unsupervised manner, pulls representations of different transformed versions of the same image together while pushing representations of different images apart. **SimCLR** (Chen et al. 2020) uses the examples in each batch to calculate the contrastive loss. **MoCo** (He et al. 2020) is another method that learns an encoder by matching an encoded query to a dictionary of encoded keys using contrastive loss. 2) *Self-distillation* methods rely on a straightforward mechanism of feeding two different views to two encoders and mapping one to the other using a predictor (Zhou et al. 2021; Caron et al. 2021; Oquab et al. 2023). They prevent collapse by introducing asynchrony in updating the encoders. **BYOL** (Grill et al. 2020) first introduced self-distillation as a means to avoid collapse. **SimSiam** (Chen and He 2021) replaces the BYOL moving average encoder with a stop-gradient mechanism. 3) *Canonical correlation analysis* methods originate from the Canonical Correlation Framework (CCA) (Hotelling 1992; Caron et al. 2018; Bardes, Ponce, and LeCun 2021). At a high level CCA aims to infer the relationship between variables by analyzing the covariance matrix. **Barlow Twins** (Zbontar et al. 2021) uses cross-correlation of the embedding features to achieve self-supervised learning goals. In this paper, we evaluate the effectiveness of our proposed generative augmentation when integrated into the augmentation pipeline of the highlighted techniques (in bold).

Generative Methods

Generative Adversarial Networks (GAN) (Goodfellow et al. 2014) have significantly advanced the field of image generation. Recently, some work has been done on conditioning GANs over a feature vector (Mangla et al. 2022; Casanova et al. 2021). In this work, we chose Instance-Conditioned GANs (ICGAN) (Casanova et al. 2021), which extends this framework by conditioning the generation and discrimination processes on a specific image. ICGAN aims to model the data distribution as a mixture of local densities around each instance, utilizing representations of these instances as additional input to both the generator and the discriminator. This conditioning allows the model to generate images that are semantically similar to a given instance. This approach addresses the optimization difficulties and mode collapse issues prevalent in traditional GANs by ensuring a more focused generation process. This method can be leveraged to create semantically consistent augmentations for SSL.

Diffusion Models, a class of likelihood-based generative models, have gained attention for their ability to generate high-quality images (Dhariwal and Nichol 2021; Ho, Jain, and Abbeel 2020; Nichol and Dhariwal 2021). They work by gradually reducing noise from an image, with their training objective expressed as a re-weighted variational lower-bound (Ho, Jain, and Abbeel 2020). Despite their ability to generate high-quality images, due to their incremental noise reduction process, these models have traditionally faced chal-

lenges of long training and inference times. While inference challenges can be mitigated with advanced sampling strategies (Salimans and Ho 2022; Song, Meng, and Ermon 2020) and hierarchical approaches (Ho et al. 2022), training on high-resolution image data remains computationally expensive. To address this, latent diffusion models (LDM) have been introduced (Rombach et al. 2022). These models operate in a latent space with lower dimensionality than the input space, making training computationally cheaper and speeding up inference with almost no reduction in synthesis quality. They also introduced cross-attention layers into the model architecture, where various conditioning inputs, such as text, images, or other representations, can be applied to get a conditional output. The conditional LDM, ϵ_θ, which is usually a UNet (Ronneberger, Fischer, and Brox 2015), and a domain-specific encoder, ϕ_τ, that provides the conditioning vector, are learned jointly by optimizing,

$$\theta^*, \tau^* = \arg\min_{\theta,\tau} \mathbb{E}_{\epsilon\sim\mathcal{N}(0,I),t,y}\left[\|\epsilon - \epsilon_\theta(z_t, t, \phi_\tau(y))\|_2^2\right],$$

where t is the timestep parameter, z_t is the noisy latent vector at timestep t, and y is the conditioning input (e.g. a text prompt). Given their ability to generate high-quality diverse samples, diffusion models stand out as a compelling option for our new generative augmentation.

Related Work

With the advance of generative models, the use of synthetic data has become a tool for training better models in various domains. In computer vision, synthetic data has been utilized to enhance model performance in tasks such as object detection (Lin et al. 2023), semantic segmentation (Chen et al. 2019; Ros et al. 2016), and classification (Azizi et al. 2023; Sariyildiz et al. 2023). In representation learning, synthetic data has been leveraged for multi-task learning (Ren and Lee 2018), and manipulating latent data representation (Liu et al. 2022; Baradad Jurjo et al. 2021; Jahanian et al. 2021). Jahanian et al. (2021) investigated learning visual representations from synthetic data generated by black-box models, emphasizing the ability of latent space transformations to facilitate contrastive learning. In contrast to our work, they rely on latent space transformations to generate multiple views of the same semantic content. In addition, past works (Astolfi et al. 2023; Zang et al. 2024; Li et al. 2025) present different methods for leveraging generative models for contrastive learning. Astolfi et al. (2023) proposes an approach similar to ours for representation learning using ICGAN models. In contrast, our work presents a more rigorous study covering both ICGAN and Stable Diffusion as generative models. The works of Donahue, Krähenbühl, and Darrell (2016); Donahue and Simonyan (2019) introduce Bidirectional Generative Adversarial Networks (BiGAN) and BigBiGAN, which incorporate an inverse mapping mechanism within the GAN framework to enable both forward and inverse mappings between latent space and data. This allows for capturing semantic variations more effectively and making the representations valuable for auxiliary tasks. More recently, StableRep (Tian et al. 2024) showed how visual representations learned from synthetic text-to-image data can be more effective than representations learned by using only real images in various downstream tasks. Unlike them, we do not replace a real dataset with a synthetic one. Instead, we leverage conditional generative models to enrich augmentations for self-supervised learning. Another distinction is that our method does not require text prompts and directly uses images as condition for the generative model.

Generative Self-Supervised Learning

Joint-embedding SSL methods create two or more views by applying a series of transformations to an image. More formally, given a batch of images, $\mathcal{B} = \{x_1, \ldots, x_N\}$, joint-embedding SSL methods learn an encoder, f_θ, by optimizing a loss function, $\mathscr{L}$, over the batch, $\mathcal{B}$:

$$\theta^*, \psi^* = \arg\min_{\theta,\psi} \mathbb{E}_{A_1,A_2,\mathcal{B}}\Big[\mathscr{L}\Big(g_\psi(f_\theta(A_1(\mathcal{B}))), g_\psi(f'_\theta(A_2(\mathcal{B})))\Big)\Big],$$

where A_1 and A_2 are augmentations applied to the images in $\mathcal{B}$ to create two views of each image, g_ψ is a projection/prediction head applied to the output of the encoder before calculating the loss, and f' is a variant of f that, depending on the SSL algorithm, could be for instance, identical to f (SimCLR), its exponential moving average (MoCo), or its frozen version (SimSiam).

The augmentations, A_1 and A_2, are a series of K sequentially-applied transformations, $A_j = T_K^{\gamma_K} \circ \ldots \circ T_1^{\gamma_1}$, where $T_i^{\gamma_i}$ is parameterized by γ_i. For example, a crop transformation could be parameterized by crop area, aspect ratio, etc. While parameterization of transformations increases the versatility of the views, they are often limited in producing semantically rich and diverse variations of the data. For example, previous SSL methods apply transformations such as random crop, color jitter, horizontal flip, grayscale (color removal), and Gaussian blur. Adding transformations that create more diversity in the output of the augmentation pipeline could potentially improve the generalization of learned representations.

Due to their limited form, standard transformations might not adequately represent the intrinsic variability found in real-world data. To overcome this limitation, we expand SSL transformations with a non-parametric transformation that is capable of enriching diversity of data. The only constraint is that to learn useful representations, a transformed image must remain semantically similar to its source image.

As illustrated in Fig. 2, the foundation of this transformation lies in conditional generative models. These models, such as conditional latent diffusion models (Rombach et al. 2022; Bordes, Balestriero, and Vincent 2021), instance-conditioned GANs (Casanova et al. 2021), and conditional Variational Autoencoders (cVAEs) (Zhang, Barbano, and Jin 2021) enable the generation of images based on specific input data. By conditioning the generation process on the input, the model ensures that the semantics of the original image is preserved to a large extent. Moreover, the use of conditional generative models allows for a more nuanced augmentation process, where the generated samples closely follow the distribution of real images while exhibiting variations that enhance the

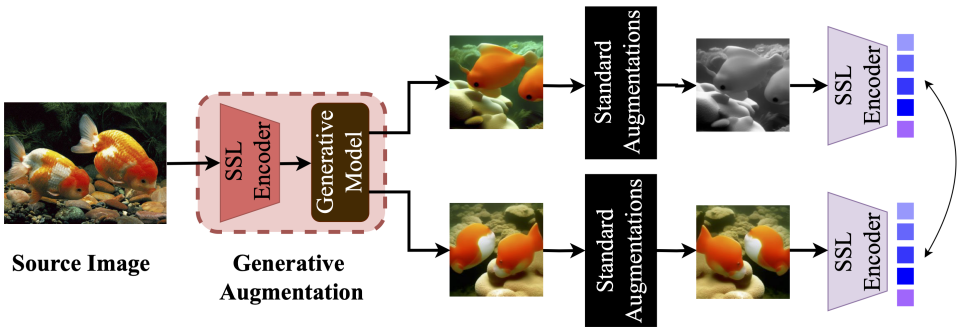

Figure 2: Our augmentation pipeline utilizes generative models, i.e., Stable Diffusion or ICGAN, conditioned on the source image representation, accompanied by the standard SSL augmentations. The components inside the *Generative Augmentation* module, i.e. the pretrained SSL encoder and the generative model remain frozen throughout the SSL training process.

diversity of the training data for self-supervised learning. Our new transformation is denoted by:

$$T_0(x) = \begin{cases} G(z; \phi(x)) & \text{if} \quad p \leq p_0 \\ x & \text{otherwise,} \end{cases}$$

where G denotes a conditional generative model, $z \sim \mathcal{N}(0, I)$ is a noise vector, ϕ is a pretrained encoder, such as a CLIP image encoder (Radford et al. 2021), that provides the condition vector for the generative model, $p \in [0, 1]$ is a random number, and p_0 is a parameter specifying the probability of applying the generative augmentation.

We emphasize that the new transformation and variants of it have been studied in a broader context of creating augmentations or views with generative models in previous work. See, for example, Astolfi et al. (2023); Zang et al. (2024); Li et al. (2025); Jahanian et al. (2021). In particular, Astolfi et al. (2023) proposed a very similar augmentation using ICGAN models. What differentiates our work is that we perform a more extensive and rigorous study of this concept in the context of joint-embedding representation learning with two generative models across five SSL techniques and a suite of in-distribution and out-of-distribution tasks.

Empirical Study

Setup

Utilizing the Solo-learn library (Da Costa et al. 2022), we evaluate the proposed generative augmentation across five joint-embedding SSL techniques: SimCLR (Chen et al. 2020), SimSiam (Chen and He 2021), BYOL (Grill et al. 2020), Barlow Twins (Zbontar et al. 2021), and MoCo (He et al. 2020). We pretrain a ResNet50 encoder using these SSL techniques for 100 epochs on the ImageNet training split. For evaluation, we follow the linear probing protocol of previous works by training a linear classifier on the output of the frozen encoder. Our downstream tasks are image classification on the ImageNet (Deng et al. 2009), Food 101 (Bossard, Guillaumin, and Van Gool 2014), Places 365 (Zhou et al. 2017), iNaturalist 2018 (Van Horn et al. 2018), CIFAR 10, and CIFAR 100 (Krizhevsky and Hinton 2009) datasets. For all datasets, except for Places 365, a linear classifier is trained on the corresponding training split for 100 epochs. For Places 365, training is performed for 45 epochs. After training the linear classifier, we evaluate the classifier on the corresponding validation set.

In the following experiments, *Baseline* refers to the model that uses standard augmentations to create the views. The sequence of transformations to create each view is as follows: (1) random crop with the relative crop area selected randomly from $[0.2, 1]$, (2) scale to 224×224, (3) color jitter, (4) grayscale, (5) Gaussian blur, (6) horizontal flip. Each augmentation is applied stochastically with a probability value. More details are available in the Appendix. For our method, we prepend the above sequence with a generative augmentation that takes a source image and, depending on the experiment, uses either Stable Diffusion, which is a conditional latent diffusion model, or ICGAN, to return a synthetic image as described in Section . To accelerate training, instead of generating synthetic images on-the-fly, we generate 10 images per each image in the ImageNet training set offline. Furthermore, with a given probability p we apply the generative augmentation by selecting and loading one of these augmentations during SSL training.

Results

We frame our experiments to answer four research questions about our proposed generative augmentation technique.

RQ1: What probability of applying the generative augmentation achieves the best visual representations? Each image transformation in the standard augmentations is applied with a certain probability value tuned to achieve the best visual representations. While higher probability values create more diversity in augmentations, they could potentially hinder fidelity or create undesirable artifacts that result in sub-optimal representations.

For this experiment we used Stable Diffusion for synthetic image generation and applied the new augmentation with probability $p \in \{0, 0.25, 0.5, 0.75, 1\}$. For each value of p, we trained a SimCLR encoder, and compared the encoders with linear probing on the ImageNet validation split. Figure 4 shows the results. We ran two sets of experiments. In one set, we applied the generative augmentation only to one view

(dark blue curve), while in the other set we apply this augmentation to both views (light blue curve). In all cases the standard augmentations are applied to both views. For the *one* view, the results show a monotonic increase in the quality of the visual representations as p increases with $p = 0.75$ and $p = 1$ achieving the best results. These results show the effectiveness of the new augmentation for representation learning. On the other hand, the best results are achieved with $p = 0.5$ for *both* view, as it strikes a balance between real and synthetic images in the views. However, increasing the probability of the generative augmentation beyond that will degrade performance as it likely creates excessive deviation from the original data. However, note that in this case $p = 1$ still outperforms no generative augmentation ($p = 0$), indicating the effectiveness of the new augmentation.

RQ2: Does the generative transformation improve representation learning consistently across different joint-embedding SSL techniques? We evaluate the effectiveness of the generative transformation across five SSL techniques. We compare an augmentation pipeline that includes the generative transformation obtained by either Stable Diffusion (Stable Diff) or ICGAN against the Baseline, which refers to pretraining the SSL encoder with only the standard augmentations. Based on the results depicted in RQ1, we apply the generative augmentation with probability $p = 0.5$ to both views. Fig. 5 shows the linear probing Top-1 accuracy for different SSL techniques. Additionally, Top-5 accuracy results are reported in the Appendix. Across all techniques, the addition of a generative transformation via Stable Diffusion consistently improves representation learning, which in turn enhances downstream classification accuracy between 3% to 10%. With the exception of Barlow Twins, we observe a similar trend with generative augmentations generated by ICGAN. However, the magnitude of improvement is smaller compared to Stable Diffusion. We conjecture that this is because the quality and diversity of images generated by Stable Diffusion better reflect the distribution of real images compared to those generated by ICGAN, as depicted by some examples in Fig. 3. These factors are often determined by the size of the model and the volume of pretraining data. Stable Diffusion is a larger model and has been trained on LAION-400M which contains 400 million images (Rombach et al. 2022), while ICGAN was trained on ImageNet with 1.2 million images.

To measure out-of-distribution performance of learned representations, we evaluate these models across other downstream classification tasks. Table 1 shows the Top-1 accuracy results. In line with the above results on ImageNet, the Stable Diffusion generative transformation outperforms other baselines across all datasets. ICGAN, with the exception of a few cases, also outperforms the baseline, but we observe larger improvement with Stable Diffusion compared to ICGAN. These results, along with the in-distribution results, indicate that the generative transformation consistently improves representation learning across different SSL techniques.

RQ3: Considering that the new generative augmentation employs a conditional generative process dependent on a pretrained encoder, does self-supervised learning with this transformation produce the same latent-space data distribution as the pretrained encoder? This question is crucial because our augmentation technique relies on conditional image generation using a pretrained encoder for image conditioning (e.g., a pretrained CLIP encoder for Stable Diffusion). If the distributions of representation vectors are identical or very similar, the new data augmentation would offer little to no advantage for representation learning, as such representations are already accessible through the pretrained SSL encoder. Note that in this case we cannot rely on downstream accuracy for two reasons: (1) comparing downstream accuracy of two encoders provides little insight into the similarity of their representation spaces, (2) the CLIP encoder is trained on 400 million image-text pairs (Radford et al. 2021), which is significantly larger than the ImageNet training split (1.2 million images) used for training our SSL encoders, rendering any downstream accuracy comparison unfair. Instead, to compare representations of different encoders we employ two dissimilarity measures, i.e. Centered Kernel Alignment (CKA) (Kornblith et al. 2019) and Orthogonal Procrustes Distance (OPD) (Ding, Denain, and Steinhardt 2021), which have been used to compare representations between layers of a network or between different trained models. Let $\mathcal{D} = \{x_i\}_{i=1}^N$ be a dataset of N examples, where each example $x_i \in \mathbb{R}^s$. Let $f_A : \mathbb{R}^s \to \mathbb{R}^p$ and $f_B : \mathbb{R}^s \to \mathbb{R}^q$ denote two encoders. Let $A \in \mathbb{R}^{p \times N}$ and $B \in \mathbb{R}^{q \times N}$ denote the representation matrices obtained by applying f_A and f_B to data points in $\mathcal{D}$ respectively. The dissimilarity measure based on CKA is defined by,

$$d_{\text{CKA}}(A, B) = 1 - \frac{\|AB^\top\|_F^2}{\|AA^\top\|_F \|BB^\top\|_F}.$$

The OPD dissimilarity measure is the solution to the following optimization problem:

$$d_{\text{OPD}}(A, B) = \min_R \|B - RA\|_F^2, \qquad \text{s.t. } R^\top R = \mathbb{I}.$$

Here, $\|.\|_F$ denotes Frobenius norm and $\mathbb{I}$ is the identity matrix. Per standard practice, we normalize each matrix by first centering the features around the origin, then dividing by the Frobenius norm of the matrix before calculating these measures.

For this analysis, we measure the dissimilarity of representations between CLIP (the encoder used for conditioning Stable Diffusion) and SimCLR encoders trained with our generative transformation. We use the ImageNet validation set as data points to build the representation matrices. Since d_{CKA} and d_{OPD} are measured between two sets, we first require to establish a baseline for the dissimilarity value between two sets of similar representations so we can compare other values against it. For this purpose, we trained two SimCLR encoders with different weight initialization. While we acknowledge that the representation spaces of these models might not be identical, the dissimilarity value between them can be used as a baseline for comparing other values. Table 2 presents the dissimilarity values between different encoders. Dissimilarity values between the SimCLR encoders measured by both CKA and OPD are significantly smaller than the values between each SimCLR encoder and the CLIP encoder.

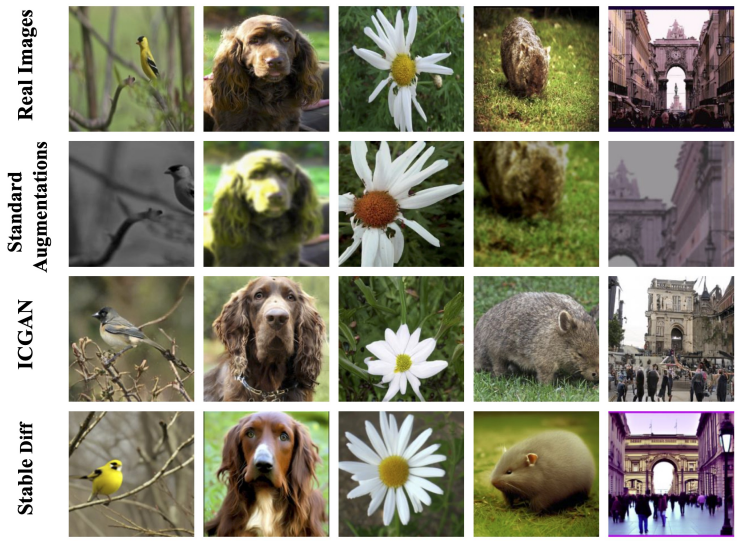

Figure 3: Examples of various augmentations. Compared to the standard augmentations (second row), instance-based generative augmentations can produce more diverse and realistic images that preserve the semantics of the original image.

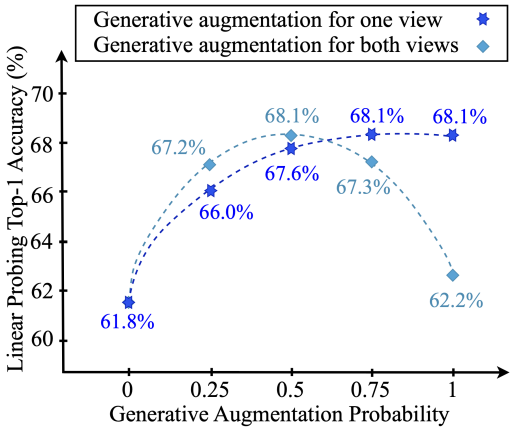

Figure 4: The effect of different probability values of applying the generative augmentation.

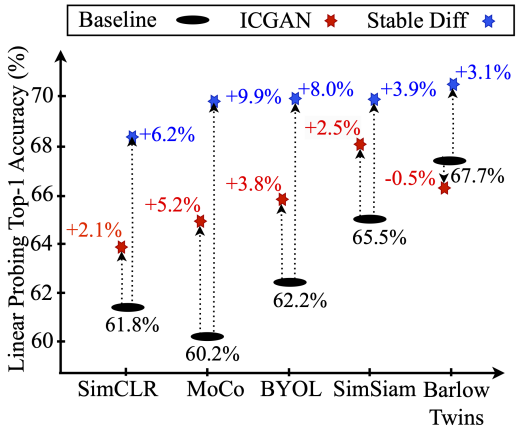

Figure 5: Top-1 accuracy improvement on ImageNet validation set obtained by the generative augmentations with Stable Diffusion and ICGAN across five SSL techniques.

These results demonstrate that the representations of SimCLR trained with the generative transformation are different from the representations of CLIP. In other words, representation learning with the proposed generative transformation does not trivially yield the same representation space of the pretrained encoder of the generative process.

RQ4: Can we replace the standard augmentations with the generative augmentation for self-supervised learning? Previous work on SSL augmentations has focused on enhancing the diversity of augmentations by adding new transformations to the pipeline. To the best of our knowledge, no attempt has been made to replace the standard augmentations with a new one, as it could significantly reduce the diversity of views and result in poor representations. Notably, Tian et al. (2024) observed that contrastive learning with a limited number of synthetic images benefits from the standard augmentations, as they help reduce overfitting. We investigate whether the new augmentation can replace the standard augmentations and whether keeping the standard augmentations in the pipeline is still beneficial for visual representation learning.

For this study we trained SimCLR models with different augmentation strategies and compared their representations by linear probing on the ImageNet validation set. Here, *Baseline* refers to the standard augmentations (crop and resize, color jitter, grayscale, Gaussian blur, and horizontal flip), and *Only Generative* refers to using only the new generative

		Food101	CIFAR10	CIFAR100	Places365	iNaturalist2018
SimCLR	Baseline	58.03	61.74	36.88	46.11	20.17
	ICGAN	57.62 (-0.41)	61.92 (**+0.18**)	38.74 (**+2.06**)	47.41 (**+1.67**)	20.01 (-0.16)
	Stable Diff	62.76 (**+4.73**)	63.35 (**+1.61**)	40.17 (**+3.29**)	48.83 (**+2.62**)	24.13 (**+3.96**)
MoCo	Baseline	54.50	58.17	34.28	45.70	16.52
	ICGAN	59.54 (**+5.04**)	63.70 (**+5.53**)	39.90 (**+4.62**)	48.82 (**+3.12**)	20.99 (**+4.47**)
	Stable Diff	63.97 (**+9.47**)	65.14 (**+6.97**)	41.21 (**+6.93**)	49.91 (**+4.21**)	26.29 (**+9.97**)
BYOL	Baseline	54.14	55.34	28.70	46.93	6.91
	ICGAN	53.12 (-0.98)	58.76 (**+3.42**)	34.49 (**+5.79**)	47.24 (**+0.31**)	7.37 (**+0.46**)
	Stable Diff	58.85 (**+4.71**)	61.90 (**+5.56**)	38.71 (**+10.01**)	49.09 (**+2.16**)	10.74 (**+3.83**)
SimSiam	Baseline	60.71	61.93	37.69	48.78	22.07
	ICGAN	65.15 (**+4.44**)	64.28 (**+0.35**)	40.21 (**+2.52**)	49.38 (**+0.60**)	27.07 (**+5.00**)
	Stable Diff	64.85 (**+4.14**)	63.88 (**+1.95**)	40.58 (**+2.89**)	49.66 (**+0.88**)	28.25 (**+6.18**)
Barlow Twins	Baseline	66.71	65.49	41.19	49.47	27.99
	ICGAN	65.50 (-1.21)	63.34 (-2.15)	41.95 (**+0.76**)	48.64 (-0.83)	26.19 (-1.80)
	Stable Diff	69.97 (**+3.26**)	65.20 (-0.29)	43.47 (**+2.28**)	50.31 (**+0.84**)	32.42 (**+4.43**)

Table 1: Top-1 accuracy (%) results of linear probing on several datasets with different augmentation strategies. Numbers inside parentheses show the absolute percentage difference from the baseline, with bold numbers indicating positive improvements.

Representation Pair	**CKA**	**OPD**
(SimCLR_1, SimCLR_2)	0.395 ± 0.006	0.022 ± 0.001
(SimCLR_1, CLIP)	0.964 ± 0.002	0.580 ± 0.002
(SimCLR_2, CLIP)	0.952 ± 0.002	0.574 ± 0.002

Table 2: CKA and OPD dissimilarity values between different representation pairs. We report the mean and 95% bootstrap confidence intervals over 100 runs.

Augmentation Strategy	**Top-1 Accuracy**
Baseline (only Standard)	61.84
Only Generative	63.80 (**+1.96**)
Generative & Random Crop	67.15 (**+5.31**)
Generative & Standard	68.11 (**+6.27**)

Table 3: The effect of different augmentation strategies on representation learning, measured by downstream Top-1 accuracy on the ImageNet validation set. Numbers inside parentheses show the absolute difference (%) from the baseline, with bold numbers indicating positive improvements..

augmentation obtained from Stable Diffusion. We also included another strategy, namely *Generative & Random Crop* that applies the generative transformation, followed by the random crop transformation. We particularly chose random crop as Chen et al. (2020) noted that this is the most effective augmentation among the standard ones. For completeness, we included *Generative & Standard* in the results, which applies the generative augmentation followed by the standard augmentations. Table 3 shows the linear probing Top-1 accuracy of these models. Applying the generative augmentation and excluding the standard augmentations results in 1.96% improvement in accuracy. However, the results of *Generative & Random Crop* and *Generative & Standard* indicate that further improvement is gained by including the standard augmentations. These results demonstrate that the standard augmentations cannot be replaced by the new generative augmentation and their combination achieves the best result.

Conclusion

Image augmentation is a crucial component of joint-embedding self-supervised learning. While many techniques rely on a limited set of transformations, recently, a few studies have proposed new parametric pixel-space transformations to enhance augmentations. In this paper, we studied a new generalized augmentation method that leverages instance-conditioned generative models to transform an image while preserving its semantics. By relying on strong generative models this transformation is capable of producing diverse and realistic images. Our empirical results demonstrate the effectiveness of the generative augmentations for self-supervised representation learning. Our method currently relies on pretrained generative models. An interesting direction to expand this work is to co-train an instance-conditioned generative model and the SSL encoder simultaneously. While care must be taken to avoid a possible collapse to trivial solutions, this approach could potentially lead to enhanced generative models for the SSL task.

Ethics Statement

Our method illustrates the potential for improving visual representation learning through the use of synthetic data. We recognize the risks of using synthetic data from pretrained generative models. Such models may have been trained on biased data, which could propagate or even amplify these biases in downstream applications. Data leakage is also possible if the generative model has not been trained with privacy-preserving methods. Our method, however, does not contribute to or exacerbate data leakage.

References

Adler, R. J.; Taylor, J. E.; et al. 2007. *Random fields and geometry*, volume 80. Springer.

Astolfi, P.; Casanova, A.; Verbeek, J.; Vincent, P.; Romero-Soriano, A.; and Drozdzal, M. 2023. Instance-conditioned gan data augmentation for representation learning. *arXiv preprint arXiv:2303.09677*.

Azizi, S.; Kornblith, S.; Saharia, C.; Norouzi, M.; and Fleet, D. J. 2023. Synthetic data from diffusion models improves imagenet classification. *arXiv preprint arXiv:2304.08466*.

Balestriero, R.; Ibrahim, M.; Sobal, V.; Morcos, A.; Shekhar, S.; Goldstein, T.; Bordes, F.; Bardes, A.; Mialon, G.; Tian, Y.; et al. 2023. A cookbook of self-supervised learning. *arXiv preprint arXiv:2304.12210*.

Baradad Jurjo, M.; Wulff, J.; Wang, T.; Isola, P.; and Torralba, A. 2021. Learning to see by looking at noise. *Advances in Neural Information Processing Systems*, 34: 2556–2569.

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

Bordes, F.; Balestriero, R.; and Vincent, P. 2021. High fidelity visualization of what your self-supervised representation knows about. *arXiv preprint arXiv:2112.09164*.

Bossard, L.; Guillaumin, M.; and Van Gool, L. 2014. Food-101–mining discriminative components with random forests. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, 446–461. Springer.

Caron, M.; Bojanowski, P.; Joulin, A.; and Douze, M. 2018. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, 132–149.

Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; and Joulin, A. 2020. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33: 9912–9924.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9650–9660.

Casanova, A.; Careil, M.; Verbeek, J.; Drozdzal, M.; and Romero Soriano, A. 2021. Instance-conditioned gan. *Advances in Neural Information Processing Systems*, 34: 27517–27529.

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

Chen, X.; and He, K. 2021. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15750–15758.

Chen, Y.; Li, W.; Chen, X.; and Gool, L. V. 2019. Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1841–1850.

Da Costa, V. G. T.; Fini, E.; Nabi, M.; Sebe, N.; and Ricci, E. 2022. solo-learn: A library of self-supervised methods for visual representation learning. *Journal of Machine Learning Research*, 23(56): 1–6.

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

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.

Ding, F.; Denain, J.-S.; and Steinhardt, J. 2021. Grounding representation similarity with statistical testing. *arXiv preprint arXiv:2108.01661*.

Donahue, J.; Krähenbühl, P.; and Darrell, T. 2016. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*.

Donahue, J.; and Simonyan, K. 2019. Large scale adversarial representation learning. *Advances in neural information processing systems*, 32.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33: 21271–21284.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Ho, J.; Saharia, C.; Chan, W.; Fleet, D. J.; Norouzi, M.; and Salimans, T. 2022. Cascaded diffusion models for high fidelity image generation. *The Journal of Machine Learning Research*, 23(1): 2249–2281.

Hotelling, H. 1992. Relations between two sets of variates. In *Breakthroughs in statistics: methodology and distribution*, 162–190. Springer.

Jahanian, A.; Puig, X.; Tian, Y.; and Isola, P. 2021. Generative models as a data source for multiview representation learning. *arXiv preprint arXiv:2106.05258*.

Kornblith, S.; Norouzi, M.; Lee, H.; and Hinton, G. 2019. Similarity of neural network representations revisited. In *International conference on machine learning*, 3519–3529. PMLR.

Krizhevsky, A.; and Hinton, G. 2009. Learning multiple layers of features from tiny images. Technical report, University of Toronto.

Li, X.; Yang, Y.; Li, X.; Wu, J.; Yu, Y.; Ghanem, B.; and Zhang, M. 2025. Genview: Enhancing view quality with pretrained generative model for self-supervised learning. In *European Conference on Computer Vision*, 306–325. Springer.

Lin, S.; Wang, K.; Zeng, X.; and Zhao, R. 2023. Explore the Power of Synthetic Data on Few-Shot Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 638–647.

Liu, H.; Zahavy, T.; Mnih, V.; and Singh, S. 2022. Palm up: Playing in the Latent Manifold for Unsupervised Pretraining. *Advances in Neural Information Processing Systems*, 35: 35880–35893.

Mangla, P.; Kumari, N.; Singh, M.; Krishnamurthy, B.; and Balasubramanian, V. N. 2022. Data instance prior (disp) in generative adversarial networks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 451–461.

Mansfield, P. A.; Afkanpour, A.; Morningstar, W. R.; and Singhal, K. 2023. Random Field Augmentations for Self-Supervised Representation Learning. *arXiv preprint arXiv:2311.03629*.

Nichol, A. Q.; and Dhariwal, P. 2021. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, 8162–8171. PMLR.

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

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ren, Z.; and Lee, Y. J. 2018. Cross-domain self-supervised multi-task feature learning using synthetic imagery. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 762–771.

Rojas-Gomez, R. A.; Singhal, K.; Etemad, A.; Bijamov, A.; Morningstar, W. R.; and Mansfield, P. A. 2023. SASSL: Enhancing Self-Supervised Learning via Neural Style Transfer. *arXiv preprint arXiv:2312.01187*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 234–241. Springer.

Ros, G.; Sellart, L.; Materzynska, J.; Vazquez, D.; and Lopez, A. M. 2016. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3234–3243.

Salimans, T.; and Ho, J. 2022. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*.

Sariyildiz, M. B.; Alahari, K.; Larlus, D.; and Kalantidis, Y. 2023. Fake it till you make it: Learning transferable representations from synthetic ImageNet clones. In *CVPR 2023–IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Tian, Y.; Fan, L.; Isola, P.; Chang, H.; and Krishnan, D. 2024. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. *Advances in Neural Information Processing Systems*, 36.

Van Horn, G.; Mac Aodha, O.; Song, Y.; Cui, Y.; Sun, C.; Shepard, A.; Adam, H.; Perona, P.; and Belongie, S. 2018. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8769–8778.

Zang, Z.; Luo, H.; Wang, K.; Zhang, P.; Wang, F.; Li, S. Z.; and You, Y. 2024. DiffAug: Enhance Unsupervised Contrastive Learning with Domain-Knowledge-Free Diffusion-based Data Augmentation. In *Forty-first International Conference on Machine Learning*.

Zbontar, J.; Jing, L.; Misra, I.; LeCun, Y.; and Deny, S. 2021. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, 12310–12320. PMLR.

Zhang, C.; Barbano, R.; and Jin, B. 2021. Conditional variational autoencoder for learned image reconstruction. *Computation*, 9(11): 114.

Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; and Torralba, A. 2017. Places: A 10 million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhou, J.; Wei, C.; Wang, H.; Shen, W.; Xie, C.; Yuille, A.; and Kong, T. 2021. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*.