Gen-SIS: Generative Self-augmentation Improves Self-supervised Learning

Varun Belagali[1*], Srikar Yellapragada[1*], Alexandros Graikos[1], Saarthak Kapse[1], Zilinghan Li[2], Tarak Nath Nandi[2,3], Ravi K Madduri[2,3], Prateek Prasanna[1], Joel Saltz[1], Dimitris Samaras[1]

[1]Stony Brook University [2]Argonne National Laboratory [3] University of Chicago
vbelagali@cs.stonybrook.edu

Abstract

Self-supervised learning (SSL) methods have emerged as strong visual representation learners by training an image encoder to maximize similarity between features of different views of the same image. To perform this view-invariance task, current SSL algorithms rely on hand-crafted augmentations, such as random cropping and color jittering, to create multiple views of an image. Recently, generative diffusion models were shown to improve SSL by providing a wider range of data augmentations. However, these diffusion models usually require pre-training on large-scale image-text datasets, which might not be available for many specialized domains like histopathology. In this work, we introduce Gen-SIS, a diffusion-based augmentation technique trained exclusively on unlabeled image data, eliminating any reliance on external sources of supervision such as text captions. We first train a vanilla SSL encoder on a dataset using only hand-crafted augmentations. We then train a diffusion model conditioned on embeddings from that SSL encoder. Once trained, this diffusion model can synthesize diverse views of a source image when conditioned on its embedding. Leveraging the ability to interpolate in the encoder latent space, we introduce a novel pretext task: disentangling the two source images from an interpolated synthetic image. We show that these 'self-augmentations', i.e., generative augmentations based on the vanilla SSL encoder embeddings, paired with our disentanglement pretext task, facilitate the training of stronger SSL encoders. We validate Gen-SIS by demonstrating performance gains across various downstream tasks in natural images, which are generally object-centric, and digital histopathology images, which are typically context-based. Furthermore, we show Gen-SIS's effectiveness across multiple SSL methods and encoder variants, highlighting its broad applicability.

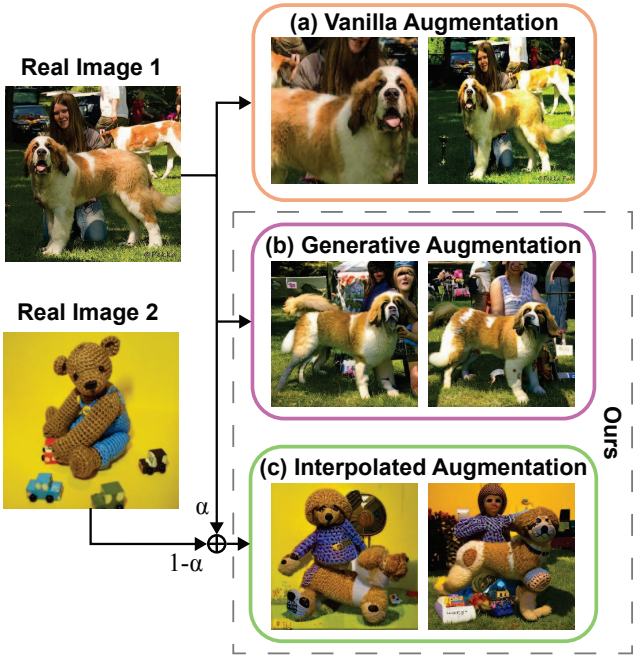

Figure 1. (a) Vanilla augmentations used in SSL, such as random cropping and color jittering. (b) Generative augmentations (ours) conditioned on a single source image. (c) Interpolated augmentations (ours) conditioned on a pair of images. In the Gen-SIS framework, we use (b) for view augmentation and (c) for the disentanglement pretext task, both in conjunction with (a).

1. Introduction

In recent years, self-supervised learning (SSL) [1, 7, 11, 19, 22, 24, 40, 51] has emerged as a standard approach for learning robust visual representations that excel across various downstream tasks. By optimizing the model weights on pretext tasks, like self-prediction or view invariance, SSL enables models to learn discriminative features without requiring labeled data. Specifically, approaches such as DINO [7], BYOL [19], and SimCLR [11] have achieved notable success, producing high-quality features that transfer effectively to diverse downstream applications. This suc-

*Equal contribution. Correspondence to vbelagali@cs.stonybrook.edu
webpage: https://histodiffusion.github.io/docs/publications/gensis

cess stems from view-invariance tasks, which encourage models to learn high-level discriminative features from the image. Formulating view-invariant tasks relies heavily on hand-crafted augmentations, such as cropping and color jittering, to create multiple views of an image. *Stronger augmentations typically lead to more robust features, as they increase the difficulty of the invariance task [19].*

In parallel, diffusion models have achieved impressive results in image generation [41, 48]. This success has led to an interest in using diffusion models, especially large foundation models like Stable Diffusion, for data augmentation [55, 56]. Given SSL's reliance on augmentations, diffusion models can significantly improve SSL by generating images with *non-trivial* variations in background, shape, and position of objects, while preserving the original high-level semantics (Fig. 1 (b))

Recent work by Tian et al. [55] has investigated using synthetic data generated from Stable Diffusion (SD) as multiple views for SSL. However, employing SD as an SSL augmenter has drawbacks: (1) It is challenging to adapt SD in domains underrepresented in its training data: LAION-5B [50]. Since SD is a general image foundation model, it often fails to generate high-quality images from specialized domains such as histopathology, rendering them ineffective for SSL training (see Supplementary Fig. 10). (2) SD-scale foundation models are usually not available for other domains outside natural images, and training them from scratch is a task beyond the scope of improving an SSL encoder. (3) Besides synthesizing variations of an image, it is not straightforward to perform other kinds of augmentations by controlling the conditioning in text-to-image models. For instance, interpolating between two images would require using a language model to first 'interpolate' the two captions and synthesize a new image. (4) As a text-conditioned model, SD is trained on paired image-text data, which can be seen as conflicting with the SSL principle of training on unlabeled data.

To address these limitations, we introduce Gen-SIS, a self-contained framework that trains the diffusion model on the same unlabeled data as an SSL encoder, using the former as an effective data augmenter for SSL without requiring additional supervision (e.g., text or class labels). We adopt the term *self-augmentation* to emphasize the distinction between our fully self-supervised approach and generative augmentations that rely on external supervision [54, 55].

Our approach begins by pre-training an SSL *encoder* on real images from the training dataset, using standard hand-crafted augmentations. Next, we train a latent diffusion model [48] (LDM) conditioned on image embeddings extracted from the initial SSL *encoder*. Once trained, the LDM synthesizes novel images, which are then used to train stronger, enhanced SSL *encoders*. Gen-SIS expands the data augmentation using self-augmentations from the diffusion model, moving beyond traditional hand-crafted augmentations. In a view-invariant setting, a pair of real and synthetic images from our diffusion model can act as different views of the same image, strengthening the augmentation process (Fig. 1 (b)).

Furthermore, we utilize the generative model's ability to interpolate between images and propose a novel pretext task that complements the base SSL objective. Our trained LDM can generate interpolated images by blending the embeddings of the two source images provided as conditioning inputs, resulting in synthetic images that semantically fuse concepts from both sources (Fig. 1 (c)). We then task the visual encoder to identify features from the original pair of images used to generate the interpolated image. This additional pretext task (termed as *disentanglement pretext task*) forces the model to learn and distinguish various object, texture, and shape-level features. Solving this task presents a more significant challenge to the encoder, significantly enhancing its performance on downstream tasks.

We validate the effectiveness of our approach by comparing vanilla SSL encoders against image encoders trained using the proposed self-augmentation methods. We benchmark the DINO [7] encoder on ImageNet classification and multiple natural image downstream tasks. We validate the broad applicability of our method by extending Gen-SIS to multiple SSL algorithms: DinoV2 [40], SimCLR [11], and I-BOT [62]. We also show that Gen-SIS can be applied to the digital histopathology domain, where there are no foundation generative models to provide large-scale generative image augmentations.

In summary, our contributions are:
(1) We introduce Gen-SIS, a generative diffusion-enhanced SSL approach that only requires unlabeled data and is effective across a wide range of SSL algorithms.
(2) We propose a novel disentanglement task as an additional pretext task in self-augmentation enhanced SSL training.
(3) We extensively evaluate our method by pretraining on ImageNet-1K and benchmark the Gen-SIS pretrained encoder across a range of downstream tasks such as classification, retrieval, copy detection, and video segmentation, achieving notable performance gains over vanilla SSL.
(4) We extend Gen-SIS to histopathology images, a domain with no *text-to-image* foundation generative models, demonstrating the effectiveness of our self-contained approach.

We acknowledge that the computational cost associated with training the diffusion model and generating synthetic samples is non-trivial. We discuss this limitation in Sec. 6.

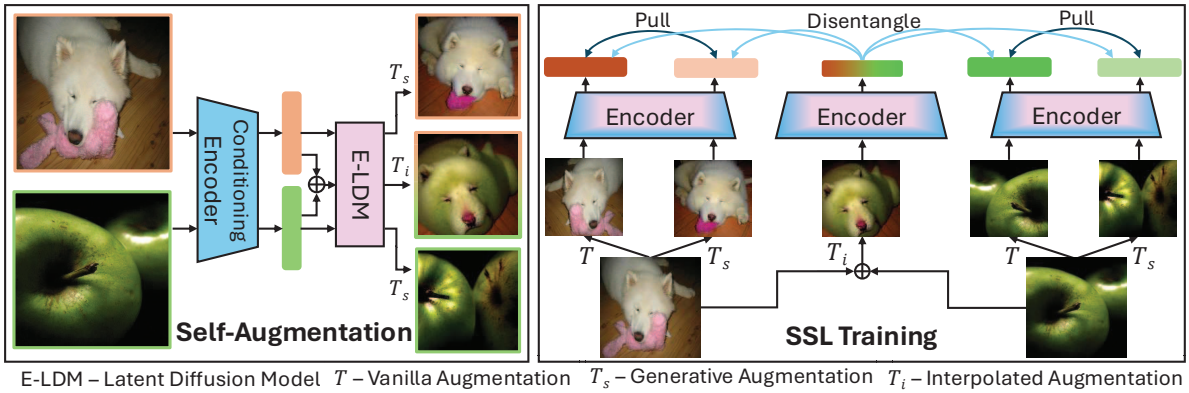

Figure 2. Overview of the Gen-SIS-framework: It contains two key steps: 1) Self-Augmentation using Embedding conditioned LDM (E-LDM), 2) SSL training with augmentations from E-LDM. T - vanilla augmentations, T_s - generative augmentation from single image, and T_i - interpolated augmentation from two images. Note that in conjunction with T_s and T_i, we applied vanilla augmentation. *Pull* represents the vanilla SSL pretext task, and *Disentangle* represents our proposed pretext task with interpolated augmentation.

2. Related work

Self-supervised Learning: Self-supervised learning [16] aims at learning generic representations from large-scale unlabeled data through a pretext task. Pretext tasks can be mainly classified into self-prediction and view-invariance tasks. Self-prediction methods (MAE [23], MaskFeat [59]) involve masking parts of an image and then training the model to reconstruct the missing information based on the remaining context. View-invariant methods task the model to output similar features for two augmented views of the same image. This involves contrastive methods like SimCLR [10], MoCo [21], NNCLR [14] and self-distillation methods like BYOL [20], DINO [7], iBOT [62], and DINOv2 [40]. View-invariant methods typically rely on handcrafted augmentations to derive multiple views of the same image for pretext tasks.

Diffusion Models : Diffusion models were first introduced in the seminal work of Ho et al. [29]. Subsequent advancements included class-conditioning and guidance techniques for more controlled generation [28, 46], and accelerated sampling techniques [52]. Latent diffusion models [8, 41, 49] enable high-resolution image generation by applying the diffusion process in a smaller latent space. In specialized domains, such as histopathology, where labeled image-text data is limited, prior works have adopted image embedding-conditioned diffusion models (E-LDM) [18, 35] to overcome these constraints. RCG [36] explored embedding-conditioned diffusion for natural image datasets. We use the E-LDM architecture from [18] in all our experiments.

Data augmentation with Diffusion models: Diffusion models have been extensively utilized for data augmentation, particularly in supervised settings [3, 17, 18, 56]. For self-supervision, recent works have explored the idea of training small-scale models [61] or using existing large-scale models (SD) to generate synthetic augmentations for SSL training [2]. The works most closely related to our method are Stable-rep [55] and SynCLR [54]. Stable-rep leverages captions from the CC-12M dataset to generate synthetic samples from Stable Diffusion [48] (SD), using them as multiple positive pairs in the SSL training. SynCLR, following a similar approach to Stable-Rep, uses ImageNet object categories to construct text prompts. However, SD-scale text-to-image models are usually unavailable in domains beyond natural images.

Moreover, models trained on large-scale internet datasets, like LAION, may accidentally contain examples from common benchmarks such as ImageNet. Previous works [6, 56] have shown that pretrained diffusion models can leak training data, thus potentially inflating SSL performance.

3. Preliminary

DINO: We use DINO [7] as the representative SSL method on which we develop our approach. In Section 5.5 we show that the ideas developed for DINO can be easily transferred to other SSL algorithms. DINO (self-**di**stillation with **no** labels) is a teacher-student framework in which two augmented views of an image, I' and I'', are processed separately by the student g_{θ_s} and teacher g_{ϕ_t} networks. The two augmented views are generated using standard augmentations, including cropping, color jittering, Gaussian blur, and solarization. Both teacher and student share the same architecture, with a backbone encoder and a projection head, and output a probability distributions P over K dimensions.

The student's output (logits L_s) is sharpened using a

low-temperature τ_s softmax (Eq. 1), while the teacher's output (logits L_t) undergoes centering with a moving average of the teacher outputs c and softmax sharpening with τ_t to prevent collapse during training (Eq. 2). The student network is optimized to match the teacher's probability distribution using a cross-entropy loss H (Eq. 3). The teacher network is updated as exponential moving average (EMA) of the student network's weights.

$$L_s = g_{\theta_s}(I'), \quad P_s^k = \frac{\exp(L_s^k/\tau_s)}{\sum_{j=1}^{K}\exp(L_s^j/\tau_s)}, \tag{1}$$

$$L_t = g_{\phi_t}(I''), \quad P_t^k = \frac{\exp((L_t^k - c^k)/\tau_t)}{\sum_{j=1}^{K}\exp((L_t^j - c^j)/\tau_t)}, \tag{2}$$

$$H(P_t, P_s) = -P_t \log(P_s), \quad \theta_s \leftarrow \text{Optimizer}(H, \theta_s) \tag{3}$$

Latent Diffusion Models: Latent Diffusion Models (LDMs) [48] synthesize images efficiently in a compressed image latent space instead of operating directly on pixels. This latent space is defined by a learned Variational Autoencoder (VAE) that maps images from pixels to latents and back. To control the generated images, the denoiser is usually conditioned on additional information about the images, such as class labels or text prompts. LDMs utilize a cross-attention mechanism between embeddings of the conditioning information and the denoiser features to guide the image synthesis, rendering the conditioning framework flexible to the choice of conditioning signals.

4. Method

In this section, we introduce Gen-SIS (see Fig. 2), a framework that leverages unlabeled data to train a diffusion model and subsequently enhances self-supervised learning (SSL) through novel self-augmentations using this learned diffusion model. First, in Sec. 4.1, we describe the embedding-conditioned Latent Diffusion Model (E-LDM), which generates synthetic images based on the embeddings of source images. Then in Sec. 4.2, we detail how synthetic images (self-augmentations) generated by the E-LDM can be integrated into SSL to improve it. We focus on two types of self-augmentations: (1) Generative augmentations, where augmentations are created from a single source image, and (2) Interpolated augmentations, where an interpolated image is generated from two source images and used in training for a novel disentanglement pretext task.

4.1. Embedding conditioned LDM

We follow the LDM [48] framework for synthetic image generation, conditioning the LDM with the embedding extracted from an image, and refer to this setup as E-LDM (embedding-conditioned LDM). Following the approach of prior work [18], we first train an image encoder on unlabeled real images using a standard SSL algorithm (DINO), and then use this encoder as the conditioning network to guide the diffusion model. This design allows our E-LDM to be trained in a fully self-supervised manner, without relying on any auxiliary information about the images. We call the synthetic images generated from E-LDM *self-augmentations*. As conditioning, we choose the output of the DINO backbone, which is a D-dimensional vector e (embedding). Once trained, we prompt the E-LDM by giving it an embedding of a real image e; it will synthesize a variation $I_s = \text{E-LDM}(z, e)$, where $z \sim \mathcal{N}(0, I)$ is an initial Gaussian noise used in sampling. We use the deterministic DDIM [52] sampling algorithm, which maps every (z, e) pair to an image I_s.

4.2. Enhancing SSL using self-augmentations

With real images as sources for E-LDM conditioning, we use two types of self-augmentations: 1) Generative Augmentations, 2) Interpolated Augmentations.

Generative Augmentations: In the generative augmentation setting, a synthetic image is generated using a single real image as the source. This involves first extracting an embedding e from the source image using the conditioning-encoder, and then guiding the image generation process with that embedding to create a synthetic image $I_s = \text{E-LDM}(z, e)$. As illustrated in Fig. 1 (b), generative augmentations introduce novel variations in the shape, size, and position of objects, as well as changes in the background, while preserving the semantic content of the objects in the image. As shown in Fig. 2, to integrate generative augmentations into SSL, we use the real image and a corresponding synthetic image as an input pair for the SSL pretext task. We also apply hand-crafted augmentations to both real and synthetic images.

Interpolated Augmentations: An interesting property of diffusion models is their ability to generate an image that partially resembles each source image when conditioned on embeddings interpolated from two sources, as demonstrated in prior works [18, 31, 57]. We leverage this property to produce an interpolated synthetic image from two real source images, which we use to perform a new pretext task during the SSL training. With embeddings e_1 and e_2 representing the two source images (I_1, I_2), and an interpolation ratio α, we compute an interpolated embedding e_{int} using spherical linear interpolation (SLERP) [57] $e_{\text{int}} = \text{SLERP}(e_1, e_2, \alpha)$. We choose SLERP over linear interpolation since high-dimensional vectors are concentrated near the surface of the unit sphere. This interpolated embedding serves as the conditioning to generate the synthetic interpolated image, $I_{\text{int}} = \text{E-LDM}(z, e_{\text{int}})$.

Since the interpolated image contains components of both source images, we propose a disentanglement task

where the network learns to separate the distinct features of each source image used in the interpolation. Specifically, given two source images (I_1, I_2), an interpolating ratio (α), and the interpolated synthetic image (I_{int}), we pass I_{int} through the student network, to obtain the student probability P_{int}.

$$L_{\text{int}} = g_{\theta_s}(I_{\text{int}}), \quad P^k_{\text{int}} = \frac{\exp(L^k_{\text{int}}/\tau_s)}{\sum_{j=1}^{K}\exp(L^j_{\text{int}}/\tau_s)} \tag{4}$$

To derive a target teacher output for the disentanglement task, we pass I_1, I_2 to the teacher network individually, and interpolate the teacher head output (logits L_{ent}) using α:

$$L_{\text{ent}} = \alpha g_{\phi_t}(I_1) + (1-\alpha) g_{\phi_t}(I_2). \tag{5}$$

This is then passed through the centering and sharpening operation to get the probability over the K dimensions

$$P^k_{\text{ent}} = \frac{\exp((L^k_{\text{ent}} - c^k)/\tau_t)}{\sum_{j=1}^{K}\exp((L^j_{\text{ent}} - c^k)/\tau_t)} \tag{6}$$

Finally, we compute the disentanglement loss (Eq.7) using the cross-entropy between the student and teacher predictions.

$$\mathcal{L}_{\text{disentangle}} = -P_{\text{ent}}\log(P_{\text{int}}) \tag{7}$$

To optimize for this loss, the student must implicitly disentangle components of the pair of source images within the interpolated image, leading us to call this a *disentanglement pretext task*. This task is more challenging and can yield better representation learning than optimizing solely for single-source augmentations. With single-source images, the student only needs to extract features for a single dominant component to minimize the loss, whereas disentangling multiple components in an interpolated image can help the model learn more discriminative features.

In Gen-SIS, we use both types of self-augmentations, generative augmentation with vanilla dino loss and interpolated augmentation with $\mathcal{L}_{\text{disentangle}}$. We provide the pseudo code in the Supplementary.

5. Experiments

5.1. ImageNet-1K

In this section, we apply the Gen-SIS framework in the natural image domain. Our experiments empirically demonstrate improvements in encoder pre-training using Gen-SIS compared to the vanilla SSL methods on ImageNet-1K [12].

E-LDM and Self Augmentations: The first step involves creating self-augmentations. This begins with training a conditioning encoder: a ViT-S/16 model using the DINO framework trained for 100 epochs on ImageNet-1K. We then train the embedding-conditioned LDM (E-LDM) for 100k iterations using embeddings from this DINO ViT-S/16 encoder as condition, following [18]. We generate self-augmentations (both generative and interpolated) using E-LDM in an offline manner, and load them from the disk during the Gen-SIS SSL training. More implementation details are provided in Supplementary 9.5.

Table 1. Top-1% accuracy on **ImageNet-1K** of DINO and Gen-DINO pre-trained for 100 epochs and evaluated using k-NN (training free) and linear probing (LP) evaluation.

Method	Network	k-NN	LP
DINO	ViT-S	69.4	74.0
Gen-DINO	ViT-S	**70.93** ($\pm$ 0.04)	**74.5**
DINO	ViT-T	59.1	65.0
Gen-DINO	ViT-T	**64.8**	**67.9**

Table 2. Performance of DINO and Gen-DINO on PANDA (6-class classification) and BRIGHT (3-class classification) datasets. For PANDA, we report the mean Top-1% accuracy over 5-fold cross-validation, and for BRIGHT, we report the mean Top-1% accuracy over three seeds.

	Dataset	
Method	**PANDA**	**BRIGHT**
DINO	47.6	64.6
Gen-DINO	**50.8**	**66.3**
UNI [9]	49.4	63.3

Evaluation: We employ standard protocols used in DINO [7], such as the training-free k-nearest neighbor classifier (k-NN) and training a linear classifier (linear-probing) on frozen features. As highlighted in the DINO paper, linear probing is sensitive to the hyperparameters, and hence we consider k-NN to be the preferred choice for evaluation given its robustness.

Comparing with DINO on ImageNet-1K: In Tab. 1, we compare the performance of ViT-S (patch size of 16) pre-trained using our Gen-DINO method against the vanilla DINO method with a 100-epoch schedule on the ImageNet-1K validation set. We observe that, compared to DINO, our method performs significantly better on k-NN evaluation, with an improvement of 1.5% in Top-1% accuracy. We report mean/std of k-NN over 3 pre-training seeds of Gen-DINO. The linear probing evaluation shows an improvement of 0.5%. This evaluation indicates that Gen-DINO enhances representation learning through generative and interpolated augmentations, particularly by learning to solve the more challenging pretext task of disentangling two objects in the object-centric images of ImageNet-1K. We discuss the importance of individual components in Sec. 5.4.

Using the same E-LDM as in the ViT-S training, we also apply Gen-DINO to train a ViT-T model on which we observe larger performance improvement. This demon-

strates that the proposed framework is applicable on different model sizes and can greatly benefit smaller ViT models.

5.2. Histopathology Imaging

Previously, we evaluated Gen-SIS in the natural image domain using ImageNet-1K. In this section, we explore its extension to histopathology, which is non-object-centric and instead involves a complex spatial layout of various tissue structures and nuclei types [9, 33]. Given the lack of large-scale (text-to-image or other) foundation diffusion models in histopathology, self-augmentations using our Gen-SIS framework have significant potential to improve SSL in this domain.

Setup: We validate Gen-SIS on two challenging histopathology datasets: the PANDA prostate cancer dataset [5] (10K Whole Slide Images (WSIs) for 6-class ISUP grading, with Karolinska slides for training and Radboud for evaluation) and the BRIGHT breast cancer dataset [4] (703 WSIs for 3-class classification). After tiling the WSIs into 256×256 pixel crops, we pre-trained ViT-S encoders using both vanilla DINO and Gen-SIS framework (Gen-DINO). We then extract frozen features from these encoders and trained an ABMIL model [32] for slide-level classification. We ensure robust evaluation by reporting mean performance from 5-fold cross-validation for PANDA and averaging results over 3 random seeds for BRIGHT. For more details, see Supplementary 9.5.6.

Results: As observed in Table 2, MIL trained with features extracted from the Gen-DINO pre-trained encoder outperforms those from the DINO pre-trained encoder across both datasets. In the PANDA dataset, our method improves performance by more than 3% in accuracy. In the BRIGHT dataset, we observe an improvement of 1.7% in accuracy. In Table 2, we also report the performance of UNI [9], a SoTA pathology foundation model. Gen-DINO outperforms UNI on both datasets despite being pre-trained only on ∼2M images instead of 100M images used in UNI.

While we focus on improving a vanilla DINO ViT-S encoder when trained on an ImageNet-scale dataset (∼1M), we believe that incorporating Gen-SIS into large-scale (>100M) SSL foundation model training can be beneficial. Since UNI and other foundation models [15, 38, 63] use DINO as the underlying algorithm, Gen-SIS can be easily integrated in their training pipeline.

5.3. ImageNet-1K downstream tasks

Having validated the effectiveness of Gen-SIS by improving k-NN classification accuracy (Sec.5.1) on ImageNet-1K, we probe the ViT-S Gen-DINO encoder on related downstream tasks. We perform copy detection, image retrieval, and video segmentation, closely following the settings described in DINO [7]. For further evaluation details, please refer to Sec. 9.5.5.

As shown in Tables 3, 4 and 5, Gen-DINO features outperform vanilla DINO across all downstream tasks. For video segmentation, we also evaluated Gen-DINO without the disentanglement task, i.e., DINO with generative augmentations only, and found that it performed worse than Gen-DINO. This is additional evidence that the disentanglement pretext task improves the model's understanding of object details.

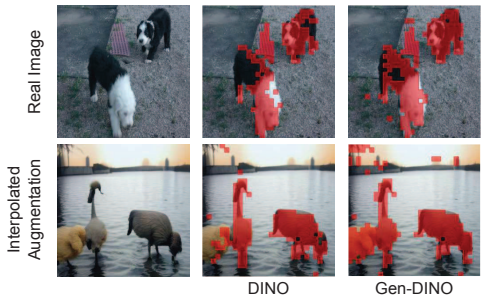

Figure 3. [CLS] token attention map of DINO and Gen-DINO averaged across all heads and overlayed on real and interpolated image. Gen-DINO's attention covers higher portion of object patches than DINO.

In Fig. 3, we visualize the self-attention of the [CLS] token overlayed on a sample real image and on a sample interpolated image using pre-trained ViT-S using DINO and our Gen-DINO model. Consistently, for both real and generated images, Gen-DINO's attention map covers the object patches (16×16 regions) better compared to DINO. This was also reflected in the significantly improved mean region similarity $\mathcal{J}_m$ in Tab 5. We posit that by training on generative and interpolated augmentations of images the SSL encoder learns to focus more on the objects of interest (e.g. animals in the image) as we are mostly introducing variations to the background and location of the object. This leads to the improvement in downstream performance.

5.4. Ablations

Here, we study the effect of the various components of Gen-DINO that are crucial for enhancing the performance of the encoder compared to vanilla DINO. All ablations are conducted using ViT-S pre-trained for 100 epochs on ImageNet-1K and evaluated on its validation set. Top-1% k-NN classifier accuracy is reported.

Importance of Disentanglement pretext task: In Tab. 6, we investigate the effect of only using generative augmentations, without the proposed disentanglement pretext task and interpolated augmentations (No disent.). We compare it to the vanilla DINO (DINO) and proposed Gen-DINO

Table 3. **Copy Detection:** mAP on Copydays [13] "strong" subset.

Method	Dim	mAP
DINO	768	80.2
Gen-DINO	768	**82.5**

Table 4. **Image retrieval.** We compare the mAP on the Oxford (ROx) and Paris (RPar) datasets using frozen features from ViT-S pre-trained with DINO and Gen-DINO on ImageNet-1K.

Method	ROx		RPar	
	M	H	M	H
DINO	30.7	10.8	55.6	26.1
Gen-DINO	**33.3**	**11.2**	**57.2**	**26.9**

Table 5. **DAVIS 2017 Video Object Segmentation.** We compared the performance of frozen features from ViT-S pre-trained (100 epochs) with DINO and Gen-DINO on ImageNet-1K for the task of video instance tracking. Mean region similarity ($\mathcal{J}_m$) and mean contour-based accuracy ($\mathcal{F}_m$) metrics are reported. We use an image resolution of 480p.

Method	$(\mathcal{J}\&\mathcal{F})_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
DINO	61.45	59.67	63.23
Gen-DINO w/o disent.	61.66	59.87	63.45
Gen-DINO	**62.07**	**60.52**	**63.62**

Table 6. Effect of disentanglement pretext task and its position.

Method	k-NN
DINO	69.4
Gen-DINO	
- No disent.	69.9 (+0.5)
- Before proj.	69.6 (+0.2)
- After proj.	**70.9** (+1.5)

Table 7. Effect of interpolation ratio α.

α	k-NN
0.2, 0.4, 0.6, 0.8	70.0
0.4, 0.6	70.1
0.5	**70.9**

encoders (After proj.). We observe that generative augmentations alone provide a 0.5% improvement over vanilla DINO, significantly less than the 1.5% improvement of Gen-DINO. This emphasizes that, beyond simple data augmentation, generative models can enhance the SSL framework in different ways – in our case the interpolation augmentation and disentanglement pretext task.

Effect of teacher entanglement position: In Tab. 6, we experiment with the entanglement position of teacher outputs used in the disentanglement pretext task. By default, we entangle the teacher head logits (after the projection head) of two source images as per Eq.5. We try performing the entanglement after the teacher backbone (before the projection head) and then passing the entangled embedding into the teacher head. Tab. 6 indicates that entangling before the projection head leads to a significant decrease in performance. This can be attributed to the low-dimensional teacher backbone output (384 in ViT-S), which allows less flexibility in feature entanglement within the low-dimensional space compared to the teacher projection head output, which is much higher-dim (typically 65K).

Effect of Interpolation Ratio: In Tab. 7, we explore the effect of interpolation ratio (α) in our framework. By default, in the ImageNet experiments, we use $\alpha = 0.5$ for interpolated image generation. Since any value can be used in the disentanglement pretext task, we experiment with $\alpha = \{0.2, 0.4, 0.6, 0.8\}$ and $\alpha = \{0.4, 0.6\}$.

Using multiple α values may seem optimal due to increased variation, but we found that using $\alpha = 0.5$ yields the best results. To understand this, in Fig. 4, we visualize the generated images with different α values. We observe that for values close to the boundaries (0.2 and 0.8), the effect of the interpolation is barely visible, with the image mostly gravitating toward the dominant side, making the pretext task noisy. Additionally, the images synthesized with values 0.4 and 0.6 are very similar, making it harder for the model to distinguish the exact α used in interpolation.

We hypothesize that the training becomes noisy when the interpolated images do not precisely reflect the interpolation ratio used. This limitation stems from the generative capabilities of the diffusion model, particularly in highly diverse datasets like ImageNet-1K [57], where generating images with the exact interpolation is difficult. Hence, both intuitively and empirically, using $\alpha = 0.5$ proves optimal, as the SSL encoder only needs to understand that the interpolated image is a combination of two others rather than estimating the exact interpolation ratio. In the histopathology domain, where the generative model can accurately capture variations across different α values (Supplementary Fig. 12 and Fig. 13), we opted to use multiple interpolation ratios - $\alpha = \{0.2, 0.4, 0.6, 0.8\}$.

5.5. Extension to other SSL methods and ViT model sizes

To demonstrate the broad applicability of Gen-SIS, we extend it to multiple SSL frameworks — I-BOT [62], DINOv2 [40], and SimCLR [11] — across ViT-Tiny (ViT-T), ViT-Small (ViT-S), and ViT-Base (ViT-B) model sizes on ImageNet-1K. I-BOT extends DINO with an additional masked image modeling pretext task. DINOv2 builds on DINO with additional losses like *ibot* loss and *koleo* loss. SimCLR is a contrastive learning based SSL approach. All models are trained for 100 epochs.

As shown in Tab. 8, Gen-SIS consistently improves $k-$NN accuracy across all settings. Notably, for SimCLR, we observe an improvement of 10.6%. For completeness, we also report the performance of StableRep [55] and SynCLR [54], which train SimCLR-like SSL algorithms on synthetic data generated from Stable Diffusion. As expected, they outperform Gen-SimCLR since they leverage large-scale text-conditioned diffusion models.

For training Gen-I-BOT, Gen-DINOv2, and Gen-SimCLR, we use the same synthetic data generated for Gen-DINO ViT-S in Sec. 5.1. This illustrates that a single diffusion model (E-LDM), trained with embeddings from a vanilla SSL encoder (DINO ViT-S in our case) as condi-

Figure 4. Interpolated augmentations ($\alpha = \{0.2, 0.4, 0.6, 0.8\}$) generated from 2 real images (α=0 and α=1). An example of interpolation between dog and stone image from ImageNet dataset is illustrated.

tioning, can generate augmentations once and be reused to train various SSL methods and ViT model sizes.

Table 8. k-NN acc. on ImageNet-1K using SSL methods DINOv2, I-BOT, and SimCLR and their Gen-SIS versions (Gen-DINOv2, ..)

SSL Method	Network	k-NN
DINOv2	ViT-S	69.4
Gen-DINOv2	ViT-S	70.9
I-BOT	ViT-T	58.6
Gen-I-BOT	ViT-T	64.1
I-BOT	ViT-S	69.8
Gen-I-BOT	ViT-S	70.8
SimCLR	ViT-B	52.3
Gen-SimCLR	ViT-B	62.9
StableRep	ViT-B	69.0
SynCLR	ViT-B	76.1

6. Limitations and Trade-offs

A potential limitation of Gen-SIS is the computational overhead associated with training the E-LDM and generating self-augmentations. In Table 9, we present the total training cost for Gen-SIS, which includes pretraining the vanilla SSL encoder, training the E-LDM, generating synthetic augmentations and training the final SSL encoder. We emphasize that this is a one-time expenditure per pretraining dataset. The ImageNet results presented in Section 5.5 use the same set of synthetic samples for Gen-SIS enhanced training across all SSL frameworks and ViT architectures. This reusability significantly amortizes the initial computational investment. We plan to release the generated synthetic data for all the datasets used in this paper, allowing future research to readily benefit from this investment.

Furthermore, recent works like RCG [36] have trained large E-LDMs on ImageNet. Using such models allows us to bypass the vanilla SSL and diffusion model training stages. We demonstrate this by re-runing Gen-SIS with RCG, reducing our incurred cost to only that of generating augmentations and training the final SSL encoder. This approach reduces our total compute time from 6.5 to 4.75 days (Table 9) while maintaining comparable k-NN accuracy. See Supplementary 9.2 for more details.

Table 9. Training cost in number of days on a single node with 8 A100 GPUs. We highlight in blue the costs that are incurred by Gen-SIS (or StableRep) training.

Method	Vanilla SSL	Diffusion Training	Aug Gen	Gen SSL	k-nn Acc
Gen-SIS	1	2 (E-LDM)	2.5	1	70.9
Gen-SIS	–	32 (RCG)	3.75	1	70.7
StableRep [55]	–	768 (SD 1.5)	34.5	3.4	69.0

Most importantly, our E-LDM training cost is considerably less than the resources involved to train an SD-scale foundation model. While methods like StableRep[55], which rely on SD, do not incur this training cost, their augmentation generation (34.5 vs 2.5 days) is still more expensive. Gen-SIS provides a more computationally efficient way to generate self-augmentations, particularly when large generative models are unavailable (histopathology).

7. Conclusion

We presented Gen-SIS, a self-augmentation technique to enhance self-supervised learning. Self-augmentations are generated from a diffusion model that does not rely on auxiliary information (text or class labels), ensuring a self-contained approach. Our Gen-SIS models, trained with generative and interpolated augmentations alongside the proposed disentanglement pretext task, outperform the vanilla SSL algorithms and show improved performance on downstream tasks. We demonstrated the effectiveness of Gen-SIS across multiple SSL methods and encoder variants, highlighting its broad applicability. Finally, we extended our framework to the non-object-centric histopathology domain, showing consistent improvement across complex cancer grading tasks. While Gen-SIS involves an up-front computational investment for the diffusion model, this is a one-time expenditure per pretraining dataset. We believe that synthetic data augmentations hold immense potential for self-supervised learning, and our framework could benefit large-scale foundation model training in both natural image and histopathology domains.

8. Acknowledgements

This research was partially supported by NCI awards 1R21CA258493-01A1, 5U24CA215109, UH3CA225021, U24CA180924, NSF grants IIS-2123920, IIS-2212046, NIH 1R01CA297843-01, NIH NCI 1R21CA258493-01A1, Stony Brook Profund 2022 seed funding, and generous support from Bob Beals and Betsy Barton. This research used resources of the Argonne Leadership Computing Facility, a U.S. Department of Energy (DOE) Office of Science user facility at Argonne National Laboratory and is based on research supported by the U.S. DOE Office of Science-Advanced Scientific Computing Research Program, under Contract No. DE-AC02-06CH11357. We thank Jingwei Zhang for engaging in discussions on histopathology setup. We thank Prof. Beatrice Knudsen for annotating the descriptions of interpolated histopathology images presented in the supplementary.

References

[1] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629, 2023. 1

[2] Sana Ayromlou, Vahid Reza Khazaie, Fereshteh Forghani, and Arash Afkanpour. Can generative models improve self-supervised representation learning? *arXiv preprint arXiv:2403.05966*, 2024. 3

[3] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J. Fleet. Synthetic data from diffusion models improves imagenet classification. *Transactions on Machine Learning Research*, 2023. 3

[4] Nadia Brancati, Anna Maria Anniciello, Pushpak Pati, Daniel Riccio, Giosuè Scognamiglio, Guillaume Jaume, Giuseppe De Pietro, Maurizio Di Bonito, Antonio Foncubierta, Gerardo Botti, et al. Bracs: A dataset for breast carcinoma subtyping in h&e histology images. *Database*, 2022: baac093, 2022. 6, 7

[5] Wouter Bulten, Kimmo Kartasalo, Po-Hsuan Cameron Chen, Peter Ström, Hans Pinckaers, Kunal Nagpal, Yuannan Cai, David F Steiner, Hester Van Boven, Robert Vink, et al. Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge. *Nature medicine*, 28(1):154–163, 2022. 6, 7

[6] Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 5253–5270, 2023. 3

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 1, 2, 3, 5, 6, 4, 7

[8] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. *arXiv preprint arXiv:2403.04692*, 2024. 3

[9] Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Bowen Chen, Andrew Zhang, Daniel Shao, Andrew H Song, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 2024. 5, 6

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 3

[11] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1, 2, 7, 4

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[13] Matthijs Douze, Hervé Jégou, Harsimrat Sandhawalia, Laurent Amsaleg, and Cordelia Schmid. Evaluation of gist descriptors for web-scale image search. In *Proceedings of the ACM international conference on image and video retrieval*, pages 1–8, 2009. 7, 6

[14] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9588–9597, 2021. 3

[15] Alexandre Filiot, Paul Jacob, Alice Mac Kain, and Charlie Saillard. Phikon-v2, a large and public feature extractor for biomarker prediction. *arXiv preprint arXiv:2409.09173*, 2024. 6

[16] Jonas Geiping, Quentin Garrido, Pierre Fernandez, Amir Bar, Hamed Pirsiavash, Yann LeCun, and Micah Goldblum. A cookbook of self-supervised learning. *arXiv preprint arXiv:2304.12210*, 2023. 3

[17] Alexandros Graikos, Srikar Yellapragada, and Dimitris Samaras. Conditional generation from unconditional diffusion models using denoiser representations. *arXiv preprint arXiv:2306.01900*, 2023. 3

[18] Alexandros Graikos, Srikar Yellapragada, Minh-Quan Le, Saarthak Kapse, Prateek Prasanna, Joel Saltz, and Dimitris Samaras. Learned representation-guided diffusion models for large-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8532–8542, 2024. 3, 4, 5

[19] 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. *Advances in neural information processing systems*, 33:21271–21284, 2020. 1, 2

[20] 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. *Advances in neural information processing systems*, 33:21271–21284, 2020. 3

[21] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 3

[22] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 1

[23] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 3

[24] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 1

[25] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021. 1

[26] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *CVPR*, 2021. 1

[27] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 3

[28] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 3

[29] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[30] Wentao Huang, Xiaoling Hu, Shahira Abousamra, Prateek Prasanna, and Chao Chen. Hard negative sample mining for whole slide image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 144–154. Springer, 2024. 7

[31] Drew A Hudson, Daniel Zoran, Mateusz Malinowski, Andrew K Lampinen, Andrew Jaegle, James L McClelland, Loic Matthey, Felix Hill, and Alexander Lerchner. Soda: Bottleneck diffusion models for representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23115–23127, 2024. 4

[32] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018. 6, 7

[33] Saarthak Kapse, Srijan Das, Jingwei Zhang, Rajarsi R Gupta, Joel Saltz, Dimitris Samaras, and Prateek Prasanna. Attention de-sparsification matters: Inducing diversity in digital pathology representation learning. *Medical Image Analysis*, 93:103070, 2024. 6

[34] Saarthak Kapse, Pushpak Pati, Srijan Das, Jingwei Zhang, Chao Chen, Maria Vakalopoulou, Joel Saltz, Dimitris Samaras, Rajarsi R Gupta, and Prateek Prasanna. Si-mil: Taming deep mil for self-interpretability in gigapixel histopathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11226–11237, 2024. 7

[35] Minh-Quan Le, Alexandros Graikos, Srikar Yellapragada, Rajarsi Gupta, Joel Saltz, and Dimitris Samaras. ∞-brush: Controllable large image synthesis with diffusion models in infinite dimensions. *arXiv preprint arXiv:2407.14709*, 2024. 3

[36] Tianhong Li, Dina Katabi, and Kaiming He. Return of unconditional generation: A self-supervised representation generation method. *Advances in Neural Information Processing Systems*, 37:125441–125468, 2025. 3, 8, 1

[37] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021. 7

[38] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874, 2024. 6

[39] L. McInnes, J. Healy, and J. Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv e-prints*, 2018. 2

[40] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1, 2, 3, 7, 4

[41] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 2, 3

[42] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2008. 6

[43] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 7

[44] Filip Radenović, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondřej Chum. Revisiting oxford and paris: Large-scale image retrieval benchmarking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5706–5715, 2018. 6

[45] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning cnn image retrieval with no human annotation. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1655–1668, 2018. 6

[46] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 3

[47] Sucheng Ren, Huiyu Wang, Zhengqi Gao, Shengfeng He, Alan Yuille, Yuyin Zhou, and Cihang Xie. A simple data mixing prior for improving self-supervised learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14595–14604, 2022. 2

[48] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3, 4

[49] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3

[50] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 2

[51] Mannat Singh, Quentin Duval, Kalyan Vasudev Alwala, Haoqi Fan, Vaibhav Aggarwal, Aaron Adcock, Armand Joulin, Piotr Dollár, Christoph Feichtenhofer, Ross Girshick, et al. The effectiveness of mae pre-pretraining for billion-scale pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5484–5494, 2023. 1

[52] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3, 4

[53] Kevin Thandiackal, Boqi Chen, Pushpak Pati, Guillaume Jaume, Drew FK Williamson, Maria Gabrani, and Orcun Goksel. Differentiable zooming for multiple instance learning on whole-slide images. In *European Conference on Computer Vision*, pages 699–715. Springer, 2022. 7

[54] Yonglong Tian, Lijie Fan, Kaifeng Chen, Dina Katabi, Dilip Krishnan, and Phillip Isola. Learning vision from models rivals learning vision from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15887–15898, 2024. 2, 3, 7

[55] Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3, 7, 8, 6

[56] Brandon Trabucco, Kyle Doherty, Max A Gurinas, and Ruslan Salakhutdinov. Effective data augmentation with diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 3

[57] Clinton J Wang and Polina Golland. Interpolating between images with diffusion models. *arXiv preprint arXiv:2307.12560*, 2023. 4, 7

[58] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pages 10506–10518, 2019. 1

[59] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14668–14678, 2022. 3

[60] Srikar Yellapragada, Alexandros Graikos, Prateek Prasanna, Tahsin Kurc, Joel Saltz, and Dimitris Samaras. Pathldm: Text conditioned latent diffusion model for histopathology. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5182–5191, 2024. 3

[61] Dewen Zeng, Yawen Wu, Xinrong Hu, Xiaowei Xu, and Yiyu Shi. Contrastive learning with synthetic positives. In *European Conference on Computer Vision*, pages 430–447. Springer, 2024. 3

[62] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021. 2, 3, 7, 4

[63] Eric Zimmermann, Eugene Vorontsov, Julian Viret, Adam Casson, Michal Zelechowski, George Shaikovski, Neil Tenenholtz, James Hall, Thomas Fuchs, Nicolo Fusi, et al. Virchow 2: Scaling self-supervised mixed magnification models in pathology. *arXiv preprint arXiv:2408.00738*, 2024. 6