# 論文 / 著書情報
# Article / Book Information

| | |
|---|---|
| Title | Diffusion-based Generative Regularization for Supervised Discriminative Learning |
| Authors | Takuya Asakura, Nakamasa Inoue, Koichi Shinoda |
| Citation | Proceedings of the Winter Conference on Applications of Computer Vision (WACV), , , pp. 8915-8926 |
| Pub. date | 2025, 3 |
| Copyright | (c) 2025 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. |
| DOI | http://dx.doi.org/10.1109/WACV61041.2025.00864 |
| Note | This file is author (final) version. |

# Diffusion-based Generative Regularization for Supervised Discriminative Learning

Takuya Asakura, Nakamasa Inoue, Koichi Shinoda
Institute of Science Tokyo
asakura@ks.comp.isct.ac.jp, inoue@comp.isct.ac.jp, shinoda@comp.isct.ac.jp

## Abstract

*Ensuring the quality and quantity of labeled training data has long been a challenge in training deep neural networks for discriminative tasks. One solution to this problem is to use a generative model to augment training data and learn a discriminative model with it. For image classification, with the recent development of diffusion models, it has become possible to generate a variety synthetic images, and there are high expectations for their use as training data. However, to obtain high-quality labeled synthetic images, the hyperparameters and prompts often need to be manually tuned, and the accuracy of the trained image classification model is highly dependent on them. To address this issue, this paper proposes diffusion-based generative regularization, a supervised discriminative learning framework that utilizes a diffusion-based image generation model as a regularizer to robustly learn discriminative representations without the need to synthesize images. Our experiments using vision transformers and stable diffusion models on ImageNet-1k demonstrate that the proposed framework improves classification accuracy on both in-distribution and distribution-shifted data.*

## 1. Introduction

The amount and variety of labeled training data are among the most important factors in improving the performance of deep neural networks for discriminative tasks. It is known that as the size of the training dataset increases, the model's discrimination performance improves [17, 94]. For image classification, training images are typically paired with ground-truth labels indicating the correct category for each image. However, the cost of preparing a large number of these pairs has long been a major challenge in this field.

A recent trend to address this challenge is the use of labeled synthetic images generated by pre-trained generative models, such as diffusion-based text-to-image generation models [18, 65, 68]. Some previous studies have reported that synthetic images effectively improve image classification performance and robustness to distribution shifts [2, 7, 27, 73, 96]. However, there is still a large dependency on manual prompt engineering or hyperparameter tuning, such as adjusting the CFG scale [22, 73], making it challenging to apply generative models to supervised discriminative learning. Moreover, the generation of synthetic images may be unsuccessful if there is a discrepancy in the data distribution between the target task and the training of the generative model.

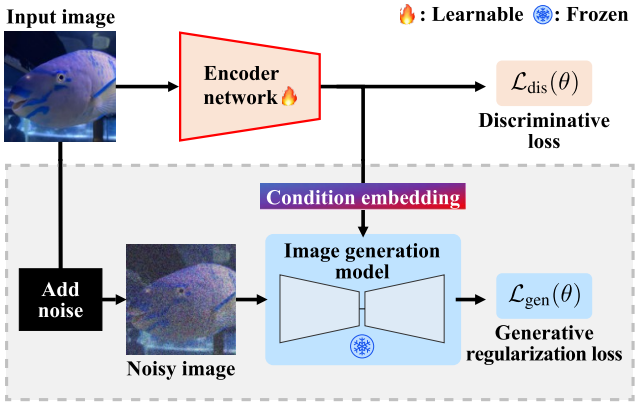



Figure 1. **Diffusion-based generative regularization.** The proposed framework uses a frozen diffusion-based image generation model as a regularizer to enhance training effectiveness and robustness in supervised learning scenarios.

Motivated by these studies, this paper investigates whether generative models can improve supervised discriminative learning without relying on synthetic images. To this end, we propose diffusion-based generative regularization, which utilizes an image generation model as a regularizer. As shown in Figure 1, when training an encoder network with a discriminative loss, such as the cross-entropy loss for image classification, the generative regularization loss is computed simultaneously by feeding the encoder output to a frozen image generation model as a condition embedding. This regularization helps in learning more robust discriminative representations for image classification when the im-

age generation model is trained to precisely reflect conditions. For example, when the image generation model is pre-trained using text embeddings for conditioning, the encoder network can learn representations that are regularized to align with these text embeddings. In this sense, our approach is related to logit distillation [30, 46, 69], a knowledge distillation technique applied to the encoder outputs. However, it differs in that there is no loss function bridging between network outputs, making our approach and investigation novel.

In experiments, we demonstrate the effectiveness of generative regularization using the vision transformer (ViT) [86] and the stable diffusion model in supervised learning scenarios on the ImageNet-1k dataset [70], and show three key results. First, we show that image classification accuracy is improved on both in-distribution data (ImageNet-1k validation set and ImageNet-V2 [66]) and distribution-shifted data (ImageNet-Sketch [90] and ImageNet-R [29]). Second, we show that the effect of generative regularization persists in downstream tasks by fine-tuning the model on ten image classification datasets. Finally, we show the superiority of our approach over other representative approaches including knowledge distillation from OpenCLIP-ViT/H [13], consistency regularization [76], image augmentation using noisy images, and training with real and synthetic SD-ImageNet images [27]. Our contributions are summarized as follows.

- We propose diffusion-based generative regularization, a supervised discriminative learning framework that utilizes a frozen generative model as a regularizer. Specifically, we propose two regularization loss functions: noise consistency loss and latent cross-entropy loss. The former minimizes the distance between sampled and predicted noise, as done in previous generative learning. The latter improves the robustness of discriminative representations through generative latent representations.

- We conduct experiments on ImageNet-1k, ImageNet-V2, ImageNet-Sketch, ImageNet-R and ten downstream image classification datasets to evaluate the effectiveness of our framework. We also demonstrate the superiority of our approach over existing methods that apply loss to the outputs of the encoder network.

## 2. Related Works

### 2.1. Diffusion Model

Diffusion models [31, 75], a type of generative model, can produce very high quality and diverse samples under conditioning such as class label and text prompts [32]. In recent years, they have achieved remarkable results in various fields such as image [4,18,57,64,71], 3D [47,52,62,82], audio [34, 41, 48, 67] and molecular conformation generation [33, 92]. The generation algorithm is inspired by the stochastic diffusion process in thermodynamics. The diffusion model adds Gaussian noise to training data according to the time step $t = 1, \dots, T$ in the diffusion process and learns the target data distribution by gradually denoising it in the reverse process. To facilitate the generation process, a number of sampling algorithms have been proposed [6, 19, 50, 51, 77, 97, 99]. It is also known that the diffusion model can be considered as a score-based model [37, 78, 79, 87].

Specifically, using a form of noise-prediction [38], the denoising learning of the noised sample $x_t$ at time step $t$ minimizes the objective function

$$\mathcal{L}_{\mathrm{D}} = \mathbb{E}_{\boldsymbol{x}_t, \epsilon \sim \mathcal{N}(0,1), t}[\|\epsilon - \epsilon_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t, c)\|_2^2], \tag{1}$$

where $\epsilon \sim \mathcal{N}(0, 1)$ is the Gaussian noise, $\epsilon_{\boldsymbol{\theta}}$ is the output of the diffusion model $\boldsymbol{\theta}$, and $c$ is a condition for sample $x$. In large-scale text-to-image diffusion models, diffusion and reverse processes are typically performed in the latent space for efficiency [9, 10, 20, 61, 68], *i.e.*, minimize Eq. (1) using the latent variable $\boldsymbol{z}$ of the training sample $\boldsymbol{x}$ obtained from an encoder such as variational autoencoder (VAE) [39].

### 2.2. Diffusion Model for Discriminative Tasks

It is known that a well-trained diffusion model, such as Stable Diffusion, can be used directly for zero-shot image classification by applying Bayes' theorem [15, 36, 45]. In the classification, each class should have a corresponding prompt, *e.g.*, "a photo of a {classname}.". These prompts are used to calculate the denoising loss in Eq. (1), and the class corresponding to the prompt with the lowest loss is selected as the estimated class. This means that a well-trained diffusion model is expected to have better denoising performance if the input image is properly conditioned. Although the diffusion model is only trained to generate images, it has comparable zero-shot image classification abilities as a discriminative model such as CLIP [63]. The diffusion model also outperforms CLIP on Winoground [83], a benchmark that measures visual-linguistic compositional reasoning ability, indicating its superior ability to capture the relationships between objects in an image. It is also shown that the diffusion model is robust to adversarial perturbations when applied to image classification tasks [12].

From an alternative perspective, numerous works have explored whether the high performance of the diffusion model in vision tasks can be used to improve discriminative models. The most dominant method for this purpose is training models using synthetic images [23, 24, 28, 55, 84, 85, 93]. In the field of image classification, training models with synthetic images has been demonstrated to improve robustness to distribution-shifted data such as ImageNet-R [5, 100]. It is known that training with synthetic images generated by off-the-shelf diffusion models

does not significantly enhance the accuracy of the target dataset [5, 27]. Therefore, fine-tuned Imagen [71] was employed to generate synthetic images in [46] for the purpose of further enhancement. However, further training of the diffusion model is a computationally expensive process. In addition, the optimal prompt and CFG scale should be selected for the generation of the synthetic image. Despite this, several studies [22, 73] have demonstrated that the performance of image classification models trained on synthetic images highly depends on the prompt and CFG scale used during generation. As a result, training image classifiers on synthetic images requires a lot of trial and error. Li et al. [45] improved performance for discriminative tasks by pre-training CNNs with knowledge distillation from diffusion models without generating synthetic images. However, its algorithm is highly dependent on CNN-based architectures and is therefore unsuitable for training ViTs. Moreover, recent studies [21, 26, 49, 101] have employed Transformer-based architectures, such as DiT [60], for the diffusion model itself in order to achieve greater scalability. Our method aims to improve the performance of image classifiers by exploiting the insight of the diffusion model into the vision task without additional training of the diffusion model, prompt engineering, selecting CFG scale, or dependence on the architecture.

### 2.3. Regularization

Regularization is commonly used to improve the performance of discriminative models. Data augmentation is one of the most common regularization methods in deep learning, and recent models are trained using multiple data augmentations on training data. A more advanced regularization method that utilizes training data is consistency regularization [3, 44, 72]. In this method, the model is given a clean training image and an image with some perturbation, such as adding noise. The discriminative model training is regularized to produce consistent representations for both images. Consistency regularization is often used in semi-supervised learning because it does not require supervised labels [1, 8, 35, 76, 80]. Tarvainen et al. [81] used exponential moving averages of the model parameters to compute the consistency loss. The perturbations to the training image can be the same as common data augmentations such as geometric or color transformations and Mixup [89, 91]. Consistency regularization also helps to stabilize and enhance the generative model [53, 74, 95]. Zhao et al. [98] proposed an improved consistency regularization for adversarial generative networks (GANs) [25]. In the context of self-supervised learning, Ko et al. [40] improved the local semantics of the generated images by learning the consistency between patches of randomly cropped real images with a discriminator of GANs.

## 3. Proposed Method

This section introduces diffusion-based generative regularization, a simple yet effective learning framework that utilizes a diffusion-based image generation model as a regularizer to improve supervised discriminative learning. As shown in Figure 2, the framework incorporates a frozen image generation model on top of a learnable encoder network to regularize its discriminative representations. Note that this work focuses on investigating whether generative models can improve supervised learning, without relying on synthetic images or prompt tuning. This is challenging because methodologies effective for supervised and self-supervised learning are often different, but we show that it is possible to design an approach that leverages the strengths of generative models in supervised learning scenarios.

### 3.1. Diffusion-based Generative Regularization

**Notation and settings.** Let $\mathcal{D}_{\text{train}} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^N$ be a labeled training dataset that consists of training images $\boldsymbol{x}_i \in \mathcal{X}$ and corresponding ground-truth labels $\boldsymbol{y}_i \in \mathcal{Y}$ for $i = 1, 2, \cdots, N$, where $\mathcal{X}$ is a set of images, $\mathcal{Y}$ is a discrete set of one-hot vectors indicating labels, and $N$ is the number of images. This paper considers supervised discriminative learning, where the goal is to train a discriminative function $F : \mathcal{X} \to \mathcal{Y}$ that can accurately predict the labels $\boldsymbol{y} \in \mathcal{Y}$ for unseen images $\boldsymbol{x} \in \mathcal{X}$. To explore the usefulness of generative models in this scenario, we assume that a pre-trained generative model $G : \mathcal{X} \times \mathcal{T} \times \mathcal{C} \to \mathcal{X}$ is given, which describes an image generation process based on a conditional reverse diffusion process. Specifically, we assume that the image sampling process is given by

$$\boldsymbol{x}^{(t-1)} = G(\boldsymbol{x}^{(t)}, t, \boldsymbol{c}) + \sigma_t \boldsymbol{u}, \tag{2}$$

where $t \in \mathcal{T}$ is a timestep, $\boldsymbol{c} \in \mathcal{C}$ is a condition embedding, $\sigma_t$ is a scheduled variance, and $\boldsymbol{u}$ is a Gaussian noise. Here, $\boldsymbol{x}^{(0)}$ indicates a clean image, $\boldsymbol{x}^{(t)}$ indicates a noisy image at timestep $t > 0$, $\mathcal{T}$ is a discrete time space, *i.e.*, $\mathcal{T} = \{0, 1, 2, \cdots, T\}$[1] with $T \in \mathbb{N}_{>0}$. For example, text-to-image models such as the stable diffusion model [68] that feed a condition embedding $\boldsymbol{c} \in \mathcal{C} = \mathbb{R}^{d_1 \times d_2}$ to the cross-attention modules of the conditional U-Net can be used as $G$.

**Framework.** Figure 1 shows an overview of the proposed framework, which utilizes two loss functions: a discriminative loss $\mathcal{L}_{\text{dis}}$ and a generative regularization loss $\mathcal{L}_{\text{gen}}$.

The discriminative loss is defined over a labeled training dataset as follows:

$$\mathcal{L}_{\text{dis}}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \ell_{\text{dis}}(\hat{\boldsymbol{y}}_i, \boldsymbol{y}_i), \tag{3}$$

$$\hat{\boldsymbol{y}}_i = h_{\theta_2}(f_{\theta_1}(\boldsymbol{x}_i)), \tag{4}$$

[1] Extension to continuous time step $\mathcal{T} = [0, 1]$ is straightforward.

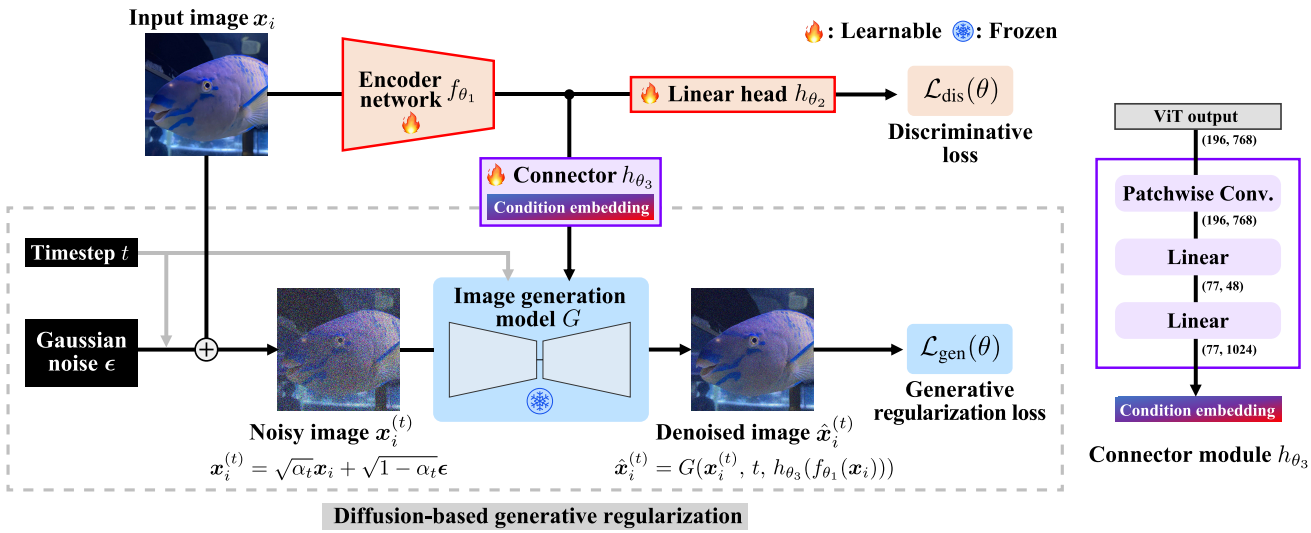



Figure 2. Overview of the proposed framework. A diffusion-based image generation model is used as a regularizer to efficiently train discriminative models. The encoder output is fed into a frozen image generation model $G$ as conditioning features through a connector module. By minimizing the generative regularization loss $\mathcal{L}_{\text{gen}}$, the encoder acquires more robust representations for image classification. The architecture of the connector module is designed for the vision transformer encoder and the stable diffusion image generation model.

where $\hat{\boldsymbol{y}}_i$ is a predicted label, $f_{\theta_1} : \mathcal{X} \to \mathbb{R}^d$ is an encoder network that maps images into $d$-dimensional vectors, $h_{\theta_2} : \mathbb{R}^d \to \mathbb{R}^{|\mathcal{Y}|}$ is a linear classification head, $\theta_1, \theta_2$ are sets of learnable parameters, and $\ell_{\text{dis}}$ is a discriminative loss such as cross entropy loss for discrete classification problems. Note that $F = h_{\theta_2} \circ f_{\theta_1}$ is the discriminative function to be trained.

The generative regularization loss is also defined over the labeled training dataset as follows:

$$\mathcal{L}_{\text{gen}}(\theta) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\ell_{\text{reg}}(\hat{\boldsymbol{x}}_i^{(t)}, \boldsymbol{x}_i, \boldsymbol{\epsilon}_t, \boldsymbol{y}_i), \tag{5}$$

$$\hat{\boldsymbol{x}}_i^{(t)} = G(\boldsymbol{x}_i^{(t)}, t, h_{\theta_3}(f_{\theta_1}(\boldsymbol{x}_i))), \tag{6}$$

$$\boldsymbol{x}_i^{(t)} = \sqrt{\alpha_t}\boldsymbol{x}_i + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}, \tag{7}$$

where $\hat{\boldsymbol{x}}_i^{(t)}$ is a denoised image, $\boldsymbol{x}_i^{(t)}$ is a noisy image at timestep $t$, $\alpha_t$ is a scheduled hyperparameter for mixing clean images with noise, $\boldsymbol{\epsilon}$ is a Gaussian noise, and $\ell_{\text{reg}}$ is a loss for regularization.

In Eq. (6), the discriminative representation $f_{\theta_1}(\boldsymbol{x}_i) \in \mathbb{R}^d$ is fed into the generative model as a condition embedding through a learnable small module $h_{\theta_3}$, which we refer to as a connector module. This regularizes the representations to align with the condition embeddings of the image generation model, potentially resulting in more robust representations for classifying objects in images if the image generation model is trained to precisely reflect conditions. For example, if the original condition embeddings are given by text embeddings, the ideal condition embedding $\boldsymbol{c}^* \in \mathcal{C}$ corresponding to an image $\boldsymbol{x}$ would be the embedding of the sentence that most appropriately describes the image's contents. Therefore, the generative regularization loss potentially predisposes the discriminative model to learn representations that satisfy $h_{\theta_3}(f_{\theta_1}(\boldsymbol{x}_i)) \simeq \boldsymbol{c}^*$, effectively embedding image's contents for classification, even if the training labels are provided for only a single object per image. We discuss the definition of $\ell_{\text{reg}}$ in the following subsections.

### 3.2. Regularization Loss

We propose two types of loss definition for $\ell_{\text{reg}}$ in Eq. (5): a noise consistency (NC) loss, and a latent cross entropy (LC) loss. The former is a generative loss inspired by previous studies on diffusion models. The latter is a cross entropy loss applied to the latent generative representations, aiming to improve the robustness of discriminative representations.

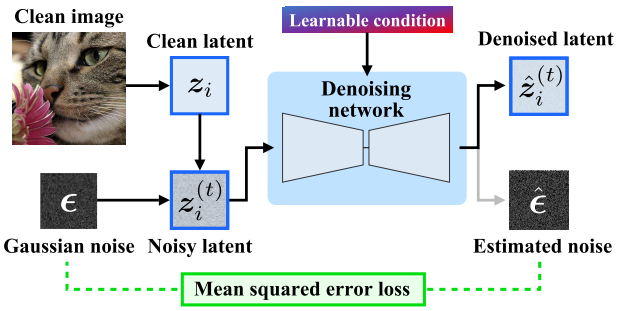
#### 3.2.1 Noise Consistency Loss

The NC loss minimizes the distance between sampled noise $\boldsymbol{\epsilon}$ and predicted noise $\hat{\boldsymbol{\epsilon}}$ following previous generative learning approaches. Specifically, we define the loss function $\ell_{\text{reg}}$ as follows:

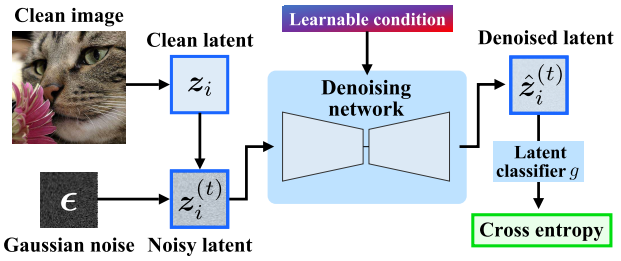
$$\ell_{\text{reg}}^{\text{NC}}(\hat{\boldsymbol{x}}_i^{(t)}, \boldsymbol{x}_i, \boldsymbol{\epsilon}, \boldsymbol{y}_i) = d(\boldsymbol{\epsilon}, \hat{\boldsymbol{\epsilon}}), \tag{8}$$

$$\hat{\boldsymbol{\epsilon}} = \frac{1}{\sqrt{1-\alpha_t}}(\boldsymbol{x}_i - \sqrt{\alpha_t}\hat{\boldsymbol{x}}_i^{(t)}), \tag{9}$$

where $d$ is a distance-based function. We choose squared $L_2$ distance for $d$, *i.e.*, $d(\boldsymbol{a}, \boldsymbol{b}) = \|\boldsymbol{a} - \boldsymbol{b}\|_2^2$. In practice, if the image generation model uses a network (*e.g.*, U-Net) to directly predict noise, $\hat{\boldsymbol{\epsilon}}$ can be obtained from the output of the network.




(a) NC loss




(b) LC loss

Figure 3. Flow of loss calculation for latent diffusion models. (a) Noise consistency (NC) loss is a mean squared error loss between a sampled noise $\boldsymbol{\epsilon}$ and an estimated noise $\hat{\boldsymbol{\epsilon}}$. (b) Latent cross entropy (LC) loss is a cross entropy loss calculated through a latent classifier $g$.

For latent diffusion models, we measure the distance in the latent space. More specifically, the flow of loss calculation is shown in Figure 3a. First, given a clean image $\boldsymbol{x}_i$, the clean latent representation $\boldsymbol{z}_i$ is computed via the encoder of the image generation model. Second, a Gaussian noise $\boldsymbol{\epsilon}$ is applied $\boldsymbol{z}_i$ to obtain noisy latent representation $\boldsymbol{z}_i^{(t)} = \sqrt{\alpha_t}\boldsymbol{z}_i + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}$ with a uniformly sampled timestep $t$. Third, the denoised latent representation $\hat{\boldsymbol{z}}_i^{(t)}$ and estimated noise $\hat{\boldsymbol{\epsilon}}$ are obtained via the image generation model. Finally, the NC loss is computed.

#### 3.2.2 Latent Cross Entropy Loss

Since our goal is to improve discriminative representations, minimizing the NC loss between noises that improves denoising performance does not directly address the main objective. From this perspective, there is still room for improvement in the loss definition for $\ell_{\text{reg}}$, particularly by utilizing the training label information.

In image classification scenarios, learning robust representations is crucial, but training labels are often limited to a single object per image. For example, when training the "cat" category with the image in Figure 4, other category labels such as "whiskers", "sniffing" and "folower" are often missing, even though some of them are relevant to the cat category and could help in learning robust discriminative representations. To address this problem, we propose to learn discriminative representations while simultaneously learning a condition embedding minimizes cross entropy loss over latent generative representations. Specifically, we introduce the LC loss:

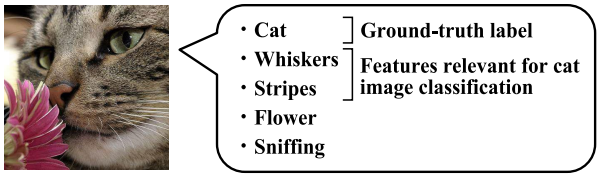



Figure 4. Example of cat image for training.

$$\ell_{\text{reg}}^{\text{LC}}(\hat{\boldsymbol{z}}_i^{(t)}, \boldsymbol{z}_i, \boldsymbol{\epsilon}_t, \boldsymbol{y}_i) = \ell_{\text{dis}}(\tilde{\boldsymbol{y}}_i, \boldsymbol{y}_i), \quad \tilde{\boldsymbol{y}}_i = g(\hat{\boldsymbol{z}}_i^{(t)}) \tag{10}$$

where $g$ is a latent classifier that predicts the label from the denoised latent representation $\hat{\boldsymbol{z}}_i^{(t)}$, and $\ell_{\text{dis}}$ is a discriminative loss used in Eq. (3). We expect that this loss contributes to improved discrimination performance on distribution-shifted images, such as sketch images, because classifying latent representations helps to find more semantically meaningful decision boundaries, such as whether a cat has whiskers or not.

The flow of loss calculation is shown in Figure 3b. Here, the latent classifier $g$ is assumed to be pre-trained on a training dataset and is frozen. This allows a fair comparison of the NC and LC losses in the sense that the number of learnable parameters when training the discriminative function $F$ is the same. We investigate two architectures for the latent classifier, ViT-based and VAE-decoder-based architectures, detailed in Section 4.1.

## 4. Experiments

In this section, we demonstrate the effectiveness of diffusion-based generative regularization in supervised image classification scenarios.

### 4.1. Experimental Settings

**Discriminative model.** We chose the vision transformer base (ViT-B) as the encoder network $f_{\theta_1}$. It consists of twelve transformer encoder blocks, producing $d = 768$ dimensional outputs. The classification head $h_{\theta_2}$ in Eq. (4) is a learnable linear layer. The connector module $h_{\theta_3}$ in Eq. (6) is a small feed-forward network that consists of a patchwise convolution layer followed by two linear layers, as detailed in Figure 2. It outputs an embedding with a size of $77 \times 1024$.

**Generative model.** The Stable Diffusion v2.1 is used as the image generation model $G$. It is a latent diffusion model consisting of a pair of VAE-based encoder and decoder, and a conditional U-Net that accepts text embeddings as conditions, where the condition embedding space is given by

| Loss | Latent classifier | ImageNet-1k | | ImageNet-V2 | | ImageNet-Sketch | | ImageNet-R | |
|---|---|---|---|---|---|---|---|---|---|
| | | Top-1 (%) | Top-5 (%) | Top-1 (%) | Top-5 (%) | Top-1 (%) | Top-5 (%) | Top-1 (%) | Top-5 (%) |
| $\mathcal{L}_{\text{dis}}$ (baseline) | - | 81.8 | 95.6 | 70.4 | 89.0 | 32.0 | 50.3 | 30.4 | 44.5 |
| $\mathcal{L}_{\text{dis}} + \mathcal{L}_{\text{gen}}^{\text{NC}}$ | - | 82.0 (+0.2) | 95.6 (+0.0) | 71.0 (+0.6) | 89.2 (+0.2) | 31.0 (-1.0) | 49.0 (-1.3) | 30.3 (-0.1) | 44.6 (+0.1) |
| $\mathcal{L}_{\text{dis}} + \mathcal{L}_{\text{gen}}^{\text{LC}}$ | VAE dec. | **82.2** (+0.4) | **95.8** (+0.2) | **71.3** (+0.9) | **89.8** (+0.8) | 32.7 (+0.7) | 51.8 (+1.5) | 31.2 (+0.8) | 45.5 (+1.0) |
| $\mathcal{L}_{\text{dis}} + \mathcal{L}_{\text{gen}}^{\text{LC}}$ | ViT | **82.2** (+0.4) | **95.8** (+0.2) | 71.0 (+0.6) | 89.5 (+0.5) | **33.2** (+1.2) | **52.3** (+2.0) | **31.8** (+1.4) | **46.3** (+1.8) |

Table 1. Comparison of top-1/top-5 accuracies on ImageNet and distribution-shifted datasets. $\mathcal{L}_{\text{gen}}^{\text{NC}}$ and $\mathcal{L}_{\text{gen}}^{\text{LC}}$ use the NC and LC loss for generative regularization, respectively. The method using the LC loss shows performance improvement on all datasets.

| Latent classifier | #Params. | Top1 (%) | Top 5 (%) |
|---|---|---|---|
| VAE decoder | 25.2M | 69.3 | 88.3 |
| ViT | 5.6M | 66.0 | 87.0 |

Table 2. Accuracies of pre-trained latent classifiers on ImageNet-1k. Although the performance is lower than that of the baseline in Table 1, these classifiers helped improve performance.

$\mathcal{C} = \mathbb{R}^{77\times1024}$. The parameters of this model is frozen. The NC and LC losses are computed in the latent space.

**Latent classifier.** The latent classifier takes as input a latent representations with four channels and a size of $28 \times 28$. The architecture is based on either a ViT or a VAE decoder. The ViT-based architecture consists of twelve transformer encoder blocks each with three attention heads and a hidden dimension of 192. The patch size is set to 4. The VAE-decoder-based architecture uses the first block of the VAE decoder of the stable diffusion model. These latent classifiers are pre-trained on ImageNet-1k independently, and are frozen. Either one of them is used to compute the LC loss.

**Noise schedule.** At each training iteration, a noise is applied to clean latent generative representations with a scheduled hyperparameter $\alpha_t$. Specifically, $\alpha_t$ is defined by $\alpha_t = \prod_{\tau=0}^{t}(1-\beta_\tau^2)$ where $\beta_t$ is a linearly scheduled parameter from $\sqrt{8.5\times10^{-4}}$ to $\sqrt{1.2\times10^{-2}}$ according to timestep $\mathcal{T} = \{0, 1, 2, \cdots, T\}$, where $T = 1000$.

**Training details.** Unless otherwise noted, follow the training setting used in [86]. Specifically, the AdamW optimizer is used for 300 epochs with an initial learning rate of $1.0 \times 10^{-5}$ and a batch size of 1024. The learning rate was gradually increased to $5.0 \times 10^{-4}$ according to the cosine scheduler [56]. The weight decay (L2 penalty) of 0.05 is used. The latent classifiers are trained for 100 epochs.

**Datasets and metrics.** The ImageNet-1k dataset [70] is used as a training dataset. It consists of 1.2 million labeled images for 1,000 objects. For testing, we use its validation set and three distribution-shifted datasets: ImageNet-V2 [66], ImageNet-Sketch [90] and ImageNet-R [29]. These datasets help evaluate the robustness beyond the original ImageNet dataset. We also evaluate fine-tuning performance on ten datasets: CIFAR-10 (CF10) [43], CIFAR-100 (CF100) [43], Oxford Flowers-102 [58], Stanford Cars [42], Oxford-IIIT Pets [59], Food-101 [11], DTD [14], STL10 [16], iNaturalist18 (iNat18), and iNaturalist19 (iNat19) [88]. These datasets help evaluate whether the effect of generative regularization remains in downstream tasks. We use accuracy as the evaluation metric.

## 4.2. Experimental results

### 4.2.1 ImageNet Evaluation

**ImageNet-1k.** Table 1 compares training with and without generative regularization loss. On the ImageNet-1k validation dataset, the top-1 accuracy was improved by 0.2 points with the NC loss and by 0.4 points with the LC loss, regardless of the choice of the latent classifier. This demonstrates that both NC and LC losses are effective when the training and testing domains are the same.

**Robustness.** The results on the three datasets for evaluating robustness are also summarized in Table 1. As shown, the NC loss improved performance on ImageNet-V2 but decreased performance on ImageNet-Sketch and ImageNet-R. This result suggests that training robust discriminative representations using image generation models is not straightforward and that minimizing the distance between sampled noise and predicted noise during training can lead to overfitting to the training image distribution. This problem was adequately addressed by the LC loss. As shown, it improved performance on all datasets. This is because training with the latent classifier using cross-entropy loss enhanced training of more robust discriminative representations as discussed in Section 3.2.2.

**Latent classifier.** To analyze the behavior of the latent classifiers, Table 2 shows the performance of the two pre-trained latent classifiers on ImageNet-1k. As shown, both underperformed compared to the baseline in Table 1, which learns a classifier directly from clean images. Nevertheless, incorporating these latent classifiers into the generative regularization loss enhanced supervised discriminative learning. These results contrast with knowledge distillation, which requires a high-performance classifier as a teacher. Furthermore, while the VAE-decoder-based latent classifier has more parameters and achieves higher performance than the ViT-based one, the LC loss using the ViT-based latent

| Pre-training loss | Latent cls. | CF10 | CF100 | Flowers | Cars | Pets | Food | DTD | STL10 | iNat18 | iNat19 | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\mathcal{L}_{\text{dis}}$ (baseline) | - | 99.0 ±0.008 | 91.0 ±0.1 | 96.4 ±0.3 | 92.0 ±0.05 | 95.0 ±0.1 | 91.4 ±0.05 | 73.3 ±0.2 | **99.4** ±0.04 | 72.7 ±0.06 | 77.3 ±0.3 | 88.8 |
| $\mathcal{L}_{\text{dis}} + \mathcal{L}_{\text{gen}}^{\text{NC}}$ | - | **99.2** ±0.02 | **91.3** ±0.01 | **97.0** ±0.2 | **92.1** ±0.08 | **95.2** ±0.2 | **91.5** ±0.05 | **73.7** ±0.5 | 99.3 ±0.04 | 73.1 ±0.1 | **77.8** ±0.2 | **89.0** |
| $\mathcal{L}_{\text{dis}} + \mathcal{L}_{\text{gen}}^{\text{LC}}$ | VAE dec. | 99.0 ±0.008 | 91.1 ±0.08 | **97.0** ±0.1 | **92.1** ±0.05 | 95.0 ±0.06 | 91.4 ±0.08 | 73.0 ±0.5 | **99.4** ±0.01 | 73.0 ±0.1 | 77.3 ±0.3 | 88.9 |
| $\mathcal{L}_{\text{dis}} + \mathcal{L}_{\text{gen}}^{\text{LC}}$ | ViT | **99.2** ±0.02 | 91.2 ±0.06 | 96.5 ±0.3 | 92.0 ±0.1 | 95.1 ±0.02 | **91.5** ±0.06 | 73.3 ±0.4 | 99.3 ±0.02 | **73.5** ±0.04 | 77.7 ±0.4 | **89.0** |

Table 3. Fine-tuning Top-1 accuracy (%) on ten image classification dataset. The effect of generative regularization in pre-training on ImageNet persists even after fine-tuning.

| Approach | ImageNet-1k | ImageNet-V2 | ImageNet-Sketch | ImageNet-R |
|---|---|---|---|---|
| Baseline | 81.8 | 70.4 | 32.0 | 30.4 |
| KD (OpenCLIP-ViT/H image encoder) | 81.6 (-0.2) | 70.7 (+0.3) | 32.9 (+0.9) | 31.5 (+1.1) |
| KD (OpenCLIP-ViT/H text encoder) | 81.9 (+0.1) | 70.9 (+0.5) | 32.2 (+0.2) | 31.2 (+0.8) |
| Data augmentation | 81.7 (-0.1) | 70.9 (+0.5) | 32.3 (+0.3) | 31.4 (+1.0) |
| Consistency regularization | 77.8 (-4.0) | 65.4 (-5.0) | 26.7 (-5.3) | 25.0 (-5.4) |
| Training with real & synthetic images | 81.9 (+0.1) | 70.8 (+0.4) | 33.1 (+1.1) | 31.9 (+1.5) |
| Generative regularization (ours) | 82.2 (+0.4) | 71.0 (+0.6) | 33.2 (+1.2) | 31.8 (+1.4) |
| w/ KD (OpenCLIP-ViT/H text encoder) | **83.0** (+1.2) | **72.4** (+2.0) | **34.3** (+2.3) | **32.9** (+2.5) |

Table 4. Comparison with other representative approaches. Our method is compared with knowledge distillation (KD) using OpenCLIP, consistency regularization, data augmentation using noisy images, and training with real and synthetic SD-ImageNet images.

classifier performed better, particularly on the distributed-shifted datasets (ImageNet-Sketch and ImageNet-R) in Table 1. This also indicates that a high-performance latent classifier does not necessarily yield better results in terms of robustness.

#### 4.2.2 Fine-tuning Evaluation

We evaluated whether the effect of generative regularization persists in downstream tasks by fine-tuning the model obtained in Section 4.2.1 for the ten image classification datasets. For all datasets, the SGD optimizer is used with an initial learning rate of 0.01 and momentum of 0.9, for 1,000 for Flowers and Stanford Cars, and for 300 for the other datasets. A weight decay of $1 \times 10^{-4}$ is used for the fine-tuning experiments. The learning rate scheduler and other hyperparameters are the same as those in Section 4.1. Fine-tuning was conducted five times for each dataset, and the average accuracy is reported. The fine-tuning results are shown in Table 3. For all datasets, the highest accuracy was achieved by models trained using the proposed method. In contrast to the findings on the ImageNet evaluation, the NC loss performed comparable with the LC loss.

#### 4.2.3 Comparison with other approaches

This subsection compares our approach with existing methods that apply loss to the outputs of the encoder network or utilize a diffusion model. In all experiments, we used the same hyperparameters as described in Section 4.1.

**Comparison with knowledge distillation using CLIP.** The Stable Diffusion model we used in the above experiments is a text-to-image translation model conditioned by text prompts encoded by the text encoder of OpenCLIP-ViT/H [13]. This raises the question of whether learning representations with knowledge distillation, using the encoder of OpenCLIP-ViT/H as a teacher, is reasonable. To investigate this, we trained a model with an $L_2$ loss between learnable representations $f_{\theta_1}(\boldsymbol{x}_i)$ and a teacher encoder $s(\boldsymbol{x}_i)$ on ImageNet, where $s$ is either the image or text encoder of the OpenCLIP-ViT/H. For the text encoder, we used the prompt "a photo of a {classname}".

The results are shown in the second and third rows of Table 4. As shown, both encoders improved accuracy, except for the image encoder on ImageNet-1k. The text encoder was more effective than the image encoder on ImageNet-1k and ImageNet-V2, while the opposite is true on ImageNet-Sketch and ImageNet-R. Overall, our approach surpassed both. This shows the effectiveness of our generative regularization loss. Furthermore, it is worth noting that our approach and knowledge distillation are complementary and benefit each other. As shown in the last row of Table 4, combining our approach with knowledge distillation significantly improved the classification accuracy across all datasets.

**Comparison with data augmentation.** It is known that

adding diffusion-like noise to the training image makes the image classifier more robust [36]. From this perspective, we added noise to the images as data augmentation while training the image classifier on ImageNet and compared the results with those of our approach. In this experiment, noise is added to 5% of the total training images. The image classifier is trained with only $\mathcal{L}_{\text{dis}}$ loss.

The results are shown in the fourth row of Table 4. Adding noise improved robustness for the distribution-shifted datasets but slightly reduced accuracy on ImageNet. In contrast, our method demonstrated superior performance across all datasets compared to data augmentation via noise addition. These results indicate that the proposed method is effective beyond noise-based data augmentation.

**Comparison with consistency regularization.** Another potential regularization method that employs noise is consistency regularization. This approach utilizes both clean image $\boldsymbol{x}_i$ and noisy image $\boldsymbol{x}_i^{(t)}$. This is analogous to the NC loss in the proposed method, which regularizes the image classifier by ensuring consistency between the input and predicted noise. To compare this approach with our method, we regularized the image classifier training with an $L_2$ loss between a clean logit $h_{\theta_2}(f_{\theta_1}(\boldsymbol{x}_i))$ and a noisy logit $h_{\theta_2}(f_{\theta_1}(\boldsymbol{x}_i^{(t)}))$ on ImageNet.

The results are shown in the fifth row of Table 4. Consistency regularization reduced the accuracy on all the datasets, while regularization with our NC loss improved the accuracy on ImageNet.

**Comparison with training using real and synthetic images.** As discussed in Section 2, adding synthetic images to the training dataset is the most prevalent method for training image classifiers with generative models. We prepared SD-ImageNet [27], a synthetic dataset created using Stable Diffusion, whose images correspond to the ImageNet class categories. The image classifier was trained on both real and synthetic images. A total of 240,000 synthetic images were created, representing approximately 20% of the number of images in ImageNet.

The results are shown in the sixth row of Table 4. Training with real and synthetic images improved performance on all datasets. Our method outperformed training with synthetic images on the in-distribution dataset and two distribution-shifted datasets. As mentioned above, our method does not require prompt engineering, which is the main bottleneck in synthetic image generation.

### 4.3. Analysis

**Distributional discrepancy.** To analyze the effectiveness of our method when the distribution of training images deviates from that of the generative model, we conducted digit classification experiments using the SVHN dataset [54], which consists of street view digit images. ViT-Ti and ViT-S are used for the encoder network $f_{\theta_1}$. The results are shown in Table 5. Interestingly, our method improved the accuracy, while the other methods did not. This demonstrates the capability of our method to enhance performance even in the presence of distributional discrepancies.

| Method | ViT-Ti | ViT-S |
|---|---|---|
| Baseline | 96.0 | 96.8 |
| Training with real & synthetic images | 93.6 | 94.7 |
| KD w/ OpenCLIP-ViT/H text encoder | 95.6 | 96.8 |
| $\mathcal{L}_{\text{dis}} + \mathcal{L}_{\text{gen}}^{\text{NC}}$ (Ours) | **96.2** | **97.0** |

Table 5. Digit classification accuracy (%) on the SVHN dataset.

| Loss | SD version | ImageNet Top-1 (%) |
|---|---|---|
| $\mathcal{L}_{\text{dis}} + \mathcal{L}_{\text{gen}}^{\text{NC}}$ | 2.0 | 81.5 |
| | 2.1 | 82.0 |
| $\mathcal{L}_{\text{dis}} + \mathcal{L}_{\text{gen}}^{\text{LC}}$ | 2.0 | 82.0 |
| | 2.1 | 82.2 |

Table 6. Results using different versions of Stable Diffusion. Higher performance is shown when using a newer version of SD.

**Diffusion model selection** The effect of different versions of Stable Diffusion, which is used to train ImageNet in the image classifier, was examined. The results are shown in Table 6. Regardless of whether the NC loss or LC loss was used, higher accuracy was achieved when using Stable Diffusion V2.1. This finding is convincing because a more trained diffusion model is likely to provide a more comprehensive understanding of the vision task. As image generation models continue to advance in the future, the effectiveness of our method is expected to further improve.

## 5. Conclusion and Future Work

We proposed diffusion-based generative regularization, a framework to train discriminative models using a pre-trained generative model. We introduced two loss functions, the noise consistency loss and the latent cross-entropy loss, which improve discriminative representations without the need for generating synthetic images. Experiments on ImageNet showed that our framework improves image classification accuracy and robustness to distribution shifts.

**Limitation and future work.** The computational cost of the proposed method increases with the size of the diffusion model used for regularization. The efficiency of training with large diffusion models is an important issue for the future. In addition, this paper does not discuss the performance of the proposed method when utilizing diffusion models other than Stable Diffusion. Our method is applicable to arbitrary conditioned diffusion models, but its evaluation is future work.

**Acknowledgements** This work was supported by JSPS KAKENHI Grant Number JP23H00490 and NEDO JPNP18002.

## References


[1] Abulikemu Abuduweili, Xingjian Li, Humphrey Shi, Cheng-Zhong Xu, and Dejing Dou. Adaptive consistency regularization for semi-supervised transfer learning. In *Proc. CVPR*, pages 6923–6932, 2021. 3

[2] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J. Fleet. Synthetic data from diffusion models improves imagenet classification. *TMLR*, 2023. 1

[3] Philip Bachman, Ouais Alsharif, and Doina Precup. Learning with pseudo-ensembles. In *Proc. NeurIPS*, 2014. 3

[4] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. eDiff-I: Text-to-Image Diffusion Models with an Ensemble of Expert Denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 2

[5] Hritik Bansal and Aditya Grover. Leaving reality to imagination: Robust classification via generated datasets. In *Proc. ICLR RTML Workshop*, 2023. 2, 3

[6] Fan Bao, Chongxuan Li, Jun Zhu, and Bo Zhang. Analytic-DPM: an analytic estimate of the optimal reverse variance in diffusion probabilistic models. In *Proc. ICLR*, 2022. 2

[7] Sara Beery, Yang Liu, Dan Morris, Jim Piavis, Ashish Kapoor, Neel Joshi, Markus Meister, and Pietro Perona. Synthetic examples improve generalization for rare classes. In *Proc. WACV*, 2020. 1

[8] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Proc. NeurIPS*, 2019. 3

[9] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, and et al. Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2

[10] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proc. CVPR*, pages 22563–22575, 2023. 2

[11] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 - mining discriminative components with random forests. In *Proc. ECCV*, pages 446–461, 2014. 6

[12] Huanran Chen, Yinpeng Dong, Shitong Shao, Zhongkai Hao, Xiao Yang, Hang Su, and Jun Zhu. Your diffusion model is secretly a certifiably robust classifier. *arXiv preprint arXiv:2402.02316*, 2024. 2

[13] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proc. CVPR*, pages 2818–2829, 2023. 2, 7

[14] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proc. CVPR*, 2014. 6

[15] Kevin Clark and Priyank Jaini. Text-to-image diffusion models are zero shot classifiers. In *NeurIPS*, pages 58921–58937, 2023. 2

[16] Adam Coates, Andrew Ng, and Honglak Lee. An Analysis of Single Layer Networks in Unsupervised Feature Learning. In *Proc. AISTATS*, 2011. 6

[17] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, Rodolphe Jenatton, Lucas Beyer, Michael Tschannen, Anurag Arnab, Xiao Wang, Carlos Riquelme Ruiz, Matthias Minderer, Joan Puigcerver, Utku Evci, Manoj Kumar, Sjoerd Van Steenkiste, Gamaleldin Fathy Elsayed, Aravindh Mahendran, Fisher Yu, Avital Oliver, Fantine Huot, Jasmijn Bastings, Mark Collier, Alexey A. Gritsenko, Vighnesh Birodkar, Cristina Nader Vasconcelos, Yi Tay, Thomas Mensink, Alexander Kolesnikov, Filip Pavetic, Dustin Tran, Thomas Kipf, Mario Lucic, Xiaohua Zhai, Daniel Keysers, Jeremiah J. Harmsen, and Neil Houlsby. Scaling vision transformers to 22 billion parameters. In *Proc. ICML*, pages 7480–7512, 2023. 1

[18] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat GANs on image synthesis. In *Proc. NeurIPS*, 2021. 1, 2

[19] Tim Dockhorn, Arash Vahdat, and Karsten Kreis. Genie: Higher-order denoising diffusion solvers. In *Proc. NeurIPS*, pages 30150–30166, 2022. 2

[20] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proc. ICCV*, pages 7346–7356, 2023. 2

[21] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proc. ICML*, 2024. 3

[22] Lijie Fan, Kaifeng Chen, Dilip Krishnan, Dina Katabi, Phillip Isola, and Yonglong Tian. Scaling laws of synthetic images for model training ... for now. In *CVPR*, pages 7382–7392, 2024. 1, 3

[23] Haoyang Fang, Boran Han, Shuai Zhang, Su Zhou, Cuixiong Hu, and Wen-Ming Ye. Data augmentation for object detection via controllable diffusion models. In *Proc. WACV*, pages 1257–1266, 2024. 2

[24] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. In *Proc. NeurIPS*, pages 50742–50768, 2023. 2

[25] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proc. NeurIPS*, 2014. 3

[26] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. In *Proc. ICCV*, pages 7323–7334, 2023. 3

[27] Ryuichiro Hataya, Han Bao, and Hiromi Arai. Will large-scale generative models corrupt future datasets? In *Proc. ICCV*, 2023. 1, 2, 3, 8

[28] Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiajuan Qi. Is synthetic data from generative models ready for image recognition? In *Proc. ICLR*, 2023. 2

[29] 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. *Proc. ICCV*, 2021. 2, 6

[30] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *Proc. NeurIPS DLRL Workshop*, 2015. 2

[31] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proc. NeurIPS*, pages 6840–6851, 2020. 2

[32] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *Proc. NeurIPS DGMDA Workshop*, 2021. 2

[33] Chenqing Hua, Sitao Luan, Minkai Xu, Zhitao Ying, Jie Fu, Stefano Ermon, and Doina Precup. Mudiff: Unified diffusion for complete molecule generation. In *Proc. LOG Conference*, volume 231, pages 33:1–33:26, 2024. 2

[34] Qingqing Huang, Daniel S. Park, Tao Wang, Timo I. Denk, Andy Ly, Nanxin Chen, Zhengdong Zhang, Zhishuai Zhang, Jiahui Yu, Christian Frank, Jesse Engel, Quoc V. Le, William Chan, Zhifeng Chen, and Wei Han. Noise2Music: Text-conditioned Music Generation with Diffusion Models. *arXiv preprint arXiv:2302.03917*, 2023. 2

[35] Sungsu Hur, Inkyu Shin, Kwanyong Park, Sanghyun Woo, and In So Kweon. Learning classifiers of prototypes and reciprocal points for universal domain adaptation. In *Proc. WACV*, pages 531–540, 2023. 3

[36] Priyank Jaini, Kevin Clark, and Robert Geirhos. Intriguing properties of generative classifiers. In *Proc. ICLR*, 2024. 2, 8

[37] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Proc. NeurIPS*, pages 26565–26577, 2022. 2

[38] Diederik P Kingma and Ruiqi Gao. Understanding diffusion objectives as the ELBO with simple data augmentation. In *Proc. NeurIPS*, 2023. 2

[39] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *Proc. ICLR*, 2014. 2

[40] Minsu Ko, Eunju Cha, Sungjoo Suh, Huijin Lee, Jae-Joon Han, Jinwoo Shin, and Bohyung Han. Self-supervised dense consistency regularization for image-to-image translation. In *Proc. CVPR*, pages 18301–18310, 2022. 3

[41] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *Proc. ICLR*, 2021. 2

[42] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proc. ICCV 3dRR Workshop*, pages 554–561, 2013. 6

[43] Alex Krizhevsky. Learning multiple layers of features from tiny images. *Master's thesis, University of Tront*, 2009. 6

[44] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *Proc. ICLR*, 2017. 3

[45] Daiqing Li, Huan Ling, Amlan Kar, David Acuna, Seung Wook Kim, Karsten Kreis, Antonio Torralba, and Sanja Fidler. Dreamteacher: Pretraining image backbones with deep generative models. In *Proc. ICCV*, pages 16698–16708, 2023. 2, 3

[46] Zheng Li, Yuxuan Li, Penghai Zhao, Renjie Song, Xiang Li, and Jian Yang. Is synthetic data from diffusion models ready for knowledge distillation? *arXiv preprint arXiv:2305.12954*, 2023. 2, 3

[47] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proc. CVPR*, pages 300–309, 2023. 2

[48] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. AudioLDM: Text-to-audio generation with latent diffusion models. In *Proc. ICML*, pages 21450–21474, 2023. 2

[49] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, Lifang He, and Lichao Sun. Sora: A Review on Background, Technology, Limitations, and Opportunities of Large Vision Models. *arXiv preprint arXiv:2402.17177*, 2024. 3

[50] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan LI, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. In *Proc. NeurIPS*, pages 5775–5787, 2022. 2

[51] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver++: Fast Solver for Guided Sampling of Diffusion Probabilistic Models. *arXiv preprint arXiv:2211.01095*, 2022. 2

[52] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proc. CVPR*, pages 2837–2845, 2021. 2

[53] Aamir Mustafa, Rafał K. Mantiuk, Horst Bischof, Thomas Brox, and Jan-Michael Frahm. Transformation consistency regularization – a semi-supervised paradigm for image-to-image translation. In *Proc. ECCV*, pages 599–615, 2020. 3

[54] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NeurIPS DLUFL Workshop*, 2011. 8

[55] Quang Nguyen, Truong Vu, Anh Tran, and Khoi Nguyen. Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation. In *Proc. NeurIPS*, pages 76872–76892, 2023. 2

[56] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proc. ICML*, pages 8162–8171, 2021. 6

[57] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photo-realistic image generation and editing with text-guided dif-

fusion models. In *Proc. ICML*, pages 16784–16804, 2022. 2

[58] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Proc. ICVGIP*, pages 722–729, 2008. 6

[59] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar. Cats and dogs. In *Proc. CVPR*, 2012. 6

[60] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proc. ICCV*, pages 4195–4205, October 2023. 3

[61] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *Proc. ICLR*, 2024. 2

[62] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *Proc. ICLR*, 2023. 2

[63] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proc. ICML*, pages 8748–8763, 2021. 2

[64] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv preprint arXiv:2204.06125*, 2022. 2

[65] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *Proc. ICML*, pages 8821–8831, 2021. 1

[66] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *Proc. ICML*, pages 5389–5400, 2019. 2, 6

[67] Julius Richter, Simon Welker, Jean-Marie Lemercier, Bunlong Lay, and Timo Gerkmann. Speech enhancement and dereverberation with diffusion-based generative models. *IEEE/ACM TASLP*, 31:2351–2364, 2023. 2

[68] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. CVPR*, pages 10684–10695, 2022. 1, 2, 3

[69] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *Proc. ICLR*, 2015. 2

[70] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 115(3):211–252, 2015. 2, 6

[71] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Proc. NeurIPS*, pages 36479–36494, 2022. 2, 3

[72] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *Proc. NeurIPS*, 2016. 3

[73] Mert Bülent Sarıyıldız, Karteek Alahari, Diane Larlus, and Yannis Kalantidis. Fake it till you make it: Learning transferable representations from synthetic imagenet clones. In *Proc. CVPR*, pages 8011–8021, 2023. 1, 3

[74] Samarth Sinha and Adji Bousso Dieng. Consistency regularization for variational auto-encoders. In *Proc. NeurIPS*, pages 12943–12954, 2021. 3

[75] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proc. ICML*, pages 2256–2265, 2015. 2

[76] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A. Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Proc. NeurIPS*, pages 596–608, 2020. 2, 3

[77] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *Proc. ICLR*, 2021. 2

[78] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Proc. NeurIPS*, 2019. 2

[79] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *Proc. ICLR*, 2021. 2

[80] Longxiang Tang, Kai Li, Chunming He, Yulun Zhang, and Xiu Li. Consistency regularization for generalizable source-free domain adaptation. In *ICCV OODCV Workshops*, pages 4323–4333, 2023. 3

[81] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Proc. NeurIPS*, 2017. 3

[82] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *Proc. ICLR*, 2023. 2

[83] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proc. CVPR*, pages 5238–5248, 2022. 2

[84] Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. In *Proc. NeurIPS*, pages 48382–48402, 2023. 2

[85] Aysim Toker, Marvin Eisenberger, Daniel Cremers, and Laura Leal-Taixé. Satsynth: Augmenting image-mask pairs through diffusion models for aerial semantic segmentation. In *Proc. CVPR*, pages 27695–27705, 2024. 2

[86] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In *Proc. ICML*, pages 10347–10357, 2021. 2, 6

[87] Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. In *Proc. NeurIPS*, 2021. 2

[88] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The INaturalist species classification and detection dataset. In *Proc. CVPR*, 2018. 6

[89] Vikas Verma, Alex Lamb, Juho Kannala, Yoshua Bengio, and David Lopez-Paz. Interpolation consistency training for semi-supervised learning. In *Proc. IJCAI*, pages 3635–3641, 2019. 3

[90] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Proc. NeurIPS*, pages 10506–10518, 2019. 2, 6

[91] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. In *Proc. NeurIPS*, pages 6256–6268, 2020. 3

[92] Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. Geodiff: A geometric diffusion model for molecular conformation generation. In *Proc. ICLR*, 2022. 2

[93] Lihe Yang, Xiaogang Xu, Bingyi Kang, Yinghuan Shi, and Hengshuang Zhao. Freemask: Synthetic images with dense annotations make stronger segmentation models. In *Proc. NeurIPS*, pages 18659–18675, 2023. 2

[94] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proc. CVPR*, pages 12104–12113, 2022. 1

[95] Han Zhang, Zizhao Zhang, Augustus Odena, and Honglak Lee. Consistency regularization for generative adversarial networks. In *Proc. ICLR*, 2020. 3

[96] Bo Zhao and Hakan Bilen. Synthesizing informative training samples with gan. In *NeurIPS SyntheticData4ML Workshop*, 2022. 1

[97] Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. UniPC: A unified predictor-corrector framework for fast sampling of diffusion models. In *Proc. NeurIPS*, 2023. 2

[98] Zhengli Zhao, Sameer Singh, Honglak Lee, Zizhao Zhang, Augustus Odena, and Han Zhang. Improved consistency regularization for gans. In *Proc. AAAI*, pages 11033–11041, 2021. 3

[99] Kaiwen Zheng, Cheng Lu, Jianfei Chen, and Jun Zhu. Dpm-solver-v3: Improved diffusion ode solver with empirical model statistics. In *Proc. NeurIPS*, pages 55502–55542, 2023. 2

[100] Yongchao Zhou, Hshmat Sahak, and Jimmy Ba. Training on Thin Air: Improve Image Classification with Generated Data. *arXiv preprint arXiv:2305.15316*, 2023. 2

[101] Zi-Xin Zou, Zhipeng Yu, Yuan-Chen Guo, Yangguang Li, Ding Liang, Yan-Pei Cao, and Song-Hai Zhang. Triplane meets gaussian splatting: Fast and generalizable single-view 3d reconstruction with transformers. In *Proc. CVPR*, pages 10324–10335, 2024. 3