MixDiff: Mixing Natural and Synthetic Images for Robust Self-Supervised Representations

Reza Akbarian Bafghi[1*] Nidhin Harilal[1*] Claire Monteleoni[1,2] Maziar Raissi[3]
[1] University of Colorado, Boulder [2] INRIA, Paris [3] University of California, Riverside
{reza.akbarianbafghi, nidhin.harilal, cmontel}@colorado.edu, maziar.raissi@ucr.edu

Abstract

This paper introduces MixDiff, a new self-supervised learning (SSL) pre-training framework that combines real and synthetic images. Unlike traditional SSL methods that predominantly use real images, MixDiff uses a variant of Stable Diffusion to replace an augmented instance of a real image, facilitating the learning of cross real-synthetic image representations. Our key insight is that while models trained solely on synthetic images underperform, combining real and synthetic data leads to more robust and adaptable representations. Experiments show MixDiff enhances SimCLR, BarlowTwins, and DINO across various robustness datasets and domain transfer tasks, boosting SimCLR's ImageNet-1K accuracy by 4.56%. Our framework also demonstrates comparable performance without needing any augmentations, a surprising finding in SSL where augmentations are typically crucial. Furthermore, MixDiff achieves similar results to SimCLR while requiring less real data, highlighting its efficiency in representation learning[1].

1. Introduction

Self-supervised learning (SSL) has enjoyed significant advancements in recent years [26]. The capability of joint-embedding SSL architectures to generate high-quality features using pretext tasks now parallels, and in some cases surpasses, that of supervised learning. Joint-embedding SSL methods can include distillation [12, 25, 76] or contrastive strategies [13, 29, 38, 81], where multiple network branches aim to learn representations by maximizing agreement between differently augmented views of the same data example in the embedding space. Even though such recent advancements in SSL save annotation costs, preparing training data is still challenging. Current research within SSL predominantly concentrates on the development of the pretext tasks, while the characteristics of the data being utilized for learning remain less explored.

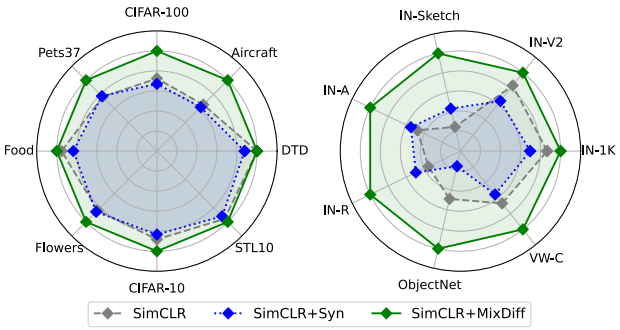

Figure 1. Comparison of SimCLR performance on real, synthetic (Syn), and mixed real and synthetic images (MixDiff). The radar charts show normalized accuracy across 8 transfer learning datasets (left) and ImageNet-1K plus 6 distribution shift datasets (right), with values from 0.5 to 1.1. MixDiff enhances in-distribution and robustness performance and generalizes better. More details in Sec. 4.

This focus on pretext tasks, however, has led researchers to consider alternative data sources to further improve model performance. In particular, there has been a growing interest in using synthetic images for self-supervision [62, 70, 74]. The appeal of synthetic image datasets lies in their ease of generation, a wider range of semantic content, and minimal human intervention, addressing key concerns in computer vision like cost efficiency and fairness in data collection and annotation. Although generative models address the issue of data scarcity, exclusive reliance on synthetic images for supervision is not without drawbacks. The main challenge has been the domain gap between synthetic and real-world data. Models trained only on synthetic images often struggle to adapt to real-world settings due to their limited exposure to the variability and complexity of natural images [66]. This issue is particularly pronounced in large-scale image recognition tasks, where models trained on synthetic data typically underperform those trained on real images [36, 67]. In response to this, our research proposes a novel training framework that integrates both real and syn-

*Joint first-authorship.

[1]We have made the source code and generated data available to the public at: https://github.com/cryptonymous9/mixing-ssl

thetic images, aiming to harness the strengths of each data source while mitigating their individual weaknesses.

In this paper, we introduce MixDiff to explore the potential of combining synthetic images generated by generative models without any labeled data with real-world images for SSL training. MixDiff is a simple framework that replaces an augmented instance of a real image in an existing joint-embedding SSL pipeline with a synthetic image from a generative model. The simplicity of the framework allows it to incorporate it in existing SSL methods like SimCLR [13], BarlowTwins [81] and DINO [12] as shown in Figure 2.

MixDiff works on the idea that synthetically generated images generate harder positive pairs that makes the overall training objective less trivial [55]. We find that SSL models pre-trained exclusively on synthetic images underperform compared to those pre-trained with real images across most scenarios. Interestingly, our proposed MixDiff, which uses both real and synthetic images, improves model performance in SSL not only on in-distribution datasets but also on various out-of-distribution tasks as we show in Figure 1, suggesting enhanced representation learning. Specifically, we observe an average increase in top-1 accuracy of about 26.92% across six distributional datasets and a 7.36% improvement in transfer learning across eight datasets. This observation suggests that while synthetic images alone may be insufficient for optimal pre-training, MixDiff capitalizes on the synergistic approach of leveraging the strengths of both image types to learn more robust SSL representations.

Building on these results, we investigated how synthetic image quality affects model performance in SSL. Prior research has shown a strong correlation between the quality of diffusion-generated synthetic images and model performance. Our study reveals that MixDiff is less sensitive to variations in synthetic image quality, potentially reducing the need for precise quality optimization. We also find that integrating synthetic data in models like DINO may de-emphasize background features, suggesting enhanced scene layout understanding beneficial for image segmentation. Despite the computational cost of generating synthetic images, MixDiff requires fewer real images to match SimCLR performance, indicating more efficient SSL pre-training. The reduced reliance on large real datasets and high-quality synthetic images, coupled with robustness to distributional shifts and transfer learning, positions MixDiff as a promising approach for enhancing SSL pre-training.

2. Related Work

Self-supervised Learning. While SSL techniques exist in different forms, one of the most successful self-supervised learning paradigms is joint-embedding SSL [13, 29, 46, 48, 78, 81]. The main focus of joint-embedding SSL is instance-based discriminative learning [3, 22], where each image is considered to be its own class, and a model is trained by discriminating different views of the same image generated using data augmentation [12, 13, 29, 81]. One such example is SimCLR [13], which uses an InfoNCE-based formulation [48] to bring in the representation of different views of the same image closer (positive pairs) and repel representations of views from different images (negative pairs) apart. In most cases, joint-embedding SSL methods work in a Siamese setting [14] where two branches have identical architectures and share weights. However, networks such as the Siamese setting are vulnerable to collapsing to trivial representations. BarlowTwins [81] brings covariance regularization to the contrastive setting to enforce a non-collapsing solution. More recently, works such as DINO [12] have shown alternative ways to prevent collapse using architectural strategies inspired by knowledge distillation [34] and addressing catastrophic forgetting [6]. We provide an improved pre-training mechanism for representation learning in such joint-embedding SSL techniques.

Learning using Synthetic Data. Recent advancements in machine learning have increasingly leveraged synthetic data across a variety of domains [28, 44, 45, 60]. This type of data is particularly crucial for tasks that demand extensive labeled datasets, such as human pose estimation [27, 42], semantic segmentation [15, 54], optical flow estimation [61, 73], and language models [1, 24, 68]. In the task of image classification, several studies have demonstrated the effectiveness of synthetic data [72]. [30, 65] illustrates its application in data-scarce settings and transfer learning; [71] explores its role in enhancing adversarial training; [23] diversifies images; [7, 59] evaluates model robustness against natural distribution shifts using synthetic data; and [4, 65] discusses augmentation of images through fine-tuned diffusion models. It is important to note that all of these studies focus on supervised learning. Our work, however, is distinct in its concentration on SSL. Recently, there has been significant interest in leveraging synthetic data for SSL [62, 70, 74]. [62] employ text-to-image diffusion models to generate multiple images from a single caption, while our approach uses image-to-image diffusion models to produce one image per source, reducing the dependency on labeled data and associated costs. [74] introduces a data generation framework to enhance contrastive learning, with the generator trained with the SSL model. [70] train diffusion models and mix real and synthetic data to boost contrastive learning through data augmentation and inflation. Unlike these methods, our approach does not require training a new generative model. Instead, we utilize off-the-shelf variants of Stable Diffusion [56, 77] to improve the quality of representations in existing SSL models.

Generative Models. The landscape of synthetic image generation has seen a significant evolution, with Generative Adversarial Networks (GANs) such as BigGAN [11] ini-

tially setting a high standard. These models have been pivotal in pushing the boundaries of image realism and quality. Recently, diffusion models have emerged as a promising alternative, demonstrating impressive results in both conditional [19, 58] and unconditional [35] synthetic image generation. Text-to-image diffusion models like DALL-E [52] and Imagen [57] are notable examples, showcasing the ability to create detailed and contextually accurate images from textual descriptions. Our research takes a unique turn by focusing on the image-to-image diffusion model, specifically a fine-tuned version of Stable Diffusion [56]. This model distinguishes itself by utilizing CLIP [51] image embeddings instead of text embeddings.

3. Method

In this work, we focus on Self-Supervised Learning (SSL) techniques, particularly those that consolidate representations from different perspectives or augmentations of the same instance [2, 13, 29, 75]. The main idea behind this technique is, through iterative processes, these representations gradually become less sensitive to the transformations generating these varied views. Consequently, this leads to the learning of image representations that are notably effective for vision tasks such as classification [13, 29, 81]. In this section, we introduce our framework, MixDiff, which uniquely employs both real and synthetically generated data through stable diffusion, and see how it can be incorporated in some of the existing SSL frameworks.

3.1. Description of MixDiff

Consider x_1 and x'_1, two augmented patches from an image, randomly selected from a dataset. These augmentations can include a variety of changes, such as altering spatial positions within an image, adding varying noise and applying random color adjustments, etc. Existing instance-based discriminative SSL methods primarily rely on real images [12, 13, 29, 81]. In these methods, the representation derived from the first augmentation, x_1, of a real image is anticipated to closely align with the representation of the second augmentation, x'_1, of the same image as shown in Figure 1. Our MixDiff framework modifies this approach by incorporating synthetically generated images alongside real ones. The primary objective of MixDiff is to synchronize the representations of real and synthetic images, thus enhancing existing SSL methodologies such as SimCLR, DINO, and BarlowTwins.

The synthetic images employed by MixDiff have a unique characteristic that they share a variation of the same semantic component or object with that of the real image. To achieve this, we employ Image Variation Diffuser [50][2], a variant of Stable Diffusion (SD) [56] tailored to generate diverse images while preserving semantic categories or in simple terms, the image class. In the SD-based generative model, represented as $g^k_{SD}(\cdot)$, where k indicates the guidance scale influencing the generative features from the input image, an input $x_i \sim D$ yields a synthetic counterpart $\tilde{x_i}$, such that $\tilde{x_i} = g^k_{SD}(x_i)$ The innovative aspect of MixDiff lies in substituting a portion of the augmentation process, i.e, the second branch of augmentation x'_i within the SSL framework with these synthetic images $\tilde{x_i}$, to learn cross real-synthetic image representations a shown in Figure 1.

[2]https://huggingface.co/lambdalabs/sd-image-variations-diffusers

3.2. Mixing in joint-embedding SSL

SimCLR + MixDiff: In SimCLR's contrastive setup [13], 'positive' and 'negative' pairs of images are identified, with the goal of either converging or diverging their representations. In SimCLR, two augmented views are generated for each image in a mini-batch, resulting in $2N$ images for a mini-batch size of N. Each view is paired with its corresponding alternate view as a 'positive' pair, while the remaining $2(N-1)$ images are treated as 'negative' pairs. We now propose to incorporate *mixing* into SimCLR. We first define a new set $\{x_k, \tilde{x}_k\}$ for $k \in [1, 2, ..., N]$, where $\tilde{x}_k$ denotes the synthetically generated counterpart of x_k, thus establishing pairs like x_1 and $\tilde{x}_1$ as positive examples. The contrastive prediction task of the modified, which we term as SimCLR+MixDiff now involves identifying x_1 and $\tilde{x}_1$ in $\{x_k, \tilde{x}_k\} \forall\, k \in [1, N]$. The modified loss function for the mixed version of SimCLR, denoted as $\mathcal{L}_{MixSR}$, for a positive pair of examples (x_i, $\tilde{x_i}$), is defined as:

$$\mathcal{L}^i_{MixSR} \triangleq -\log \frac{\exp(\text{sim}(z_i, \tilde{z_i}))/\tau}{\Sigma_{z \in \{z_j, \tilde{z_j} \forall j \in [1,N]\}} \mathbb{1}_{z \neq z_i, \tilde{z_i}} \exp(\text{sim}(z_i, z))}$$

where for a given two feature vectors, u and v, their 'sim' refers to the cosine similarity and is calculated as $\text{sim}(u, v) = \frac{u^T v}{||u||,||v||}$, representing the dot product of the l_2 normalized vectors u and v. And z_i and $\tilde{z_i}$ are the outputs of the network $f(\cdot)$ we are developing, expressed as:

$$z_i = f(x_i) \;\text{ and }\; \tilde{z_i} = f(\tilde{x_i}) = f(g^k_{SD}(x_i))$$

Barlow Twins + MixDiff: The Barlow Twins framework [81], while maintaining a Siamese network structure similar to SimCLR [13], adopts a distinct approach to representation alignment. The difference lies in the Barlow Twins' objective function, which assesses the cross-correlation matrix between the embeddings from two identical networks. These networks process distorted versions of a batch of samples, with the aim of aligning this matrix closely with the identity matrix. This alignment ensures that the embeddings of distorted versions of a sample are similar, while simultaneously reducing redundancy among the components of these embeddings.

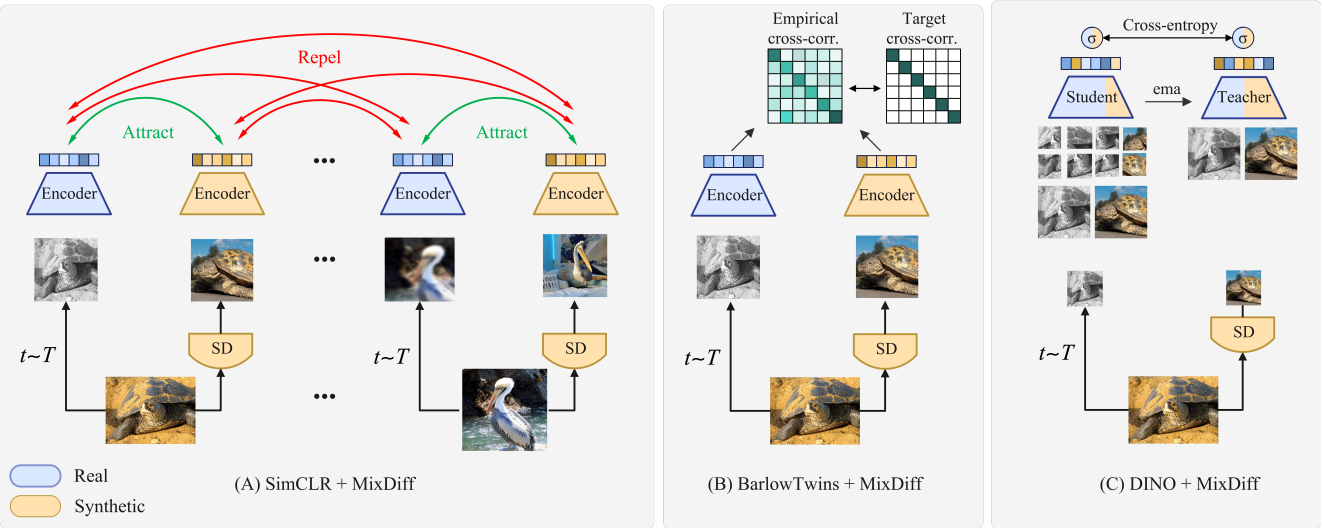

Figure 2. Existing SSL methods, including (A) SimCLR, (B) Barlow Twins, and (C) DINO, have been enhanced with our novel MixDiff approach. In both (A) SimCLR and (B) Barlow Twins, we replace a branch representing the positive pair with a synthetic image generated without the label using Stable Diffusion. This modification enables the learning of real-synthetic view prediction. (C) DINO utilizes a distillation framework with two global views for the teacher and a mix of two global and eight local views for the student. Our adaptation integrates a blend of global and local synthetic and real images facilitating learning correspondences between global-to-local on top of real-to-synthetic image views.

In our modified approach, as we show in Figure 2 (B), we innovate by introducing a synthetic element into this framework. Instead of solely using distorted versions of the same real image, we integrate a distorted version of a synthetic image. Following the notation from the previous section, let z_i represent the distorted version of a real image and $\tilde{z}_j$ that of a synthetic image. The objective function for this adapted version of Barlow Twins, denoted as $\mathcal{L}_{BT}$, is formulated as:

$$\mathcal{L}_{BT} \triangleq \Sigma_i (1 - \mathcal{C}_{ii})^2 + \lambda \Sigma_i \Sigma_{j \neq i} C_{ij}^2$$

Here, λ is a positive constant that balances the first and second terms in the loss function. The cross-correlation matrix, $\mathcal{C}$, is computed between the outputs of the two identical networks, one fed with real images and the other with synthetic images, as follows:

$$\mathcal{C}_{ij} \triangleq \frac{\Sigma_b z_{b,i} \, \tilde{z}_{b,j}}{\sqrt{\Sigma_b (z_{b,i})^2} \sqrt{\Sigma_b (\tilde{z}_{b,j})^2}}$$

where b indexes the batch samples, while i and j index the vector dimensions of the networks' outputs. This updated approach, which integrates synthetic images into the Barlow Twins framework, focuses on aligning the representations of real (z_i) and synthetic images ($\tilde{z}_j$) via the cross-correlation matrix. We give more details regarding mixing in joint-embedding SSL in the appendix B.

3.3. Mixing in Distillation SSL

DINO + MixDiff: In contrast to other SSL methods, DINO [12] uses a multi-crop strategy to create multiple views at different scales, including two high-resolution global views (x_1^g, x_2^g) and multiple lower-resolution local views (x_k^l, $k = 1$ to 8). It employs a knowledge distillation (KD) framework where a student network g_{θ_s} learns to match the output of a teacher network g_{θ_t}, with the student processing both local and global views, while the teacher focuses on global views to enhance 'local-to-global' learning. Building on this foundation, we introduce image mixing in DINO, termed DINO + MixDiff, the model is adapted to integrate both real and synthetic images. This is accomplished by adjusting the view composition to include one global and six local views from a real image, plus one global and two local views from a synthetic image. Consequently, our modified set includes a global view from a real image (x_1^g), a global view from a synthetic image ($\tilde{x}_2^g$), and four local views each from the real ($x_r^l \; \forall \; r \in [1, 6]$) and synthetic ($\tilde{x}_q^l \; \forall \; q \in [1, 2]$) images.

With a student network $g(\theta_s)$ and a fixed teacher network $g\theta_t$ updated via *Exponential Moving Average* (EMA), the learning objective is to align these distributions. This is achieved by minimizing the cross-entropy loss with respect to the student network's parameters θ_s, expressed as: $\min_{\theta_s} H(P_t(X_t), P_s(X_s))$ where $H(a, b) = -a \log b$ denotes the cross-entropy function. Both the student and teacher networks generate probability distributions denoted as P_s and P_t, respectively, derived by normalizing the networks' outputs using a softmax function. This learning process involves passing all crops through the student network, while the teacher network processes only the global

views. This design fosters 'local-to-global' as well as 'real-to-synthetic' learning correspondences. The loss objective becomes:

$$\min_{\theta_s} \; \Sigma_{X_t \in \{x_1^g, \tilde{x}_2^g\}} \; \Sigma_{X_s \in \tilde{V},\; X_s \neq X_t} H(P_t(X_t), P_s(X_s))$$

We provide more details regarding mixing in distillation SSL such as DINO in appendix C.1.

4. Experiments

In this section, we present experiments testing robustness to distribution shifts, domain transfer across datasets, and performance on low-quality images. We provide insights into the learned representations using and without using MixDiff in SSL as described in Section 3.

Training Algorithms and Data. As the proposed solution is a simple change in one data branch of the SSL pipeline, we can easily incorporate it into any existing joint-embedding SSL methods. The substantial size of the ImageNet-1K (IN-1K) [18] dataset, which contains approximately 1.3 million images, presents challenges for extensive experimentation. Consequently, we primarily utilize the more manageable ImageNet-100 (IN-100) dataset [63] for our studies, which include 100 classes and 1300 images per class. This dataset's smaller scale enables us to efficiently run multiple variations of each synthetic dataset and thoroughly evaluate the impact of various design choices. Nonetheless, we extend our experiments to IN-1K with SimCLR to validate our findings on a larger scale.

We trained variants of DINO, SimCLR, and Barlow Twins models using only real, synthetic, and our proposed mixed version of both image types (MixDiff) on the IN-100 dataset. For classification using the pre-trained features, unless stated, we always train the linear probes on the training set of real images. For example, the models trained with synthetic IN-100 images use the training set from real IN-100 images to train the linear probes to evaluate them on the IN-100 test set and other datasets. We leverage FFCV [39] to accelerate training. We provide further details regarding configurations in appendix D.1.

4.1. MixDiff boosts robustness to distribution shifts

To evaluate performance under domain shifts, we choose a set of four datasets including ImageNet-A (IN-A) [32], ImageNet-Sketch (IN-Sketch) [69], ObjectNet [8], ImageNet-V2 (IN-V2) [53], VizWiz-Classification (VW-C) [5], and ImageNet-R (IN-R) [31].

Figure 3 shows the average accuracy on the four distribution shift datasets (excluding ObjectNet and WV-C due to the lack of common objects with IN-100) and compares it to the in-distribution IN-100. Models utilizing MixDiff (green) demonstrate superior accuracy on both IN-100

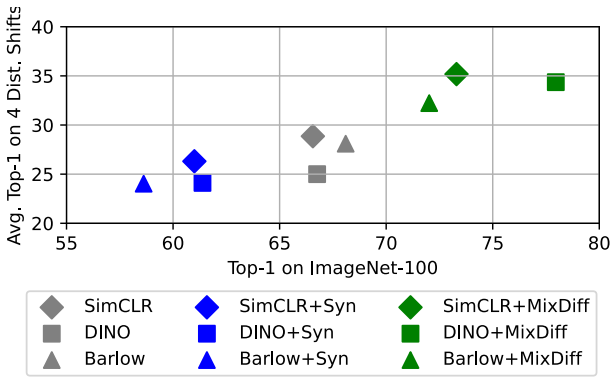

Figure 3. Top-1 classification accuracies (%) for various models on ImageNet-100 (x-axis) and the average of four domain shift datasets (y-axis). This figure compares the performance of models trained on real, synthetic (Syn), and an equal combination of real and synthetic images (MixDiff). Models in the top-right quadrant exhibit better in-distribution and out-of-distribution accuracies.

and the distribution shift datasets on average, outperforming models trained exclusively on real (grey) or synthetic (blue) images. This suggests that MixDiff not only enhances in-distribution performance but also enhances robustness against distribution shifts. SimCLR+MixDiff is the most robust, while DINO+MixDiff excels in in-distribution accuracy. In Table 1, we drill down on the SimCLR variants using IN-1K dataset, and these findings align with our previous observations from the IN-100 dataset. While DINO outperforms on IN-100, we chose SimCLR for this experiment owing to its markedly quicker training time, attributed to its use of fewer crops and simpler overall setup. When pre-trained on ImageNet-1K, we observe similar improvements using mixing in SimCLR (SimCLR+MixDiff). Additionally, The performance boost is consistent across the four distribution-shift datasets.

SimCLR trained on synthetic images (SimCLR+Syn) shows greater robustness on datasets like IN-Sketch and IN-R compared to SimCLR trained on real images. This improvement can be attributed to the datasets exhibiting properties similar to synthetic images. For instance, IN-Sketch contains human-drawn sketches, and IN-R includes artistic and stylized object renditions. These characteristics align well with the diverse and sometimes abstracted nature of the synthetic images. We believe MixDiff's effectiveness is due to synthetic images acting as hard positive samples. It is more challenging to bring generated images closer than to bring augmented samples closer. These hard positive samples prevent the model from learning trivial features, which enhances its ability to learn effective representations [55, 74]. However, it is crucial to highlight that accuracy on more challenging datasets like ImageNet-A is still low [20, 64]. This may be attributed to the backbone of the SimCLR model being ResNet-50. While ImageNet-A was curated specifically as images that fool the ResNet-50

Model	IN-1K	Distribution Shifts Datasets							Mean
		IN-V2	IN-Sketch	IN-A	IN-R	ObjectNet	VW-C	Mean	
SimCLR	63.34	50.10	14.08	1.64	23.71	14.64	24.96	21.52	27.92
SimCLR+Syn	57.58	44.70	16.19	1.72	26.12	11.35	23.25	20.55	26.28
SimCLR+MixDiff	**67.90**	**54.53**	**22.57**	**2.22**	**34.97**	**19.65**	**29.93**	**27.31**	**33.64**

Table 1. Top-1 classification accuracies (%) of models trained on ImageNet-1K, evaluated on domain shift datasets. SimCLR+Syn improves accuracy on IN-R and IN-Sketch, and this robustness extends to SimCLR+MixDiff. MixDiff enhances both in-distribution accuracy and out-of-distribution robustness.

Model	CIFAR-10	CIFAR-100	Aircraft	DTD	Flowers	Food	Pets37	STL10	Mean
SimCLR	79.75	55.49	30.48	**62.71**	81.27	64.80	67.51	93.75	66.97
SimCLR+Syn	77.67	53.74	29.82	58.83	82.16	60.91	67.92	92.02	65.38
SimCLR+MixDiff	**84.76**	**64.33**	**36.84**	**62.71**	**88.43**	**66.40**	**76.45**	**95.80**	**71.90**

Table 2. Comparison of transfer learning performance on eight diverse datasets for models trained on ImageNet-1K. MixDiff outperforms all other models, demonstrating superior generalization across these datasets

model [32]. More information and detailed numerical results are available in the appendix D.2.

4.2. MixDiff improves transfer learning

We evaluate the feature generality of the models by conducting transfer learning experiments across various image datasets and compare MixDiff's effectiveness with other models. The datasets include: Aircraft [43], DTD [16], Flowers102 [47], Food101 [10], Pets37 [49], STL10 [17], CIFAR-10 [37], and CIFAR-100 [37].

We pre-train the SimCLR model on the IN-1K dataset using the three data variants: real images, synthetic images, and MixDiff. For each variant, after pre-training the SimCLR backbone, we subsequently froze these layers to train linear probes on real data from the domains listed above. Table 2 presents the top-1 accuracy results for each dataset. Our approach consistently achieves higher top-1 accuracy across all datasets compared to models trained solely on real or synthetic images. We would like to point out that our training iterations are lower (100 epochs) than the original SimCLR (1000 epochs), resulting in slightly lower numbers, but our method still consistently outperforms the original SimCLR model in relative terms. The superior transferability of representations learned using MixDiff can be attributed to its exposure to a broader range of visual inputs. By combining real and synthetic images, MixDiff allows the model to learn from a more diverse set of visual features and variations. This expanded visual vocabulary likely contributes to the development of more robust and generalizable representations, which in turn transfer more effectively to other datasets.

4.3. MixDiff learns more from limited data

We explore MixDiff's effectiveness in limited data scenarios for SSL. We trained SimCLR models using both real images and MixDiff on 25%, 50%, and 100% subsets of the IN-100 dataset, supplemented with proportional synthetic images generated at a guidance scale of 8. Linear probes were subsequently trained on the real images of each subset. As shown in Figure 4, we observe noticeable performance gains across different data-size regimes. As an example, using 25% of the images, MixDiff achieves a 5.14% increase in accuracy. This trend extends well to robustness, where a model trained with MixDiff on 50% of the data matches the accuracy on distribution shift datasets of a model trained on 100% real images.

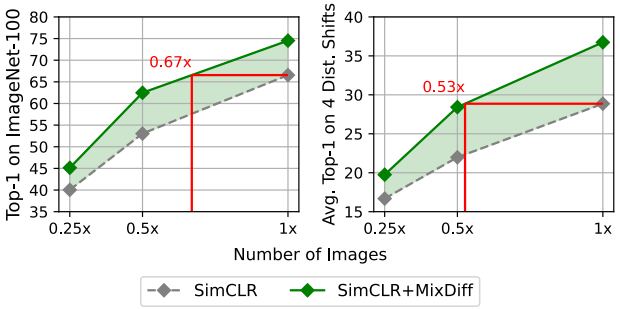

Figure 4. **Left:** Top-1 accuracy on IN-100 for SimCLR models trained with and without MixDiff at different scales of training images. **Right:** Average top-1 accuracy on four distribution shift datasets. SimCLR+MixDiff outperforms SimCLR, as indicated by the green area showing the performance gap.

Notably, as indicated by the red line in Figure 4, MixDiff achieves comparable in-distribution accuracy and robustness to SimCLR trained on 100% real images, while using only 53% and 67% of the real data, respectively. These findings demonstrate that MixDiff not only enhances performance but also enables robust training with reduced data requirements. It is important to note that while MixDiff offers significant efficiency gains in terms of data usage, it does require a one-time computational investment for gen-

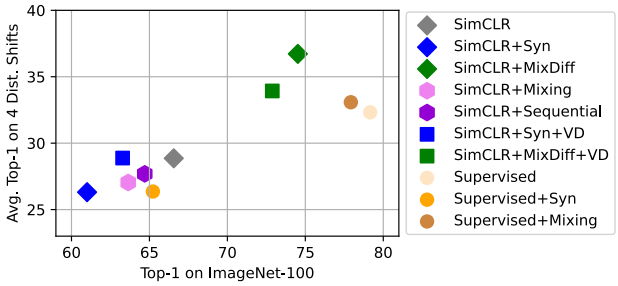

Figure 5. Top-1 classification accuracies for various models on ImageNet-100 (x-axis) and the average of four domain shift datasets (y-axis). The models use different approaches for mixing real and synthetic images, as well as varying generation models.

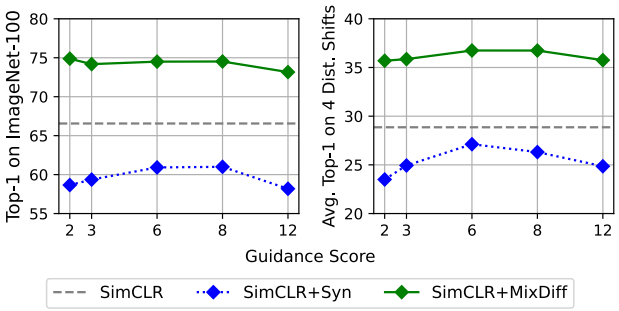

Figure 6. **Left:** Top-1 accuracy on IN-100. **Right:** Average top-1 accuracy on four distribution shift datasets. SimCLR+MixDiff is more robust to changes in guidance scale than SimCLR+Syn.

erating synthetic images. We provide more discussion on the strategies for reducing training times using MixDiff with minimal performance impact in section E.5 in the appendix.

5. Ablation study

Here, we empirically study the properties of generated synthetic images under varying guidance scales, mixing configurations, augmentations and evaluate how these factors affect on MixDiff's performance compared to other SSL methods. We further visualize self-attention maps of DINO over IN-100, and observe that DINO trained with MixDiff tend to attend and segment objects well with much less focus on the background (See section E.1).

5.1. Analysis of different configurations vs. MixDiff

We consider two new setups: training SimCLR on real images for 50 epochs followed by synthetic images for 50 epochs (Sequential), and training on a combined dataset of real and synthetic images for 50 epochs (Mixing), keeping the total number of images the same. Figure 5 presents the performance of these new setups alongside our baseline configurations: SimCLR trained on synthetic images, real images, and our proposed MixDiff pipeline. While the new configurations surpassed SimCLR+Syn, they underperformed compared to SimCLR and SimCLR+MixDiff. We also replicated the MixDiff experiment using images generated by Versatile Diffusion (VD) [77] with a guidance scale of 8, as an alternative to Stable Diffusion (SD) [50]. Results in Figure 5 demonstrate our MixDiff with VD-generated images (SimCLR+MixDiff+VD) outperforms the original SimCLR, indicating its generalizability across a different generative model. Finally, Figure 5 shows the accuracy of a supervised ResNet-50 model trained on real, synthetic, and mixed data for 100 epochs, similar to the training setup of prior works [4, 80]. The model with real and mixed data achieves higher in-distribution accuracy, but SSL models trained with MixDiff are more robust to distribution shifts, which can be attributed to stronger SSL augmentations and learning from synthetic images. We show that data mixing can make the supervised model more robust at the cost of in-distribution accuracy, while MixDiff improves both. MixDiff narrows the gap between SimCLR and the supervised setting. Details on numerical results are in section E.4 of the appendix.

5.2. Impact of varying guidance scales

The guidance scale in generative models plays a crucial role in balancing the diversity and quality of synthesized images, which subsequently affects the learned representations, as demonstrated by [62]. To investigate this phenomenon in our context, we generated images using various guidance scales $\{2, 3, 6, 8, 12\}$ and trained different SimCLR configurations with these images.

Figure 6 illustrates our findings. Consistent with results from [62], we observe that altering guidance scales impacts the robustness and in-distribution accuracy of models trained on synthetic images. This consistency is noteworthy, given our use of different generative models. Interestingly, MixDiff exhibits less variability to different guidance scales, maintaining a consistent performance across most scales, with a slight drop at a guidance scale of 12. This consistency offers a significant computational advantage, as it eliminates the need to fine-tune the guidance scale as a hyperparameter. Furthermore, our results consistently demonstrate that the SimCLR+Syn model underperforms compared to the original SimCLR. This finding indicates the constraints of using only synthetic data for training and suggests the advantages of our combined method. These findings collectively emphasize the importance of carefully considering guidance scale selection in generative models and suggest that our MixDiff method offers a more stable and efficient alternative for using synthetic data in representation learning tasks.

5.3. Properties of generated images

Building on our previous analysis of guidance effects on MixDiff's performance, we now investigate the properties of the generated synthetic images. We use cosine distance

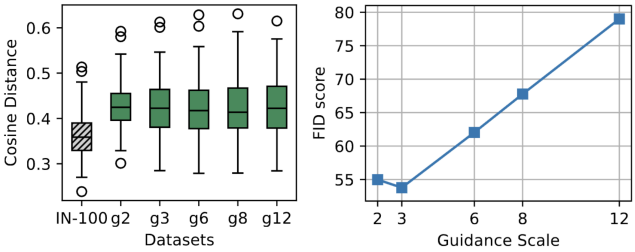

Figure 7. **Left:** Mean cosine distance distribution between feature vectors of IN-100 categories and generated images with varying guidance scales. **Right:** Relationship between guidance scales and FID scores. Generated images are more diverse than real images, and higher guidance scales reduce image quality.

to measure diversity and Frechet Inception Distance (FID) scores [33] to assess image quality in comparison to IN-100 across different guidance scales. The left panel of Figure 7 reveals that generated images exhibit higher diversity than real images, likely due to the generative model's inherent randomness. The right panel illustrates the similarity in quality between generated images and IN-100 source images across guidance scales. Drawing from the observations in previous section, we observe an inverse relationship between the SimCLR+Syn model's accuracy and image quality (as indicated by FID). For instance, a guidance scale of 8 yielded better performance despite a higher FID compared to a scale of 3 with the lowest FID. This aligns with findings from [62], suggesting that lower-quality generated images with higher FID scores may be more beneficial for learning solely from synthetic data. Interestingly, MixDiff's performance showed no clear trend relative to the FID scores of synthetic data. This robustness across guidance scales suggests that MixDiff may be effectively combining information from both real and synthetic sources, potentially offsetting image quality variations. Further research is needed to understand this phenomenon and its implications for using synthetic data in representation learning.

5.4. Impact of data augmentation

Our study shows the effect of data augmentation on the SimCLR model, comparing its performance when trained with MixDiff versus solely real images of IN-100. Specifically, for models trained without augmentation, we retain only the random flip and remove all other augmentations.

As depicted in Figure 8, an interesting finding is that omitting data augmentations, which is a key component in self-supervised learning models, does not significantly affect the performance of models trained with both real and synthetic images. Surprisingly, MixDiff model without augmentation generally outperforms the original SimCLR, on all except for the ImageNet-Sketch dataset. Conversely, SimCLR models trained exclusively on real images experience a significant drop in performance due to the absence

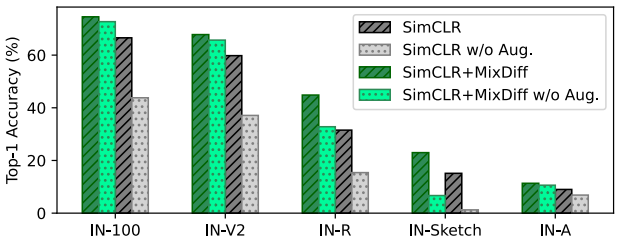

Figure 8. Top-1 accuracy comparison of SimCLR models trained with and without MixDiff, and the impact of including versus omitting SSL data augmentations.

of data augmentations. Specifically, the average accuracy drop across five validation datasets is 15.52% for SimCLR without augmentation, compared to a lesser accuracy drop of 6.60% for SimCLR+MixDiff. Notably, removing these augmentations can lead to faster training times, which is significant considering that data augmentations are often a major bottleneck in the training process, as highlighted in [9]. Also, we observed that, for SimCLR+MixDiff, jittering is the most effective augmentation for in-distribution and robustness performance (See section E.3).

6. Conclusion

This work proposes MixDiff, a framework that demonstrates the potential of mixing synthetic images generated without any labels, with real images for fully self-supervised pre-training. We show that MixDiff consistently enhances the performance of existing SSL techniques across various benchmarks. Notably, MixDiff exhibits reduced variability to synthetic image quality and requires a smaller quantity of real data to achieve performance comparable to models trained on real data, thus offering a more robust and efficient SSL pre-training mechanism. More importantly, these results suggest that the integration of synthetic data with real images may serve as a viable alternative to augmentation techniques in existing SSL methods.

Further investigation reveals a noticeable performance gap between models trained solely on synthetic images and those trained on real images. This indicates a significant unexplored potential in the development of generative models to improve efficiency and produce images that more closely mimic the distribution of real images. In future work, we aim to investigate advanced generative diffusion models that could facilitate better mixing and enhance the robustness of visual features.

Acknowledgements. This work utilized the Alpine high performance computing resource at the University of Colorado Boulder. Alpine is jointly funded by the University of Colorado Boulder, the University of Colorado Anschutz, Colorado State University, and the National Science Foundation (Award 2201538).

References

[1] Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. 2

[2] Dosovitskiy Alexey, Philipp Fischer, Jost Tobias, Martin Riedmiller Springenberg, and Thomas Brox. Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE TPAMI*, 38(9):1734–1747, 2016. 3

[3] 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. 2, 13

[4] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J. Fleet. Synthetic data from diffusion models improves imagenet classification. *ArXiv*, abs/2304.08466, 2023. 2, 7, 14

[5] Reza Akbarian Bafghi and Danna Gurari. A new dataset based on images taken by blind people for testing the robustness of image classification models trained for imagenet categories. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16261–16270, 2023. 5, 15

[6] Reza Akbarian Bafghi, Nidhin Harilal, Claire Monteleoni, and Maziar Raissi. Parameter efficient fine-tuning of self-supervised vits without catastrophic forgetting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3679–3684, 2024. 2

[7] Hritik Bansal and Aditya Grover. Leaving reality to imagination: Robust classification via generated datasets. *ArXiv*, abs/2302.02503, 2023. 2

[8] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Joshua B. Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Neural Information Processing Systems*, 2019. 5

[9] Florian Bordes, Randall Balestriero, and Pascal Vincent. Towards democratizing joint-embedding self-supervised learning. *ArXiv*, abs/2303.01986, 2023. 8, 14

[10] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 - mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014. 6

[11] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *ArXiv*, abs/1809.11096, 2018. 2

[12] 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, 4, 13, 14, 15

[13] 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, 3, 13

[14] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021. 2, 13

[15] Yuhua Chen, Wen Li, Xiaoran Chen, and Luc Van Gool. Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1841–1850, 2018. 2

[16] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3606–3613, 2013. 6

[17] Adam Coates, A. Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *International Conference on Artificial Intelligence and Statistics*, 2011. 6

[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, K. Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 5

[19] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. *ArXiv*, abs/2105.05233, 2021. 3

[20] Josip Djolonga, Jessica Yung, Michael Tschannen, Rob Romijnders, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Matthias Minderer, Alexander D'Amour, Dan I. Moldovan, Sylvain Gelly, Neil Houlsby, Xiaohua Zhai, and Mario Lucic. On robustness and transferability of convolutional neural networks. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16453–16463, 2020. 5

[21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2020. 15

[22] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. *Advances in neural information processing systems*, 27, 2014. 2, 13

[23] Lisa Dunlap, Alyssa Umino, Han Zhang, Jiezhi Yang, Joseph E. Gonzalez, and Trevor Darrell. Diversify your vision datasets with automatic diffusion-based augmentation. *ArXiv*, abs/2305.16289, 2023. 2

[24] Lijie Fan, Kaifeng Chen, Dilip Krishnan, Dina Katabi, Phillip Isola, and Yonglong Tian. Scaling laws of synthetic images for model training... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7382–7392, 2024. 2

[25] 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, 13

[26] Jie Gui, Tuo Chen, Jing Zhang, Qiong Cao, Zhenan Sun, Hao Luo, and Dacheng Tao. A survey on self-supervised learning: Algorithms, applications, and future trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1

[27] Xi Guo, Wei Wu, Dongliang Wang, Jing Su, Haisheng Su, Weihao Gan, Jian Huang, and Qin Yang. Learning video representations of human motion from synthetic data. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20165–20175, 2022. 2

[28] Hasan Abed Al Kader Hammoud, Hani Itani, Fabio Pizzati, Philip Torr, Adel Bibi, and Bernard Ghanem. Synthclip: Are we ready for a fully synthetic clip training? *arXiv preprint arXiv:2402.01832*, 2024. 2

[29] 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, 2, 3, 13

[30] Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip H. S. Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition? *ArXiv*, abs/2210.07574, 2022. 2

[31] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Lixuan Zhu, Samyak Parajuli, Mike Guo, Dawn Xiaodong Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8320–8329, 2020. 5

[32] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Xiaodong Song. Natural adversarial examples. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15257–15266, 2019. 5, 6

[33] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Neural Information Processing Systems*, 2017. 8

[34] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2, 13

[35] Jonathan Ho, Chitwan Saharia, William Chan, David J. Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.*, 23:47:1–47:33, 2021. 3

[36] Mehran Jeelani, Noshaba Cheema, Klaus Illgner-Fehns, Philipp Slusallek, Sunil Jaiswal, et al. Expanding synthetic real-world degradations for blind video super resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1199–1208, 2023. 1

[37] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009. 6

[38] Phuc H Le-Khac, Graham Healy, and Alan F Smeaton. Contrastive representation learning: A framework and review. *Ieee Access*, 8:193907–193934, 2020. 1

[39] Guillaume Leclerc, Andrew Ilyas, Logan Engstrom, Sung Min Park, Hadi Salman, and Aleksander Madry. FFCV: Accelerating training by removing data bottlenecks. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. https://github.com/libffcv/ffcv/. commit 6c3be0cabf1485aa2b6945769dbd1c2d12e8faa7. 5, 14

[40] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv: Learning*, 2016. 14

[41] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017. 14

[42] Jianxin Ma, Shuai Bai, and Chang Zhou. Pretrained diffusion models for unified human motion synthesis. *ArXiv*, abs/2212.02837, 2022. 2

[43] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew B. Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *ArXiv*, abs/1306.5151, 2013. 6

[44] Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. Generating training data with language models: Towards zero-shot language understanding. *ArXiv*, abs/2202.04538, 2022. 2

[45] Masato Mimura, Sei Ueno, Hirofumi Inaguma, Shinsuke Sakai, and Tatsuya Kawahara. Leveraging sequence-to-sequence speech synthesis for enhancing acoustic-to-word speech recognition. *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 477–484, 2018. 2

[46] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6707–6717, 2020. 2

[47] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729, 2008. 6

[48] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2, 13

[49] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3498–3505, 2012. 6

[50] Justin Pinkney. Stable diffusion image variations model card. https://huggingface.co/lambdalabs/sd-image-variations-diffusers, 2024. Accessed: 2024-07-04. 3, 7

[51] 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 *International Conference on Machine Learning*, 2021. 3

[52] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *ArXiv*, abs/2102.12092, 2021. 3

[53] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, 2019. 5

[54] Pitchaporn Rewatbowornwong, Nontawat Tritrong, and Supasorn Suwajanakorn. Repurposing gans for one-shot semantic part segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:5114–5125, 2021. 2

[55] Joshua Robinson, Li Sun, Ke Yu, Kayhan Batmanghelich, Stefanie Jegelka, and Suvrit Sra. Can contrastive learning avoid shortcut solutions? *Advances in neural information processing systems*, 34:4974–4986, 2021. 2, 5

[56] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2021. 2, 3

[57] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, Seyedeh Sara Mahdavi, Raphael Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *ArXiv*, abs/2205.11487, 2022. 3

[58] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:4713–4726, 2021. 3

[59] Mert Bulent Sariyildiz, Alahari Karteek, Diane Larlus, and Yannis Kalantidis. Fake it till you make it: Learning transferable representations from synthetic imagenet clones. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8011–8021, 2022. 2, 14

[60] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy P. Lillicrap, Fan Hui, L. Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of go without human knowledge. *Nature*, 550:354–359, 2017. 2

[61] Deqing Sun, Daniel Vlasic, Charles Herrmann, V. Jampani, Michael Krainin, Huiwen Chang, Ramin Zabih, William T. Freeman, and Ce Liu. Autoflow: Learning a better training set for optical flow. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10088–10097, 2021. 2

[62] Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. *ArXiv*, abs/2306.00984, 2023. 1, 2, 7, 8

[63] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European Conference on Computer Vision*, 2019. 5

[64] Nenad Tomasev, Ioana Bica, Brian McWilliams, Lars Holger Buesing, Razvan Pascanu, Charles Blundell, and Jovana Mitrovic. Pushing the limits of self-supervised resnets: Can we outperform supervised learning without labels on imagenet? In *First Workshop on Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022*, 2022. 5

[65] Brandon Trabucco, Kyle Doherty, Max Gurinas, and Ruslan Salakhutdinov. Effective data augmentation with diffusion models. *arXiv preprint arXiv:2302.07944*, 2023. 2

[66] Jonathan Tremblay, Aayush Prakash, David Acuna, Mark Brophy, Varun Jampani, Cem Anil, Thang To, Eric Cameracci, Shaad Boochoon, and Stan Birchfield. Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 969–977, 2018. 1

[67] Bram Vanherle, Steven Moonen, Frank Van Reeth, and Nick Michiels. Analysis of training object detection models with synthetic data. *arXiv preprint arXiv:2211.16066*, 2022. 1

[68] Veniamin Veselovsky, Manoel Horta Ribeiro, Akhil Arora, Martin Josifoski, Ashton Anderson, and Robert West. Generating faithful synthetic data with large language models: A case study in computational social science. *ArXiv*, abs/2305.15041, 2023. 2

[69] Haohan Wang, Songwei Ge, Eric P. Xing, and Zachary Chase Lipton. Learning robust global representations by penalizing local predictive power. In *Neural Information Processing Systems*, 2019. 5

[70] Yifei Wang, Jizhe Zhang, and Yisen Wang. Do generated data always help contrastive learning? *ArXiv*, abs/2403.12448, 2024. 1, 2

[71] Zekai Wang, Tianyu Pang, Chao Du, Min Lin, Weiwei Liu, and Shuicheng Yan. Better diffusion models further improve adversarial training. *ArXiv*, abs/2302.04638, 2023. 2

[72] Zhicai Wang, Longhui Wei, Tan Wang, Heyu Chen, Yanbin Hao, Xiang Wang, Xiangnan He, and Qi Tian. Enhance image classification via inter-class image mixup with diffusion model. *ArXiv*, abs/2403.19600, 2024. 2

[73] Yo whan Kim, SouYoung Jin, Rameswar Panda, Hilde Kuehne, Leonid Karlinsky, Samarth Mishra, Venkatesh Saligrama, Kate Saenko, Aude Oliva, and Rogério Schmidt Feris. How transferable are video representations based on synthetic data? In *Neural Information Processing Systems*, 2022. 2

[74] Yawen Wu, Zhepeng Wang, Dewen Zeng, Yiyu Shi, and Jingtong Hu. Synthetic data can also teach: Synthesizing effective data for unsupervised visual representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2866–2874, 2023. 1, 2, 5

[75] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018. 3

[76] Guodong Xu, Ziwei Liu, Xiaoxiao Li, and Chen Change Loy. Knowledge distillation meets self-supervision. In *European conference on computer vision*, pages 588–604. Springer, 2020. 1

[77] Xingqian Xu, Zhangyang Wang, Eric Zhang, Kai Wang, and Humphrey Shi. Versatile diffusion: Text, images and variations all in one diffusion model. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7720–7731, 2022. 2, 7, 15

[78] Mang Ye, Xu Zhang, Pong C Yuen, and Shih-Fu Chang. Unsupervised embedding learning via invariant and spreading instance feature. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6210–6219, 2019. 2

[79] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv: Computer Vision and Pattern Recognition*, 2017. 14

[80] Zhuoran Yu, Chenchen Zhu, Sean Chang Culatana, Raghuraman Krishnamoorthi, Fanyi Xiao, and Yong Jae Lee. Diversify, don't fine-tune: Scaling up visual recognition training with synthetic images. *ArXiv*, abs/2312.02253, 2023. 7

[81] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021. 1, 2, 3, 13