SS-GEAR: Improving Self-Supervised Learning with - Generative Augmentation for Representation Enhancement

Aaditya Baranwal
IIT Jodhpur
baranwal.1@iitj.ac.in

Mohit Mathuria
IIT Jodhpur
mathuria.1@iitj.ac.in

Nakul Sharma
IIT Jodhpur
sharma.86@iitj.ac.in

Abstract

This paper addresses the imperative for increased data in emerging self-supervised methodologies, highlighting the potential benefits of advancing the field of artificial intelligence (AI). The focus is on exploring generative data augmentation to augment datasets and improve the performance of self-supervised models. The report provides a detailed analysis of existing approaches, shedding light on the current landscape of generative augmentation. Additionally, it delves into related work, establishing the context for the proposed methodology.

1. Introduction

Self-Supervised Learning has emerged as a promising technique for learning powerful visual representations from unlabeled data. By defining pretext tasks such as image rotation prediction or contrastive learning, self-supervised methods can learn representations that transfer well to downstream tasks.

However, the performance of self-supervised learning algorithms heavily depends on the amount and diversity of the unlabeled dataset. Recent models are routinely pre-trained on dataset sizes in the millions or even billions of images, which is costly to acquire and store. Therefore, developing techniques to augment existing datasets could substantially improve self-supervised learning.

This report proposes a novel generative data augmentation [11] approach to enhance self-supervised visual representations by synthesizing additional training data. We first discuss the motivation behind this work and how it can advance artificial intelligence. We then review existing generative data augmentation techniques and discuss their limitations. Finally, we describe the details of the proposed methodology and related work.

1.1. Motivation

While self-supervised learning has shown promise, state-of-the-art algorithms require massive pre-training datasets, which presents data acquisition, storage, and transmission challenges. By effectively augmenting existing datasets with synthetic samples, the proposed generative augmentation approach can help address these challenges and make self-supervised learning more practical.

Additionally, intelligently augmenting the training distribution can make self-supervised models more robust and invariant to nuisance variables. This could make the learned representations more generalizable and performant on downstream tasks.

This generative augmentation approach could lower the data requirements for pre-training self-supervised models. Democratizing access to powerful unsupervised representations could greatly benefit researchers and practitioners with limited computational resources. This could advance artificial intelligence applications for social good where data collection is challenging.

1.2. Objective

We propose our work with the following broad objectives in mind.

Explore the necessity for more data in self-supervised methodologies. Self-supervised approaches hold immense promise for learning transferable visual representations from unlabeled data. However, state-of-the-art algorithms require massive datasets, often in the hundreds of millions of images, to realize their full potential. We aim to motivate the critical need for techniques to augment limited self-supervised learning datasets. Key questions include: How does dataset scale drive recent progress? Can we reduce this dependence through impactful augmentation? Discuss the potential impact of increased data on advancing AI applications.

By enabling robust self-supervised learning from fewer samples, our work can greatly expand the reach of mod-

ern machine perception. We will highlight opportunities around democratization and applications such as medical imaging, satellite analysis, robotics, and more, where collecting supervision is challenging, but self-supervised approaches could enable solutions. The discussion will center on how progress in data-efficient self-supervised learning can serve inclusion and social good. Investigate existing approaches in generative data augmentation.

While conventional augmentation is well-explored, complex augmentation based on modern generative models remains relatively nascent. We will provide a comprehensive background on pioneering work that harnessed adversarial networks and other generative models to synthesize additional training data across problem domains. The discussion will provide context around the limitations of prior art, laying the foundation for the proposed innovations.

Provide a comprehensive overview of related work in the domain. Via an exhaustive literature review, we will highlight connections between our core research areas of self-supervised representation learning, data augmentation, and generative modelling. The goal is to provide a holistic view of precedent across these interconnected domains while positioning our work at this essential intersection with the potential to advance the collective field.

2. Related Work

While data augmentation has become standard practice in supervised learning [4], its exploration in self-supervised representation learning has been more limited. Simple augmentation techniques like cropping and flipping have been integrated into existing self-supervised algorithms [3], but more complex generative augmentation remains relatively unexplored.

Initial attempts at generative self-supervised augmentation have demonstrated some promise but also limitations. [6] proposed using an adversarial image translation model before feeding data to the self-supervised network. However, directly transferring samples from other datasets can introduce artifacts and fail to capture intra-dataset variability critical for generalization.

Concurrent work by [12] trains a GAN alongside the self-supervised model to introduce novel samples without explicitly optimizing the generator for beneficial augmentations. As a result, variability in representations learned on real vs synthetic samples is not guaranteed.

A common limitation of these preliminary approaches is that the generator objective does not consider the impact on representation quality. As such, they produce samples aimed at general realism rather than tailored augmentation. This motivates our proposed technique of directly incorporating the representation loss of synthetic images into the generator's training procedure.

We can produce a generator that augments more meaningfully by explicitly encouraging synthetic samples that lead to better representations of the real data. Coupling the adversarial and self-supervised objectives also enables co-evolution, where the generator synthesizes increasingly useful data distributions as the representation model gains competency. We hypothesize that this targeted augmentation approach will learn superior representations compared to standard and prior generative augmentation strategies.

3. Approaches

3.1. GAN + SSL

For our purposes, we choose SimCLR for SSL and styleGANv2 [5] / cDCGAN for GAN.

3.1.1 Objective Function

Our objective is to seamlessly integrate the strengths of generative adversarial networks (GANs) and self-supervised learning (SSL) to create a unified training framework that optimizes the generator for both realism and meaningful augmentation. The overall objective function is formulated as a min-max optimization problem, seeking to train the generator G and discriminator D through adversarial training while concurrently leveraging SSL for representation learning. This novel coupling aims to produce synthetic samples that fool the discriminator and enhance the self-supervised model's representation quality.

$$\mathcal{L}_{objective} = \phi * \min_G \max_D \mathcal{L}_{\text{GAN}} + \mathcal{L}_{\text{SSL}}$$

where

$$\mathcal{L}_{\text{GAN}} = \mathbb{E}_{\mathbf{z}\sim\mathcal{P}_{\mathbf{z}},\mathbf{c}\sim\mathcal{P}_{\mathbf{c}}} \log D(G(\mathbf{z},\mathbf{c}))$$

$$L_{\text{SSL}} = -\log \frac{\exp\left(\frac{z_i \cdot z_j}{\tau}\right)}{\sum_{k=1}^{2N} \mathbb{1}_{[k\neq i]} \exp\left(\frac{z_i \cdot z_k}{\tau}\right)}$$

If we have a controlling condition "c":

$$\phi = \mathbb{E}_{\mathbf{z}\sim\mathcal{P}_{\mathbf{z}},\mathbf{c}\sim\mathcal{P}_{\mathbf{c}},\mathbf{x}\sim\mathcal{D}_{\text{s}}} \mathcal{L}_{\text{SSL}}(G(\mathbf{z},\mathbf{c}),\mathbf{x})$$

Else:

$$\phi(t) = f(\frac{t}{T})$$

Here, the GAN loss $\mathcal{L}_{\text{GAN}}$ encourages the generator to produce samples indistinguishable from real data, fostering realism. Concurrently, the SSL loss $\mathcal{L}_{\text{SSL}}$ evaluates the quality of synthetic samples by assessing their ability to fool the SSL model. The hyperparameter λ modulates the importance of SSL during training, ensuring a gradual transition from purely adversarial to SSL-focused learning.

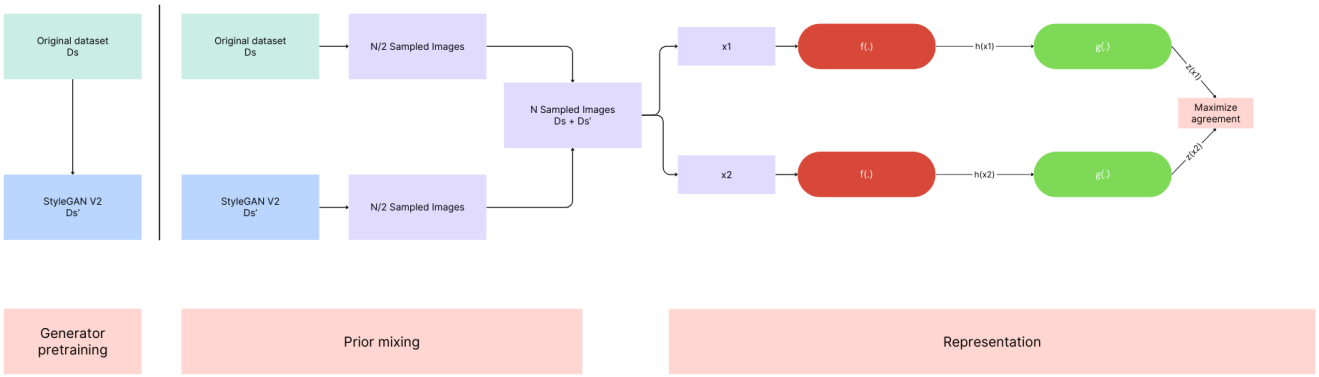

Figure 1. Schematic of GAN + SSL

The rationale behind this combined objective is twofold: (1) to train the generator to produce diverse and realistic samples, and (2) to optimize the generator specifically for augmentation by considering its impact on SSL representation learning. This integrated approach facilitates the co-evolution of the generator and the self-supervised model, resulting in synthetic samples that not only deceive the discriminator but also contribute meaningfully to the downstream representation learning task.

This dual optimization is inspired by the intuition that generating samples beneficial for SSL should inherently lead to more effective data augmentation, ultimately improving the generalization capabilities of the self-supervised model on downstream tasks. The curriculum learning strategy for λ further refines this process, allowing the generator to adapt progressively to the SSL model's growing competence over training. https://www.overleaf.com/project/656f9b3bd14b69259eb7a9a1

3.1.2 Implementation Challenges and Resolutions

Challenge 1: Balancing Objectives

Resolution: To address the challenge of balancing adversarial and self-supervised objectives, we employ curriculum learning for λ, gradually increasing its value during training. This approach is inspired by the insight that, as the generator becomes more adept at producing realistic samples, emphasizing SSL learning becomes increasingly beneficial. The curriculum learning is defined as follows:

$$\phi(t) = f(\frac{t}{T}) \quad \text{where} \quad t \in [0, T]$$

Here, t represents the current training step, and T is the total number of training steps. This adaptive change ensures a smooth transition, allowing the generator to adapt to the growing competency of the SSL model.

Challenge 2: Stability during Training

Resolution: To ensure stable training, we introduce gradient penalties and intermediate supervision. The gradient penalty prevents mode collapse by penalizing large gradients in the discriminator. The Wasserstein [1] GAN objective with gradient penalty is formulated as:

$$\mathcal{L}_{\text{WGAN-GP}} = \mathcal{L}_{\text{GAN}} + \lambda_{\text{GP}}\mathcal{L}_{\text{GP}}$$

Where λ_{GP} controls the strength of the penalty term:

$$\mathcal{L}_{\text{GP}} = \mathbb{E}_{\hat{\mathbf{x}}\sim\mathcal{D}_{\text{mix}}}[(\|\nabla_{\hat{\mathbf{x}}} D(\hat{\mathbf{x}})\|_2 - 1)^2]$$

This penalizes the discriminator when the gradients deviate from a unit norm.

Intermediate supervision involves introducing auxiliary heads in the generator for SSL loss computation at different layers. This helps stabilize training by providing learning signals at multiple stages.

Challenge 3: Sample Diversity

Resolution: To ensure diversity in the generated samples, we design the SSL loss to consider both real and synthetic samples. The SSL loss encourages the model to recognize augmented instances as positive pairs. This formulation is based on the contrastive loss, defined as:

$$L_{\text{SSL}}(i,j) = -\log \frac{\exp(\text{sim}(\mathbf{a},\mathbf{b})/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k\neq i]} exp(\text{sim}(\mathbf{a},\mathbf{b})/\tau)}$$

Where $\text{sim}(\mathbf{a},\mathbf{b})$ measures the similarity between feature vectors $\mathbf{a}$ and $\mathbf{b}$, and τ is the temperature parameter controlling the scale of similarity.

This formulation ensures that the SSL loss encourages representations to be close for positive pairs while effectively distinguishing them from negative instances, promoting diversity in the learned features.

3.1.3 Experiments

Dataset and Setup:

Our experiments aim to evaluate the proposed approach on two benchmark datasets: Tiny ImageNet and CIFAR-10. We choose these datasets for their diverse image content and varying degrees of complexity. The setup involves comparing our method against established self-supervised algorithms, including SimCLR [3], BYOL, and SwAV [2]. We also include baselines using standard augmentation techniques and recent generative approaches.

Metrics:

We employ a comprehensive set of metrics to assess the performance of our approach:

1. **Linear Probing:** - Linear evaluation on frozen features to measure the quality of learned representations.

$$\text{Accuracy}_{\text{linear}} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\text{argmax}(f(\mathbf{x}_i)) = y_i)$$

2. **Fine-Tuning:** - Fine-tuning on downstream tasks to evaluate transferability and practical utility.

3. **Mean Average Precision (mAP):** - Assessing invariance and discrimination capabilities.

$$\text{mAP} = \frac{1}{C}\sum_{c=1}^{C}\frac{\sum_{i=1}^{N_c}\text{AP}_i}{N_c}$$

Here, N is the total number of samples, y_i is the ground truth label, $f(\mathbf{x}_i)$ is the learned representation, C is the number of classes, and AP_i is the average precision for class i.

Anticipated Results:

We hypothesize that our approach will demonstrate superior performance compared to baselines, particularly in scenarios with limited labeled data. The anticipated results include:

1. **Improved Representation Quality:** - Higher accuracy in linear probing, indicating improved feature representations.

2. **Enhanced Transfer Learning:** - Improved fine-tuning performance on downstream tasks, showcasing the practical utility of the learned representations.

3. **Increased Mean Average Precision:** - Higher mAP values, indicating superior invariance and discrimination capabilities.

3.2. SimCLR + Diffusion

The proposed methodology integrates GIT (Generative Image2Text Transformer) [9], Stable Diffusion [8], and SimCLR-inspired self-supervised learning methodology into a cohesive pipeline for representation learning from weak generative image augmentation. An additional step before GIT, for super-resolving images through ESRGAN [10], was deemed necessary for better captioning of the Tiny ImageNet200 dataset. This section details the sequential steps involved in the overall pipeline, emphasizing the holistic approach that combines image-to-text conversion, diffusion-based augmentation, and contrastive learning.

3.2.1 Image-to-Text Conversion using GIT Transformer:

For each batch of sample images I, we first upscale them using ESRGAN (I_{Scaled}) then the GIT is employed to convert these up-scaled images into textual descriptions or captions (S_{cap}). The GIT model, a generative image-to-text transformer, is pre-trained on a combination of image-text datasets to generate text conditioned on images and fine-tuned to provide contextualized captions for the MS-COCO [7] images. Our initial experiments with using the GIT model with TinyImageNet failed due to the low resolution of images. Thus, we concluded that up-scaling step is required to ensure better caption generation for the TinyImageNet dataset. The original Tiny ImageNet200 images have a resolution of 64x64, we super-resolve them 4x using the ESRGAN to obtain images of size 256x256.

$$S_{\text{cap}} = f_{\text{GIT}}(I_{\text{Scaled}})$$

Here, f_{GIT} represents the GIT Transformer function.

3.2.2 Augmented Image Generation using Stable Diffusion V2:

The obtained textual descriptions (S_{cap}) are utilized to generate augmented images (I') through Stable Diffusion model. The diffusion model is conditioned on the obtain textual descriptions during the synthesis process, allowing for the creation of diverse and relevant augmented samples. To save memory during training, we save these augmented images offline, instead of generating them on-the-fly. The images obtained from this step have a resolution of 512x512.

$$I' = f_{\text{Diffusion}}(S_{\text{cap}})$$

The function $f_{\text{Diffusion}}$ represents the Stable Diffusion V2 model.

3.2.3 Contrastive Learning using SimCLR:

The training process begins with two augmented views, denoted as I'_{aug} and I_{aug} generated by applying random horizontal crop, resize, normalization augmentations on I' and I_{Scaled} respectively. These views undergo encoding by a neural network f_{SimCLR} to produce high-dimensional representations: $h_i = f_{\text{SimCLR}}(I_{\text{aug}})$ and $h_j = f_{\text{SimCLR}}(I'_{\text{aug}})$

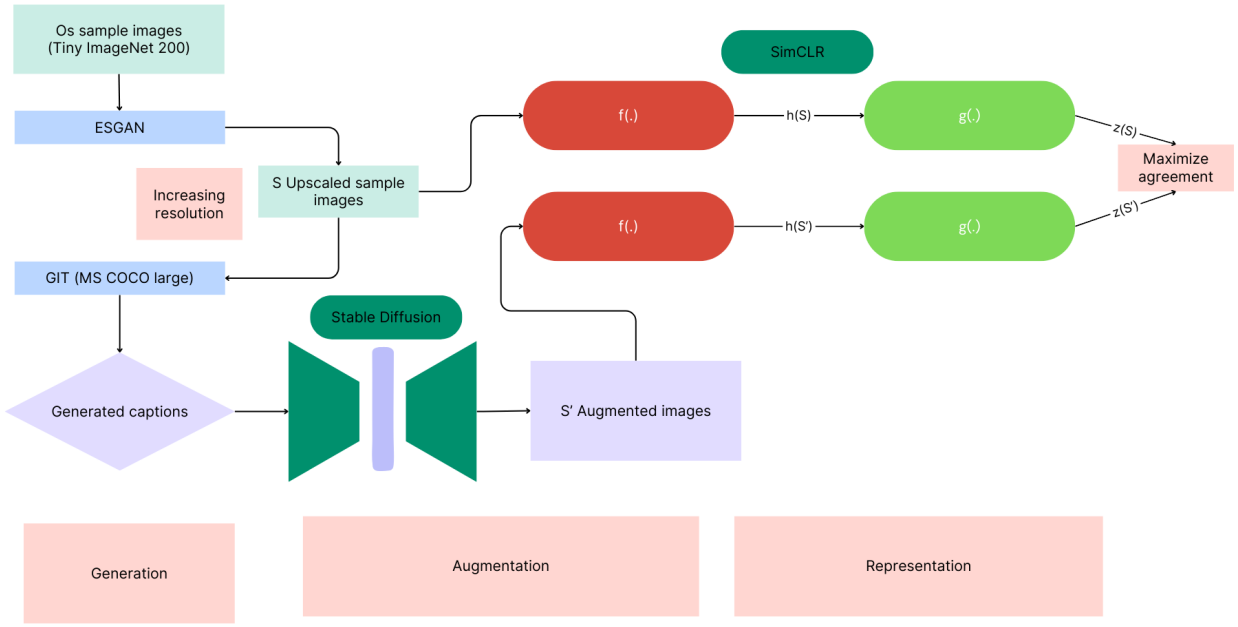

Figure 2. Overview of our pipeline utilizing the weak augmentation of images generated by Stable Diffusion model. To augment the image, we represent it as its caption, this caption is fed to Stable Diffusion Model to generate a corresponding image. Thus, the original image and the generated image serve as two views for our SSL framework.

A subsequent projection head g further maps these representations to a lower-dimensional space $z_i = g(I_{\text{aug}})$ and $z_j = g(I'_{\text{aug}})$ The core of SimCLR's training objective lies in the contrastive loss function, formulated to encourage the model to distinguish between positive and negative pairs in the representation space. The contrastive loss, denoted as $L_{contrastive}(i, j)$, is defined as follows:

$$L_{\text{contrastive}}(i,j) = -\log \frac{\exp\left(\frac{z_i \cdot z_j}{\tau}\right)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp\left(\frac{z_i \cdot z_k}{\tau}\right)}$$

$$L_{\text{final}} = \frac{1}{2N} \Sigma_{k=1}^{N} [l(2k-1, 2k) + l(2k, 2k-1)]$$

where $\mathbb{1}_{[k \neq i]} \in \{0, 1\}$ represents an indicator functizon, evaluating to 1 if $k \neq i$, τ denotes the temperature parameter. and N is the number of training samples. The loss function aims to maximize the similarity between positive pairs and minimize the similarity between negative pairs. The optimization process involves using stochastic gradient descent or variants like Adam to update the parameters of the encoder f_{SimCLR} and the projection head g. Through iterations of this training process, SimCLR learns representations that capture meaningful features of the input data, enabling downstream tasks to benefit from these learned features without the need for labeled data during training. Final loss function is:

3.2.4 Implementation Challenges and Resolutions

Loss Balancing in SimCLR:

The SimCLR framework relies on balancing the contrastive loss, which involves tuning the temperature parameter (τ) and ensuring proper scaling. An imbalance in loss terms may hinder the model's ability to learn meaningful representations.

Resolution: The contrastive loss function ($L_{\text{contrastive}}$) is defined as:

$$L_{\text{contrastive}}(i,j) = -\log \frac{\exp\left(\frac{z_i \cdot z_j}{\tau}\right)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp\left(\frac{z_i \cdot z_k}{\tau}\right)}$$

Properly tune the temperature parameter (τ) to balance the exponential terms, ensuring stable learning. A mathematical analysis of the impact of τ on the loss function can guide the selection of an optimal value.

Generator Stability in Stable Diffusion V2:

Training diffusion models can be sensitive to the stability of the generator, and destabilization may lead to poor sample quality.

Resolution: To stabilize the generator, introduce a gradient penalty term (GP) in the loss function:

$$L_{\text{gen}} = -\mathbb{E}_{\hat{x}\sim p_{\hat{x}}}[\log D(\hat{x})] + \lambda \cdot \mathbb{E}_{\hat{x}\sim p_{\hat{x}}}[(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2]$$

Here, $\hat{x}$ represents the generated samples. The gradient penalty (GP) term penalizes large gradients, promoting stability during training. Mathematically, this stabilizes the generator while maintaining the diversity of synthesized samples.

3.2.5 Implementation Challenges and Resolutions

3.2.6 Experiments

We conduct experiments to evaluate the proposed methodology on the Tiny ImageNet dataset, comparing it with baseline methods. The primary focus is on assessing the quality of augmented images, the robustness of learned representations, and downstream task performance.

Experimental Setup

- Datasets: We use the Tiny ImageNet dataset, consisting of a diverse set of images across multiple classes.

- Baselines: We compare our approach against standard augmentation techniques and recently proposed generative augmentation methods.

Evaluation Metrics:

1. Augmented Image Quality:

- Metric: Frechet Inception Distance (FID) - Definition: Measures the similarity between the distribution of real and generated images using features from an InceptionV3 model. - Mathematical Formulation:

$$FID(P_r, P_g) = \|\mu_r - \mu_g\|_2^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{0.5})$$

where μ and Σ represent the mean and covariance of the feature representations.

2. Robustness of Representations:

- Metric: Linear Evaluation Accuracy - Definition: Evaluates the quality of learned representations by training a linear classifier on top of frozen features. - Mathematical Formulation: Accuracy of a linear classifier trained on representations obtained from the self-supervised model.

3. Downstream Task Performance:

- Metric: Classification Accuracy - Definition: Measures the accuracy of the self-supervised model when fine-tuned on a downstream classification task. - Mathematical Formulation:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

Anticipated Outcomes

We anticipate that our proposed methodology will outperform baselines in the following aspects:

1. Augmented Image Quality:

- Explanation: The tailored augmentation strategy, combining GIT Transformer, Stable Diffusion V2, and SimCLR, is designed to produce diverse and high-quality synthetic samples aligned with textual descriptions. - Mathematical Insight: The lower FID score indicates a closer match between real and generated image distributions, highlighting the efficacy of our approach in producing realistic augmented images.

2. Robustness of Representations:

- Explanation: The coupling of SimCLR for contrastive learning with Stable Diffusion V2 encourages the generation of augmented samples that enhance the robustness and invariance of learned representations. - Mathematical Insight: Higher linear evaluation accuracy indicates that the representations learned through our approach are more effective for downstream tasks.

3. Downstream Task Performance:

- Explanation: The improved quality and robustness of representations are expected to translate into better performance on downstream classification tasks. - Mathematical Insight: Higher classification accuracy in fine-tuning demonstrates the practical utility of the self-supervised model trained with our generative augmentation approach.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan, 2017. 3

[2] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments, 2021. 4

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020. 2, 4

[4] Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation policies from data, 2019. 2

[5] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan, 2020. 2

[6] Donghyun Kim, Kuniaki Saito, Tae-Hyun Oh, Bryan A. Plummer, Stan Sclaroff, and Kate Saenko. Cross-domain self-supervised learning for domain adaptation with few source labels, 2020. 2

[7] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 4

[8] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 4

[9] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang.

Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022. 4

[10] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *International Conference on Computer Vision Workshops (ICCVW)*. 4

[11] Chenyu Zheng, Guoqiang Wu, and Chongxuan Li. Toward understanding generative data augmentation, 2023. 1

[12] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks, 2020. 2