SynthAugment: Diffusion-based Data Augmentation for Low-Data Image Classification

Reinout Wijnholds

Abstract

Deep learning models often require large labeled datasets to achieve good performance, which is challenging in data-scarce scenarios. Data augmentation is a common strategy to expand limited training data, but traditional augmentation (e.g. flips, rotations) only provides limited diversity and cannot create new semantic content. We propose **SynthAugment**, a data augmentation approach leveraging generative diffusion models to synthesize realistic and semantically diverse training images. By using a pre-trained text-to-image diffusion model, we generate new examples that vary in high-level attributes (e.g. object appearance or background) while retaining the original class label. Experiments on image classification tasks with very limited training data (e.g. 10 samples per class) demonstrate that SynthAugment significantly improves accuracy compared to both standard augmentation and GAN-based augmentation. For instance, on CIFAR-10 with 10 images per class, our approach achieves up to 10% higher accuracy than baseline augmentation. We also show that diffusion-generated augmentation yields more diversity and better generalization than prior approaches. Our results highlight that advanced generative models can serve as powerful data augmenters for improving model performance in low-data regimes.

1 Introduction

Training state-of-the-art deep neural networks typically requires large amounts of labeled data. In many domains, collecting such big datasets is difficult or impossible, leading to models that overfit and fail to generalize. Data augmentation is a widely used technique to alleviate data scarcity by artificially expanding the training set with transformed copies of existing images [1]. Common image augmentations include geometric transformations (flips, rotations, crops), color jitter, and other perturbations that yield new training samples without requiring new labels. Augmentation has been essential in improving generalization for tasks like image classification, especially when data is limited [1].

However, traditional augmentation methods have inherent limitations: they only produce images that are variations of the existing ones and cannot change high-level semantic content. For example, flipping or rotating an image will not

introduce new object categories or significant novel appearance variations. Thus, the diversity of augmented data along crucial semantic dimensions remains limited. In scenarios with very few training examples, classical augmentations may be insufficient to capture the variability needed for robust learning.

Generative models offer a promising avenue to go beyond simple transformations by synthesizing entirely new images. In particular, Generative Adversarial Networks (GANs) [2] have been explored for data augmentation. GANs can learn to generate realistic images following the distribution of a training set, potentially providing unlimited new samples. Prior works (e.g., DAGAN [3]) have shown that augmenting with GAN-generated images can improve performance in low-data regimes. However, GAN-based augmentation can be challenging to deploy in practice. Training a GAN on a small dataset can be unstable and prone to mode collapse, often yielding limited diversity. In some cases, synthetic images from GANs did not significantly outperform traditional augmentations [4]. Moreover, GANs trained from scratch for each task require substantial computational effort and can struggle to reproduce fine details of complex data.

Diffusion models have recently emerged as a powerful class of generative models, demonstrating the ability to generate high-fidelity, diverse images that rival or surpass GANs [5, 6]. Diffusion models gradually transform noise into images through a learned denoising process, and have been shown to cover complex image distributions without mode collapse [6]. Importantly, large diffusion models pre-trained on broad image datasets (such as Stable Diffusion [7]) encapsulate a wide variety of visual concepts. This presents an opportunity to leverage knowledge from these models for data augmentation: instead of training a generative model on limited data, we can use a powerful pre-trained generator to create synthetic examples for our target task.

In this paper, we introduce **SynthAugment**, a diffusion-based data augmentation method designed for image classification in low-data settings. Our approach uses a pre-trained text-to-image diffusion model to generate new images for each class in the dataset. By providing the model with appropriate textual prompts or class identifiers, we can synthesize novel images that belong to the same classes as the training data but with different appearances or contexts. For instance, given a class "cat" with only a few real images, we can prompt the diffusion model to produce additional images of cats in various environments, poses, or colors (while still recognizable as cats). These synthetic images are then added to the training set to improve the classifier's generalization.

A key advantage of SynthAugment is that it can introduce new intra-class variation along semantic dimensions that are unreachable by standard augmentations (such as changing the background scene or object attributes). Compared to GAN-based augmentation, using a large pre-trained diffusion model allows us to generate high-quality images even for very small original datasets, since the generative model has learned from a much larger external corpus. We show that classifiers trained with SynthAugment achieve substantially higher accuracy than those with only traditional augmentation or even GAN augmentation. Notably, our method yields improvements up to 8-10 percentage points in ac-

curacy on benchmark datasets with extremely limited training data.

We conduct experiments on CIFAR-10 and CIFAR-100 in few-shot settings (e.g. 10 images per class). We also compare against baseline augmentation techniques and a conditional GAN approach. Our results demonstrate that SynthAugment provides the largest boost in performance, confirming the effectiveness of diffusion-based augmentation. In analysis, we observe that diffusion-generated images offer greater diversity in terms of object attributes and backgrounds. We also discuss strategies to ensure the generated data is relevant and does not introduce labeling errors or unrealistic samples.

In summary, the contributions of this work are:

- We introduce SynthAugment, a novel data augmentation framework that leverages text-to-image diffusion models to generate semantically diversified training images for low-data image classification tasks.
- We empirically show that SynthAugment outperforms traditional augmentation and GAN-based augmentation on CIFAR-10 and CIFAR-100 with limited data, achieving up to 10% absolute improvement in test accuracy.
- We analyze the augmented data and demonstrate that diffusion models can produce realistic variations without custom training on the target dataset, highlighting a practical and reproducible way to utilize pretrained generative models to boost discriminative performance.

The rest of the paper is organized as follows. Section 2 reviews related work on data augmentation and generative models. Section 3 describes the SynthAugment methodology in detail. Section 4 presents experimental results and comparisons. Section 4.3 provides additional discussion, and Section 5 concludes the paper.

2 Related Work

Data Augmentation in Deep Learning. Augmentation techniques have long been used to expand datasets and improve model generalization [1]. Basic image augmentations such as flips, rotations, scaling, cropping, and color jitter are standard in training pipelines for image classification. Automated augmentation policies have also been developed, for example AutoAugment [8] uses reinforcement learning to find an optimal set of transformations, and RandAugment [9] simplifies this by randomly selecting augmentations from a predefined set. While these methods can improve performance by optimizing augmentation parameters, they still rely on the same set of geometric or photometric transformations of existing images.

Another line of augmentation methods creates new images by mixing or interpolating existing ones. MixUp [10] generates convex combinations of pairs of images and their labels, and CutMix [11] combines patches from different images. These approaches can regularize training and yield better generalization. However, they do not introduce fundamentally new visual content; they

produce hybrids of the original data and still cannot generate novel examples outside the span of the training set's support.

GAN-based Data Augmentation. Generative Adversarial Networks [2] learn to model the data distribution and can sample entirely new images. Using GANs for data augmentation has shown promise in low-data scenarios. Antoniou et al. [3] introduced DAGAN, a GAN that learns to augment a dataset by generating within-class variations. They reported significant accuracy gains in few-shot learning by adding GAN-generated images to the training set. Subsequent works have applied GAN augmentation in specific domains (e.g. medical imaging) where data is scarce. For instance, GANs have been used to generate medical scans (like chest X-rays) to improve classification of rare conditions. Nonetheless, the success of GAN augmentation is mixed. Fedoruk et al. [4] found that for a limited COVID-19 X-ray dataset, images generated by StyleGAN2 did not outperform classical augmentation; the classifier trained on GAN-augmented data performed worse than with just traditional transformations. Challenges include the difficulty of training GANs on very small datasets, and the risk that GANs may produce only a narrow range of samples or artifacts that do not help the classifier.

Diffusion Models. Diffusion probabilistic models [5] are a class of generative models that have recently gained attention for their excellent image synthesis quality. Unlike GANs which learn a direct mapping from noise to image via adversarial training, diffusion models learn to iteratively denoise data starting from random noise. This process tends to cover the data distribution more completely and avoids mode collapse, at the cost of longer generation time. Dhariwal and Nichol [6] showed that diffusion models can achieve image sample quality superior to GANs on standard benchmarks (e.g. CIFAR-10, ImageNet). Large-scale diffusion models have been trained on enormous datasets; a notable example is Stable Diffusion [7], a latent diffusion model trained on billions of images with a text-to-image interface. Such models can generate a wide variety of realistic images guided by text prompts.

The ability of diffusion models to produce diverse high-fidelity images makes them attractive for data augmentation. Concurrent to our work, Trabucco et al. [12] proposed a diffusion-based augmentation method (DA-Fusion) focusing on few-shot image classification. Their approach uses an off-the-shelf text-to-image diffusion model to perform semantic edits on existing images (using a technique known as SDEdit and textual inversion to introduce new concepts) in order to generate variant images that expand the dataset. They report improved accuracy on few-shot benchmarks using this strategy. Our work is aligned with this idea of leveraging diffusion models for augmentation, but we explore a somewhat different approach: rather than editing existing images, we primarily generate new images for each class using class-specific prompts. We demonstrate similar benefits of diffusion augmentation and provide additional evaluation against GAN-based augmentation and standard baselines. The general notion of using pre-trained generative models to supply additional training data is a new and promising direction, and our results further confirm diffusion models as effective data augmenters.

3 Methodology

SynthAugment uses a pre-trained text-to-image diffusion model to create synthetic training images for each class in a classification task. In this section, we describe the components of our approach: the generative model, the procedure for synthetic data generation, and how we integrate generated data into training.

3.1 Pre-trained Diffusion Model

We build on a publicly available diffusion model capable of generating images from text descriptions. In our implementation, we use the Stable Diffusion model [7], which is a Latent Diffusion Model trained on the LAION-5B dataset of image-text pairs. This model can produce 512×512 pixel images given a text prompt. It has learned a rich representation of a vast range of visual concepts (objects, scenes, styles) from its training data. By using such a model, we bypass the need to train a generative model on our small target dataset.

Because Stable Diffusion is a text-to-image model, we can guide it to generate instances of a particular class by providing a suitable textual prompt. For example, to generate images for the "cat" class, one can use a prompt like "a photo of a cat" or a more detailed description ("a photo of a small kitten playing with a ball"). The diffusion model will then sample a new image that matches this description. We craft prompts that specify the target class and possibly some generic context (e.g. "a clear photo of a [$class_name$]")$toencouragerealisticoutputs$.

In cases where the class is a concept not well-known to the pre-trained model (for instance, a very specific type of object or an uncommon animal species), the base model might struggle to generate relevant images. Techniques such as *textual inversion* [13] could be applied to teach the diffusion model a new concept by learning an embedding for the new class using the few available images. However, in our experiments on standard datasets like CIFAR, the classes are generic (e.g. "cat", "car", "tree") which are generally well-represented in the diffusion model's knowledge, so we did not require additional fine-tuning of the generative model.

3.2 Synthetic Data Generation Procedure

For each class in the dataset, we generate a set of synthetic images to augment the training data. The procedure can be summarized as follows:

1. **Prompt Design:** We create a text prompt for the class. Typically, we use a simple template such as "*a photo of a [$class_name$]*" $fornaturalobjectclasses.Forclassesthatmightbeamb$ $-200imagesperclass, dependingonhowmanyrealimagesareavailable.Thegenerationprocessisstochastic$ $freeguidancetechniquewithamoderateguidancescale(e.g.7.5asusedinStableDiffusion)toensuretheoutp$

2. **Post-processing:** The generated images are then resized or cropped to the resolution needed for the classifier (e.g. 32×32 for CIFAR). We also inspect the images for obvious failures. In rare cases, the diffusion model

might produce an image that does not actually contain the intended class or has severe artifacts. We discard any such clearly unsuitable images to avoid confusing the classifier. In practice, we found the majority of generated samples to indeed depict the target class when using straightforward prompts.

3. **Labeling:** Each synthetic image is labeled with the class whose prompt was used. Since we control the generation per class, we know the label for each generated image.

By following this process for all classes, we obtain an augmented dataset consisting of the original real images plus the newly generated synthetic images for every class.

It is worth noting that our approach generates images from scratch for each class. An alternative approach (as explored by Trabucco et al. [12]) is to perform *image-to-image* generation, where one starts from an existing image and uses the diffusion model to create a variant (for example, using the SDEdit method which injects noise into an image and then denoises it with a conditioning that can alter certain features). We focused on text-guided generation, but one could also experiment with image-guided diffusion to produce augmentations that preserve some attributes of a specific input image while changing others. We leave such explorations to future work and use the simpler class-conditional generation via text prompting in this study.

3.3 Training with Augmented Data

Once the synthetic images are generated and labeled, we combine them with the original training images. We then train a classifier on this enlarged dataset using standard supervised learning. In our experiments, we use a ResNet-18 [14] architecture as a backbone for CIFAR classification. We train with cross-entropy loss on the combined set of real and synthetic images.

One important consideration is the balance between real and synthetic data during training. If an overwhelming number of synthetic images are added, the classifier might learn to overly rely on or identify artifacts from the generative model, or the effective distribution might shift. We maintain a ratio such that the number of synthetic images per class is at most on the order of the number of real images times a small factor (e.g. if 10 real images per class, we might add 50 synthetic per class, which is a 5:1 ratio). In preliminary trials, we found that having a few times more synthetic than real data was beneficial, but extremely large amounts of generated data yielded diminishing returns and slightly worse performance, likely due to noise or some synthetic bias. In our final experiments, we chose a fixed augmentation amount (like 5x the real data) for consistency.

During training, we apply the usual preprocessing and any standard augmentations to all images (real and synthetic) as is typical. For example, for CIFAR we still apply random horizontal flip and random crop (from padded images) even to the synthetic images. This can further amplify the benefit of the synthetic set.

We do not use any special weighting for synthetic vs real data in the loss; each image is treated equally with its label. An implicit assumption here is that the synthetic images are as trusted as real ones in terms of label correctness and relevance. By manual verification, our generated images were generally correct for the class (especially for clearly defined objects). There is a risk that generative models might sometimes produce out-of-distribution images or incorrect instances (like an image that looks like a blend of two classes). Filtering blatant errors helps maintain label fidelity.

3.4 Reproducibility and Implementation Details

Our implementation uses the HuggingFace Diffusers library to access the Stable Diffusion model. The augmentation process (generating images and adding to training) is automated via a script. The entire pipeline is thus reproducible given the random seeds and the pre-trained model. We note that the use of a pre-trained diffusion model essentially injects external knowledge (from its training data) into our task, which is a reason for its strong performance. This is akin to using a pre-trained feature extractor in transfer learning, but here we use a pre-trained generator to supply new data.

Training of classifiers was done on a single GPU. Generating the synthetic dataset is the most time-consuming step; for example, creating 500 images (10 classes * 50 each) at 512^2 resolution took a few minutes per class using an NVIDIA A100 GPU. We then downsampled to 32^2 for CIFAR. The classification training on CIFAR with augmented data (a few thousand images total) is relatively fast (tens of epochs to convergence).

4 Experiments

We evaluate SynthAugment on image classification benchmarks in low-data settings and compare it with baseline augmentation strategies.

4.1 Datasets and Setup

We use two standard image classification datasets:

- **CIFAR-10**: 10 classes of natural images (airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck) with 32×32 pixel resolution. The full training set has 50,000 images (5000 per class) and test set has 10,000. We simulate a low-data scenario by randomly sampling a small subset of the training data. In our experiments, we use 10 images per class (100 images total) as the training set, unless otherwise specified.
- **CIFAR-100**: 100 classes of 32×32 images (600 images per class in full training). We again sample 10 images per class (total 1000 images) for training in the low-data setting, and evaluate on the standard 10,000 test images.

We intentionally choose the extremely low number of 10 images per class to test the effectiveness of augmentation in a severe data-starved regime (which we refer to as "10-shot" learning here, though not using meta-learning, just standard training on 10 examples per class).

For each dataset, we compare the following training configurations:

1. **No Augmentation (baseline):** Train the classifier only on the given few images per class, with no extra data and only minimal preprocessing (we still normalize images and can use random crop/flip).

2. **Standard Augmentation:** Train on the few images per class but use traditional data augmentation during training (random crops, flips, etc. for each epoch). This does not add new images permanently, but augments on-the-fly.

3. **GAN Augmentation:** Train on an expanded dataset where we add synthetic images generated by a GAN. We use a conditional Deep Convolutional GAN (DCGAN) that we trained on the small dataset to generate images for each class. We add an equal number of GAN-generated images as the diffusion method for fairness (i.e. if diffusion adds 50 per class, GAN also adds 50 per class). Note: training a DCGAN with so few images is not ideal, but we attempt it to simulate a naive GAN augmentation approach.

4. **Diffusion Augmentation (SynthAugment - Ours):** Train on the expanded dataset including synthetic images generated by the diffusion model (Stable Diffusion) as described in Section 3. We use 50 generated images per class for CIFAR-10 and CIFAR-100 in the main results.

All classifiers use the same architecture (ResNet-18) and are trained with identical optimization hyperparameters (we use SGD optimizer, learning rate 0.1, for 100 epochs, with learning rate decay at 50 and 75 epochs by factor 0.1). We ensure that the only difference between runs is the augmentation strategy and data provided.

We run each experiment 3 times with different random seeds for sampling the few-shot training subset (and for GAN/Diffusion generation) and report the average accuracy to account for variability.

4.2 Results

Table 1 shows the classification accuracy on the test set for CIFAR-10 and CIFAR-100 under the different augmentation conditions. We present results averaged over 3 trials (with standard deviation in parentheses).

As seen in the table, SynthAugment (diffusion augmentation) achieves the highest accuracy on both datasets. On CIFAR-10, our method improves the test accuracy from 47.8% with standard augmentation to 59.6%, an absolute gain of about 11.8%. On CIFAR-100, which is a much harder task with 100 classes, the baseline accuracy is very low (around 8-11%) due to the extreme lack of

Training Data	CIFAR-10 (10-shot)	CIFAR-100 (10-shot)
No Augmentation	41.3% ± 2.5	8.5% ± 0.6
Standard Augmentation	47.8% ± 1.8	11.2% ± 0.7
+ GAN Augmentation	53.4% ± 2.1	13.5% ± 0.9
+ **Diffusion Augmentation (Ours)**	**59.6% ± 1.7**	**18.0% ± 1.1**

Table 1: Test accuracy on CIFAR-10 and CIFAR-100 with only 10 training images per class, comparing different augmentation methods. Standard augmentation refers to using flips/crops during training. GAN and Diffusion augmentation use additional generated images (50 per class) added to training data.

data. Diffusion augmentation boosts accuracy to 18.0%, which is a relative improvement of roughly 60% over the standard augmentation baseline.

Using GAN-generated images also provided improvements over standard augmentation, but not as large as diffusion. For example, on CIFAR-10, GAN augmentation reached 53.4%, which is about 6% better than standard aug but still 6% worse than diffusion augmentation. We observed that the quality of GAN images was not very high given the limited data (they were often blurry or lacked variety), which likely limited their usefulness. In contrast, diffusion-generated images were generally sharp and diverse, as qualitatively observed.

Standard augmentation (on-the-fly transforms) did improve over no augmentation, confirming that even basic techniques help a bit in few-shot training (particularly for CIFAR-10, from 41.3% to 47.8%). But the addition of new synthetic data yields a much larger benefit.

The results demonstrate that leveraging a pre-trained diffusion model for data augmentation is highly effective for low-data classification. Particularly noteworthy is that on CIFAR-100, even with only 10 images per class, diffusion augmentation allowed the model to achieve 18% accuracy, whereas without it the model nearly collapsed to 8-11% (only barely above random guessing which is 1%). This suggests the model was able to learn much more about the classes from the synthetic images.

4.3 Analysis and Discussion

We analyzed the synthetic images generated for augmentation to understand what the diffusion model was contributing. Figure **??** shows some example generated images for a few CIFAR-10 classes. We found that for each class, the diffusion model created instances that, while not identical to any training image, captured the general concept of the class. For example, for the "truck" class, the model generated images of trucks with various colors, backgrounds (city streets, highways, etc.), and types (pickup trucks, cargo trucks). This kind of variation is extremely valuable for a classifier, as it exposes it to different possible appearances of "truck".

Another observation was that the diffusion images often had higher resolu-

tion details or more complex backgrounds than CIFAR images. We downsampled them to 32×32, which sometimes made them a bit unclear or noisy at pixel level, but the overall object shape and context still enriched the training data. Potentially, one could generate at the target resolution or use a model specialized for low-res generation, but even using the 512px model and shrinking it worked in our pipeline.

We also considered whether the classifier might be picking up on generative artifacts. In our experiments, adding synthetic data did not cause any overfitting to peculiar artifacts; in fact it reduced overfitting by providing more data. The diffusion model's outputs, being relatively realistic, did not contain obvious common artifacts that the classifier could latch onto (unlike, say, if one used a very poor generator that produced some consistent noise pattern).

One potential concern is that using an external generative model trained on a large corpus effectively leaks extra information into the training process. For example, Stable Diffusion has seen many images of cats, so when we generate cat images, we are tapping into a broader distribution than just the few cats we had. This could be viewed as a form of transfer learning. However, since we do not use any labels beyond our original few, this approach still operates under the assumption of no additional labeled data. It's leveraging unlabeled prior knowledge in the generative model. This is analogous to using pre-trained networks for feature extraction (common in low-data regimes), but here in the data space.

We compare our results qualitatively to those reported by Trabucco et al. [12]. They noted improvements in few-shot classification with diffusion augmentation (DA-Fusion) and in some cases up to 10% absolute accuracy gain, which is in line with what we found (we saw 12% on CIFAR-10). This reinforces that diffusion-based augmentation is a general technique applicable across datasets.

Finally, we stress the reproducibility: since our method uses an open pre-trained model and a straightforward generation process, anyone can replicate our augmentation and results. The code release will include prompt configurations and random seeds for transparency.

5 Conclusion

We presented SynthAugment, an approach to enhance training data for image classification by utilizing generative diffusion models. In low-data settings, where traditional augmentations offer limited gains, our method can inject new, semantically meaningful variation into the training set by generating synthetic images. Through experiments on CIFAR benchmarks, we demonstrated that diffusion-based augmentation significantly outperforms both standard augmentation and GAN-based augmentation, leading to notable accuracy improvements.

This work highlights the potential of pre-trained generative models, such as large diffusion models, as tools for data augmentation in machine learning workflows. As these generative models become more powerful and accessible,

we expect their integration into data-scarce learning problems will become increasingly common. Future work could explore diffusion augmentation for other tasks (e.g. object detection or segmentation) and investigate automated prompt engineering or concept learning to further improve the relevance of generated data. Additionally, combining image generation approaches (diffusion) with other techniques (such as active learning to decide which extra images would be most useful) could amplify benefits.

In conclusion, SynthAugment offers a practical, high-impact solution for improving model performance when data is limited, by tapping into the creative capacity of state-of-the-art AI-generated content. We encourage the community to build on these findings and continue to bridge generative and discriminative modeling for more robust AI systems.

References

[1] C. Shorten and T. M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, 2019.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, et al. Generative adversarial networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2672–2680, 2014.

[3] A. Antoniou, A. Storkey, and H. Edwards. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*, 2018.

[4] O. Fedoruk, K. Klimaszewski, A. Ogonowski, and R. Możdżonek. Performance of GAN-based augmentation for deep learning COVID-19 image classification. *arXiv preprint arXiv:2304.09067*, 2024.

[5] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 2020.

[6] P. Dhariwal and A. Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 34, 2021.

[7] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022.

[8] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le. Autoaugment: Learning augmentation strategies from data. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 113–123, 2019.

[9] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le. Randaugment: Practical automated data augmentation with a reduced search space. In

Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020.

[10] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations (ICLR)*, 2018.

[11] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 6023–6032, 2019.

[12] B. Trabucco, K. Doherty, M. A. Gurinas, and R. Salakhutdinov. Effective data augmentation with diffusion models. *arXiv preprint arXiv:2302.07944*, 2023.

[13] R. Gal, O. Patashnik, A. Hertz, et al. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.