Semantic Data Augmentation with Generative Models

Shivashankar
Shane Miller
csshank,shanem@deltabluelabs.com

Abstract

Data augmentation is an important and widely used technique for training deep neural networks, underlying many recent advances in vision including those from classification, generative models, and representation learning. The standard approach to data augmentation mixes a fixed set of hard-coded transformations such as rotations and flips to generate new images from existing ones. However, such an approach faces two problems; a) it requires encoding domain knowledge of the problem and b) the transformed inputs are lacking in semantic and content diversity. For example, in a species classification task, an ideal data augmentation process would produce images with slightly different anatomical features such as bill length, or number of tusks, horns etc. The standard augmentation techniques cannot produce incorporate such changes in high-level semantic attributes. Moreover doing so realistically for a large number of inputs is outside human feasibility. In this work, we propose to address both these issues by using a pre-trained diffusion model to produce semantically diverse image transformations. Our experiments show that our method can outperform commonly used data augmentation techniques, especially in the low data regime; as well as show improvement in image classification in few-shot setting.

1. Introduction

Deep Neural Networks (DNNs) have shown impressive results across several machine learning tasks [13, 29], and have revolutionized the field of computer vision. However, these successes hinge on the availability of large sets of annotated images. When data is comparatively limited, training modern neural networks relies a lot on implicit regularization and augmentation techniques developed over the last decade such as dropout [52], early stopping, cutmix [62], shufflemix [17], cutout [6] and others—as such data augmentation remains a crucial component of image processing pipelines in practice [33, 47].

Data augmentation (DA) techniques alleviate the scarcity of training examples, by producing new training examples from existing ones. The classical data augmentation (DA) techniques increase the number of training examples by using a predefined set of transformations to the training samples that do not change the corresponding class label. These fixed transformations can be composed together to produce an even wider set of images with the same labels. Such label-preserving operations include geometric transforms such as rotations and flips as well as photometric transforms like blurring, sharpening and jittering. Since these approaches are limited to making the classifier robust only to the fixed set of hard-coded transformations, advanced methods incorporate more loosely defined transformations in the data space. For example, *mixup* [63] uses convex combinations of pairs of examples and their labels, *cutout* [6] randomly masks square regions of the input sample and *patchshuffle* [19] uses a kernel filter to randomly swaps the pixel values in a sliding window. These methods still require domain-specific knowledge that, for example, are applicable only to images and often produce inputs that do not make sense to humans.

Despite these failings, the current methods have been demonstrated to improve the generalization and robustness of deep neural networks. However, they are still not robust with respect to details of visual appearance. A natural way to imbue models with such invariances is to provide inputs created by augmentations which preserve the semantic content of the images. However, creating such augmentations requires a lot of human effort and is infeasible at scale.

Recently, improvements in both generative modeling techniques, multimodal learning and availability of truly large-scale datasets, have led to great progress in image generation. Now we have access to large generative networks that are capable of synthesizing photo-realistic images across a wide range of inputs. Our aim is to utilize such generative models, to augment a training dataset with synthetic images. These generated images can exhibit natural variations in appearance that standard data augmentation methods cannot capture, and hence can be used to improve visual recognition models without significant human effort.

Contribution This work proposes a novel data augmentation strategy that utilizes pre-trained deep generative models to generate new variants of a given real image. Our approach leverages a pre-trained generative model in the following ways:

a) We employ latent space perturbation to generate alternative, yet semantically consistent images of a given visual concept. This allows using an image in order to produce new, realistic views of the same object. For example, we can take an image of a cat and generate multiple realistic images of the cat in different poses and orientations.

b) We can adapt pre-trained text-to-image generative models, by adding new tokens to incorporate novel concepts. This allows us to generate synthetic images of previously unseen objects, which can be used to augment training data for few-shot learning scenarios. For example, we can add new tokens to generate synthetic images of imaginary creatures like a ”unicorn” or a ”dragon.”

c) We leverage diffusion models to change the appearance of both foreground and background while preserving the semantic characteristics of objects. This allows us to generate diverse synthetic images that preserve the essential features of the original image while introducing realistic variations. For example, we can generate synthetic images of a beach with different lighting conditions or weather patterns, while still preserving the essential features of the beach such as sand, water, and palm trees.

Overall, our approach enables the generation of diverse, realistic images that can potentially be used for various computer vision tasks,such as object detection, image classification and segmentation, and few-shot learning.

2. Preliminaries

2.1. Related Work

Data augmentation techniques are routinely used to create synthetic data for improving classifiers [22, 49]. There are primarily two ways to generate such synthetic data: 1) composing various geometric and photometric transforms in a traditional simulation pipeline; 2) producing images from a generative models. While traditional techniques utilize domain knowledge to create synthetic examples similar to training data, modern deep learning techniques often also use other techniques that use interpolation [63], masking [6], patch shuffling [17, 62] etc. These augmentations by construction are limited and can suffer from a real-sim gap where they do not cover the variations of real-world data. Moreover, depending on the type of data source, the amount of data they can generate is bounded.

Compared to these techniques, generative models as a source of synthetic images can be used to create images that capture real world variations (as they are often trained on real-world data). Moreover they can be easier to store and simulate, offer potentially unlimited images, and need not rely on a domain expert. Moreover, with advances in generative modeling, the potential for improvements is large. Neural network assisted augmentation has been explored previously using specifically trained augmentation networks [32, 55] or using GANs [1, 34, 56, 64]. Normalizing flows have also been used to generate semantic perturbations to improve classifiers [61]. Other than improving classifiers [2, 5, 11, 54, 57, 59, 60], generative models have also been used to create data augmentations for segmentation [20, 66],inverse graphics [65] and representation learning [18].

Generative models have garnered significant attention from researchers in both probabilistic modeling and applications work. Early generative models such as Variational Autoencoders (VAEs) [21] and Generative Adversarial Networks (GANs) [10] have been scaled up in terms of resolution and quality [3, 39] with advancements in training these models. Continuous probability flows based on differential equations have recently gained popularity as they have been shown to be better at generative modeling compared to their earlier counterparts [7]. Recent successes in photorealistic image generation have resulted from stochastic diffusion models [45]. Other developments including multimodal learning [36], classifier-free guidance [16], and internet-scale datasets like LAION-5B [46] have led to spectacular successes in text-to-image generation [30, 37, 40, 44]. These developments have unlocked many application areas for generative models.

Text-to-Image generation is a conditional sampling task of producing an image based on a natural language caption. Diffusion models [15, 31] have been combined with multimodal text-image embeddings like CLIP [36] to build models that can sample an image matching a given text. Recent text-to-image models such as Stable diffusion [41], DALL-E2 [38], Imagen [44] and GLIDE [30] have demonstrated incredible results. However, their potential for data synthesis, augmentation and similar tasks is relatively underexplored.

2.2. Background

2.2.1 Diffusion Models

Diffusion models [40, 50, 51] are latent variable models which define the target distribution via a Markov chain (called the reverse diffusion) of learnable transitions starting from an initial distribution $p(x_T) = \mathcal{N}(x_T; 0, I)$. The joint distribution is given as:

$$p_\theta(x_{0:T}) = p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t) \tag{1}$$

The approximate posterior is also given via a Markov chain (called the forward process) of Gaussian transitions with a fixed variance schedule $\beta_1, \ldots, \beta_T$ What distinguishes diffusion models from other types of latent variable models is that the approximate posterior q, called the *forward process* or *diffusion process*, is fixed to a Markov chain that gradually adds Gaussian noise to the data according to a variance schedule $\beta_1, \ldots, \beta_T$:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \tag{2}$$

Parameterizing the transitions of the reverse process, via learned Gaussian distributions with mean $\mu_\theta(x_t, t)$, and fixed covariance σ_t, leads to the following denoising objective:

$$\sum_t \mathbb{E}_{x_t \sim q_t}[\frac{1}{\sigma_t^2}||\tilde{\mu}_t(x_t, x_0) - \mu_\theta(x_t, t)||^2] \tag{3}$$

where $\tilde{\mu}_t$ is the mean of the forward process. By taking μ_θ of the following form

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{\beta_t}{\sqrt{1-\tilde{\alpha}_t}} \epsilon_\theta(x_t, t) \right) \tag{4}$$

and reparameterizing Equation 3 in terms of a standard normal variable ϵ leads to the following objective.

$$\mathbb{E}_{t,x_0,\epsilon}||\epsilon - \epsilon_\theta(\psi_t(x_0, \epsilon), t)||^2 \tag{5}$$

where t is uniformly sampled from $[0, T]$ to match the unweighted sum over time in 3, and x_t are samples from the forward process obtained as

$$\psi_t(x_0, \epsilon) = \sqrt{\tilde{\alpha}_t}x_0 + \sqrt{1 - \tilde{\alpha}_t}\epsilon \tag{6}$$

where $\alpha_t = 1 - \beta_t$ and $\tilde{\alpha}_t = \prod_{s=1}^{t} \alpha_t$. These components allow training and sampling from a diffusion model, which is used as the generative backbone in this work. In this work, we use a pretrained Stable Diffusion model.

A powerful application enabled by diffusion models which is relevant for this work is editing a given image. For example, inpainting with allows the user to mask out a portion of the image and fill it with a different visually similar pixels [26, 43]. SDEdit [27], is a guided image synthesis method, which hijacks a the reverse diffusion process by inserting the user provided input in the middle of the generative process. [14] has used SDEdit for generating synthetic data. We too will use a similar insertion based methodology to create augmentations of an image.

2.2.2 Flow Matching

While diffusion models are state of the art in terms of generative modeling, they are sensitive to training choices. Even more importantly they do not provide a latent representation of the image. Flow matching is an alternative version which instead of relying on stochastic processes, works directly with probability path, with standard diffusion models being a special instance of it. Along with a deterministic latent representation, flow matching also avoids discretization errors induced by a time-grid and is simpler to train. The aim of the flow matching is to learn a temporal vector field $v_t(x) : [0, 1] \times \mathbb{R}^d \to \mathbb{R}^d$, such that the following ODE

$$\begin{aligned} \dot{\phi}_t(x) &= v_t(\phi_t(x)) \\ \phi_0(x) &= x \end{aligned} \tag{7}$$

defines a flow $\phi_t(x) : [0, 1] \times \mathbb{R}^d \to \mathbb{R}^d$ that transforms an initial noise distribution $p_0(x) = \mathcal{N}(x \,|\, 0, I)$ towards the target distribution.

Given a probability path $p_t(x)$ and a vector field $u_t(x)$ that generates p_t, one can learn a generative model by parameterizing its corresponding vector field with a neural network $v_\theta(t, x)$and solve

$$\min_{v_t} \mathbb{E}_{t,p_t(x)} \|v_\theta(t, x) - u_t(x)\|^2 \tag{8}$$

[24] show that by defining conditional flows $p_t(x \,|\, x_1)$ and the corresponding conditional vector field $u_t(x \,|\, x_1)$ such that all intermediate distributions are Gaussian $p_t(x \,|\, x_1) = \mathcal{N}(x \,|\, \mu_t(x_1), \sigma_t^2(x_1))$ and $\mu_0(x_1) = 0, \mu_1(x_1) = x_1, \sigma_0(x_1) = 1, \sigma_1(x_1) = \sigma_{\text{min}}$; one can get the trainable objective from Equation 8.

$$\mathbb{E}_{t,x_1,\epsilon}||v_\theta(t, \psi_t(x_1, \epsilon)) - \frac{\partial \psi_t(x_1, \epsilon)}{\partial t}||^2 \tag{9}$$

where $\psi_t(x_1, \epsilon) = (1 - (1 - \sigma_{\text{min}})t)\epsilon + tx_1$

Comparing Equation (9) and (5), we can see that the two objectives are very similar. The main difference between the two is that (9) models the tangent vector to the path between x and ϵ while (5) matches the vector difference. As such one can consider v_θ in (9) to be the time derivative of ϵ_θ in (5). [1]. This is partly due to the choice of linear time path chosen for ψ. However, one can choose alternate probability paths for the continuous flow between the noise and target distribution. Sampling from the learned model can be obtained by first sampling $x_0 \sim \mathcal{N}(x \,|\, 0, 1)$ and then numerically solving (7). Similarly a latent representation for an input image x_1 can be obtained by solving the same ODE backward for x_0.

3. Method

Most common augmentation techniques can be applied to real images independent of their visual content. Our goal

[1] Note that for flow matching x_1 corresponds to target distribution and x_0 corresponds to noise; which is the opposite convention to diffusion models

is to replicate such flexibility with diffusion models. This suggests the following desired features: 1. produce on-manifold augmentations for most real images 2. be able to work with novel or rare classes 3. be able to control what to augment based on segmentation information 4. be composable with with other data augmentations, including those based on generative models

3.1. Randomized Latent Augmentations

The invertibility of vector field in flow matching, enables bidirectional transitions between image and latent spaces. This, in turn, allows for applying perturbations directly in the latent space rather than image space. We denote the push forward operation induced by the vector field ϕ_t as $[\phi_t]_*$, then $[\phi_1]_*$ is a function mapping from the latent space to the data space, while the pull-back $[\phi_1]^*$ maps the data manifold to latent vectors. Given a perturbation function $\mathcal{P}$ defined over the latent space, we want to create identity-preserving semantic modifications over the original image x in the image domain. To this end, we consider only incremental perturbationsand use a scaling parameter δ to control the size of perturbation change More precisely, we have:

$$[\phi_1]_*\big([\phi_1]^*(x) + \mathcal{P}([\phi_1]^*(x), \delta)\big) .$$

At training time, given an image x_i, we use the flow matching model to get the corresponding latent code $z_i = [\phi_1]^*(x_i)$. We consider a simplistic Gaussian noise in the latent space:

$$\mathcal{P}_{rand}(\cdot, \delta) = \delta \cdot \mathcal{N}(0, \mathbf{I}) ,$$

which is independent from z_i. If the base distribution for the generative model is Gaussian, the above perturbation is equivalent to sampling from the learned manifold.

3.2. Modelling Novel Visual Concepts

The previous augmentation technique, which relies solely on the capabilities of the generative model, may not suffice when dealing with novel labels. While the generative model can invert an image of an unknown class to a latent vector, there is no guarantee that perturbations to it will result in semantically consistent changes. This is particularly true for compact base distributions where the latent vector may not even lie in the support of the base distribution. Thus, it is necessary to adapt the generative model to be able to use novel concepts that were previously unknown to it. This adaptation requires incorporating new tokens in the text encoder that correspond to these novel concepts, as well as fine-tuning the diffusion model to generate plausible augmentations for these images. This increases the flexibility of our approach and enables us to achieve semantically consistent data enhancement across a wide range of categories.

Existing work often uses a given class label for generating synthetic images. While this can work for common objects, in an ideal setting we want to provide an entirely new category to the generative model. This can be addressed by incorporating new tokens in the text encoder that correspond to these novel concepts. Specifically given a hyperparameter corresponding to new classes c, we introduce one new token for each class in the vocabulary. And inversion attack [9] is then appplied to each new token, to obtain their embeddings. We then fine-tune these embeddings via standard training of the generative model.

Rather than generate synthetic images from scratch, we use the splicing technique proposed in SDEdit [27]. Given a time-dependent transformation acting on the initial input, one can insert a real image x_0^{ref} modified with gaussian noise ϵ. The exact moment of insertion t_0 is a user specified criteria, and the variance of the added noise depends on the timestep t_0 . SDEdit [27] chose the reverse diffusion process to interrupt and modify with the real input, however we can accomplish the same with the Flow matching process as well. The reverse diffusion process is nothing but transforming the latent vector along the probability path, and an analogous change can be made to the latter process.

$$x_{t_0} = \sqrt{\tilde{\alpha}_{t_0}}\epsilon + \sqrt{1 - \tilde{\alpha}_{t_0}}x_0^{\text{ref}} \tag{10}$$

3.3. Object-Centric Augmentations

Our approach thus far applies global transformation to an image, regardless of its visual content. While traditional data augmentations [35, 48] follow the same pattern of global transformations, this poses a challenge with respect to semantic augmentations. Real images often contain multiple objects with different fundamental invariances. Capturing these differences using a single global transform might not be an ideal. Applying transformations at the object level is more intuitive and flexible.

Unlike standard augmentation methods, using diffusion based generative model allows us to leverage inpainting [26, 43] to selectively modify parts of the image. Thus with user guidance in form of a mask, we can selectively sample only a new background for example. Given a pixelwise mask $v \in [0, 1]^{H \times W}$ specifying which image content to modify, we insert content into the diffusion process at locations where v_{ij} is close to one. In particular, after every timestep t_0 along the probability path, we can assign the pixels in the masked part of the input image a new value sampled from the transition model. This process is entirely analogous to the process used in [42], except it has been applied to the ODE transform rather than the stochastic diffusion process.

3.4. Composing Augmentations

A natural metric when augmenting images is their diversity. This is also reflected in a variety of SOTA training pipelines where simple transformations are typically

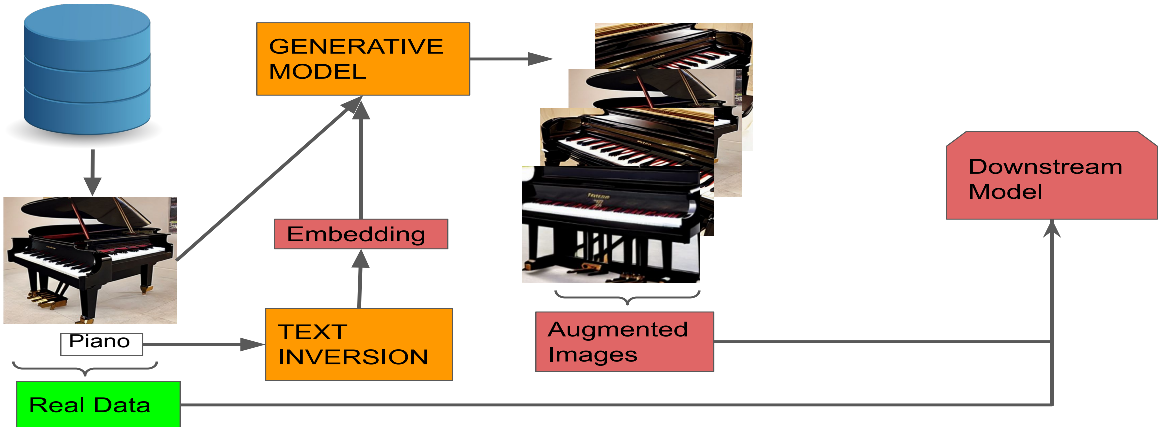

Figure 1. Given a dataset of images we generate versions using a pretrained generative model. These augmented images are included with real data for training downstream models. For few-shot/novel-class learning we use the class label and perform Textual Inversion [9] to get token embeddings for novel concepts which are provided to the image generator.

composed, yielding more sophisticated and diverse data. [4,48,58]. The validity of this concept has also been demonstrated with synthetic images recently. Given a set of image transformations $D_1, D_2, ..D_k$, we randomly sample one augmentation D_a with probability p_a. We consider D_i to be an augmentation adding random noise at time $t = i/k$ in the image generating transformation. Since $t = 0$ corresponds to the fully latent space, while $t = 1$ is the image space, this induces a natural hierarchy of augmentations with lower t corresponding to higher level augmentations.

4. Experiments

In this section, we study the performance and flexibility of our method in both standard classification as well as in a few-shot setting.

4.1. Evaluating Image Classification

We first evaluate whether using synthetic data generated by our method is useful in the standard image classification setting. We evaluate our latent augmentation method on CIFAR as we are primarily interested in the performance boost obtained in the low-data regime. Two other reason to consider this dataset is that a) with current models and full training set this problem is widely considered as solved; and b) existing works have looked at the effect of synthetic data generated from different models for this data.

Experimental details Following earlier literature [61] we train ResNet classifiers on both 5% anf 100% of the full training set and evaluate models on the test set. We report results of common augmentation methods such as *Cutout* [6] and *Mixup* [63] from existing literature. Along with these augmentations, we also report comparisons with normalizing flow [61] and VAEGAN [53]. Table 1 summarizes the results obtained from these experiments.

Method	Low-data	Full-set
Vanilla	49.8	89.7
Classical DA	64.1	95.2
VAE-GAN [53]	58.9	94.2
Cutout [6]	66.8	96.0
Mixup [63]	73.4	95.9
Normalizing Flow [61]	70.1	96.3
Ours	**77.5**	**96.7**

Table 1. Test accuracy (%) on CIFAR, in the *low-data regime* compared to the *full train set*. For the former, we use 5% and 100% of the training and test set, respectively. In addition to standard training, we consider standard training with commonly used data augmentations (DA) in the image space, as well as *Cutout* [6] and *Mixup* [63] methods.

Results Unsurprisingly, the performance of the neural networks is substantially lower, when dealing with limited data. With very deep networks and limited data, overfitting is an issue as the model is too flexible to correctly capture the underlying structure of the data. All augmentation methods, help alleviate this problem with more than 10% improvement in all cases. However, there is a wide gap between augmentations produced by modern generative models over other methods. Our approach shows significant performance improvements over older generative models as wel as mixup like augmentations. However, we also note that even with the use of large-scale generative models, the performance improvement is not enough to completely compensate for the lack of training data.

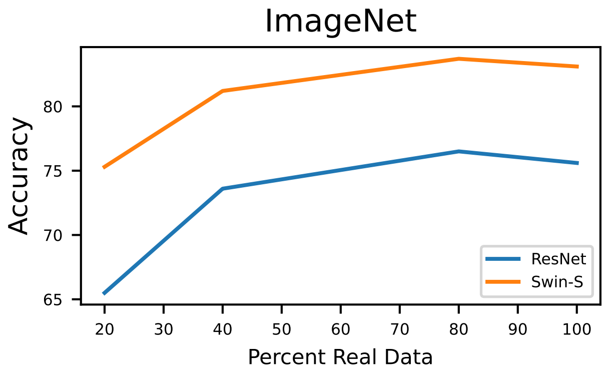

Figure 2. Analysis for different fraction of synthetic to real data ratio. We test on ImageNet with both ResNet and Swin models. The plot shows adding too many synthetic images hurts standard image classification.

4.1.1 Further Analysis

We next take look at possible cause of failure in generalizing with generative models. A natural suspect is distribution shift between training data and the images generated by the generative model. While visual inspection of a few samples, does suggest differences, the image types and details, make it hard to analyse this clearly. As such we shift to the larger and more detailed ImageNet dataset. We specifically use the ILSVRC-2012 subset which contains over 1,000 categories. We used ResNet [13] and SwinTransformer-S (Swin-S) [25], for this purpose. We evaluated these models on varying mix of training and synthetic data. Figure 2 depicts these results where we plot the accuracy of the two models across varying level of synthetic data.

As can be seen, while performance increases by adding a small amount of synthetic data, as the fraction of synthetic images increases the performance starts deteriorating. For a small amount of synthetic data (20%) we see a flat or mild increase in performance, however as the amount of generated images crosses 40%, the performance drops become significant. At 100% synthetic data, the performance is substantively lower (around 20% [2]). However, a small amount of real training data to 100% synthetic data improves the results by a lot. This is a clear indication of a distribution shift between the training and synthetic images. Similar performance drop has also been noted by [12]. However, they have reported a consistent drop in performance, instead of small gains with low levels of augmentation. We believe this to be due to the fact that they generate images purely from textual description. This causes the model to lose visual content from the real image, leading to larger distribution shift than when using latent perturbations.

4.2. Evaluating Few-Shot Learning

Unlike stadard classification, few shot learning requires a model to learn entirely new classes with few (sometimes even 0 examples). From results in the previous section, where 100% synthetic data had a significantly above random performance, gives credence to the hypothesis that synthetic examples can significantly improve performance in this setting. Recently, [14] has shown this to the case. In these set of experiments, we further build upon their results. One key difference from their approach is that we will focus on a greater exploration of the type and nature of augmentations for improving few shot learning against their work, which has looked at a wider setting with general synthetic data.

Pascal Visual Object Classes challenge [8], consists of 11,530 images and 6,929 object segmentation masks. We repurpose this dataset for object classification by removing out images that do not have at least one object segmentation mask. The remaining images are then labeled according to the object class with the largest area in the image. This results in a total of 20 classes for our classification task. We use the official training and validation sets for the 2012 challenge. These images are used to measure few-shot classification accuracy.

COCO is another widely used dataset in computer vision [23]. This dataset contains 330K images with 1.5M object segmentation masks. We perform similar processing as descried earlier for Pascal, and adopt the same evaluation methodology.

Note that since both Pascal and COCO datasets contain common objects, like household items, cars , planes etc.; a true few shot evaluation on these models with large generative models is not possible. This is because these generative models are trained on publically available images from the internet, which would contain a multitude of these common objects. In an ideal scenario, one would need to force the generative model to forget such information. However this is a very hard problem [28]. As such we would use the method adopted in [14], of using prompts that does not contain the class name.

Experimental Details In this experiment, we compare the few-shot classification of our method against two data augmentation strategies. The Real Guidance baseline is based on [14], and uses SDEdit on real images. The DA-Fusion method uses similar method as ours, but does not use latent augmentations. For this experiment we use a pre-trained ResNet [13] classifier and fine tune it on a mixture of real and synthetic examples

Results Figure 3 shows our results. We observe a consistent improvement in accuracy by as much as 2% on the Pas-

[2] Not plotted for better visualization in Figure2

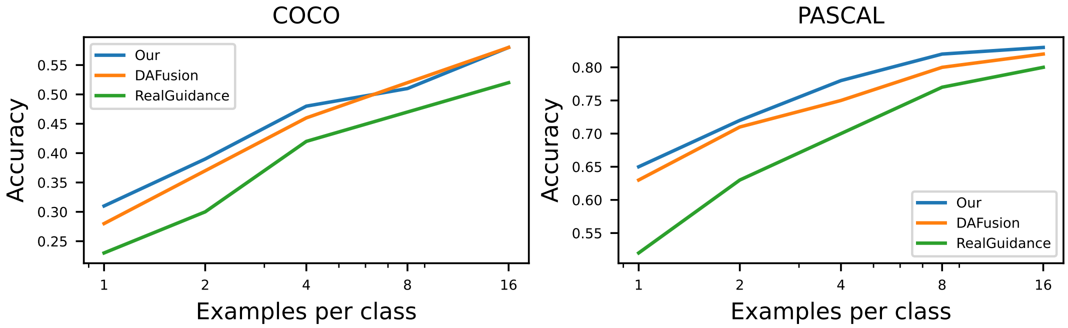

Figure 3. Few-shot classification performance. We evaluate our method on COCO and PASCAL, and find it outperforms other baselines including a recent generative model based method. The figure reports validation accuracy as a function of the number of images in the observed training set.

cal and COCO datasets compared to DA-Fusion which itself exceeds the performance of standard augmentation methods by close to 9% on all domains. We also exceed the performance of Real Guidance [14]. This experiment shows that generative model based augmentation can be used to improve few-shot learning and better generalize to out-of-vocabulary concepts. In the following sections, we ablate these results to understand how important each part of the method is to these gains in performance.

4.2.1 Further Analysis

Our previous results show synthetic images obtained via diffusion models provide effective data augmentation for few-shot learning. In the following experiments, we perform further analysis with respect to object localization and composition of augmentations.

Ablation with Object Masks We wish to analyse how our method behaves when dealing with changes localized to a single input object. Augmentations based on generative models can achieve this behavior using inpainting [26, 43], and stable performance when various masks are given is a desirable trait for any such method.

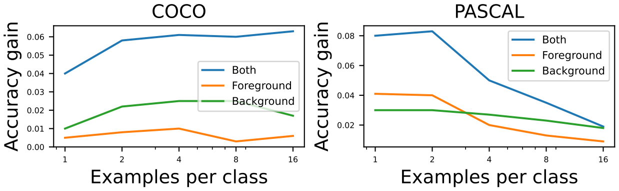

Figure 4. Ablation for masking. We evaluate our method when applied to foregrounds and backgrounds separately. We report the improvement in few-shot classification accuracy on a validation set for our method compared to a Real Guidance baseline using object segmentation masks from the Pascal and COCO datasets. Results show a consistent improvement over the baseline when masks are used.

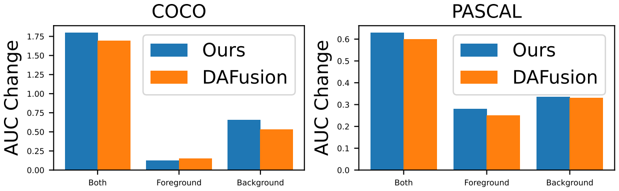

Figure 5. We compare the performance of our method against the DAFusion. We evaluate all three setting for masking: foreground, background, and both. Results show a consistent improvement over the baseline DAFusion.

We assess the robustness of the few-shot learning outcomes by employing mask-based data augmentation to produce novel training images. The experiment incorporates two categories of masks - foreground object masks and background masks - which are utilized to generate masked renditions of the original images. The foreground object mask constitutes a binary overlay that encompasses the focal object in each image. This mask is subsequently employed for inpainting using a pretrained diffusion model. Inpainting aims to reconstruct the obscured portions of the image veiled by the mask, resulting in an altered version of the image with a modified appearance. To guarantee that the principal object is entirely enclosed within the mask, the mask undergoes dilation by 16 pixels prior to its utilization for inpainting. The background mask is generated by inverting the foreground mask for each image, meaning it covers everything excluding the primary object. Analogous to the foreground mask, the background mask is harnessed for inpainting to produce a transformed version of the image. We test three types of masked augmentations: one that augments the whole image, one that augments only the foreground, and one that augments only the background. We train a classifier on the samples generated by the masked augmentation and compare the accuracy to the accuracy of the classifier trained on samples from the Real Guidance [14] baseline. These results are plotted in Figure 4, where we can see that our method consistently outperforms prior work given masks for objects and backgrounds.

Ablation with Composition Next, we present results demonstrating how compositionality in our method (as described in Section 3.4) affects performance. For this we use the same experimental setting of Section 4.2; however we vary the number of augmentations used during training. We report the improvement in area under the curve (AUC) versus standard data augmentation for our method. While all augmentations improve over the baseline, stacking more augmentations lead to greater improvements.

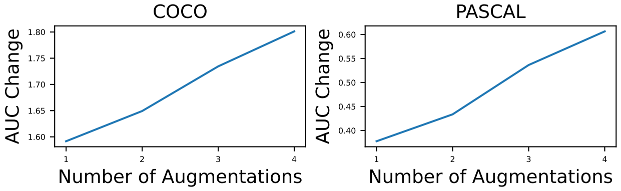

Figure 6. Ablation for number of augmentations. We vary the number of augmentations used for generating images and report the change in AUC. Results show that adding more augmentations improve the performance

5. Conclusion

We propose a versatile method for data augmentation-grounded in diffusion models. Our method tailors a pretrained diffusion model to produce high-quality augmentations while maintaining image semantics. This strategy is adaptable, fitting a wide array of images regardless of their labels or visual characteristics. Our experimental findings reveal that harnessing generative models can significantly bolster performance in both standard and zero-shot classification contexts.

In the standard classification scenario, there exists a limit on the enhancements provided by additional augmentations due to the distribution divergence between real examples and generated samples. Despite this limitation, synthesized instances prove highly beneficial for generalization in few-shot classification settings. Our proposed technique has bolstered the accuracy of few-shot classification across all examined domains, surpassing other recent generative model-based augmentation methods.

We perform an exploratory analysis to assess the robustness of our approach under various conditions. Our results indicate that even when portions of images are masked, our method can still yield performance improvements. Furthermore, we discover that latent augmentations, facilitated by employing invertible flows, contribute to additional advancements, likely owing to their capacity to synthesize images more closely aligned with a given input. In conclusion, our proposed method presents a flexible and effective means of augmenting classification models across diverse domains.

Limitations Effectiveness of these generative models for data augmentation is highly dependent on the specific dataset and task at hand. Furthermore, few shot learning is much more affected by such augmentations than standard learning. This suggests that in some cases, using certain generative models may not be effective at all, while in others, they may show significant improvement. Therefore, it is important to carefully evaluate the performance of generative models in the context of the specific task and dataset.

References

[1] Antreas Antoniou, Amos Storkey, and Harrison Edwards. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*, 2017.

[2] Victor Besnier, Himalaya Jain, Andrei Bursuc, Matthieu Cord, and Patrick Pérez. This dataset does not exist: training models from generated images. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2020.

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[4] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Simclr. `https://github.com/google-research/simclr`, 2020.

[5] Pham Thanh Dat, Anuvabh Dutt, Denis Pellerin, and Georges Quénot. Classifier training from a generative model. In *2019 International Conference on Content-Based Multimedia Indexing (CBMI)*, pages 1–6, 2019.

[6] Terrance DeVries and Graham W. Taylor. Improved Regularization of Convolutional Neural Networks with Cutout. *arXiv:1708.04552*, 2017. arXiv: 1708.04552.

[7] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat gans on image synthesis. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 8780–8794, 2021.

[8] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88:303–308, September 2009. Printed version publication date: June 2010.

[9] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion, 2022.

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and

Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[11] Ayaan Haque. EC-GAN: low-sample classification using semi-supervised algorithms and gans (student abstract). In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 15797–15798. AAAI Press, 2021.

[12] Ryuichiro Hataya, Han Bao, and Hiromi Arai. Will large-scale generative models corrupt future datasets? *arXiv preprint arXiv:2211.08095*, 2022.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition?, 2022.

[15] 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):1–33, 2022.

[16] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022.

[17] Hiroshi Inoue. Data augmentation by pairing samples for images classification. *arXiv preprint arXiv:1801.02929*, 2018.

[18] Ali Jahanian, Xavier Puig, Yonglong Tian, and Phillip Isola. Generative models as a data source for multiview representation learning. *arXiv preprint arXiv:2106.05258*, 2021.

[19] Guoliang Kang, Xuanyi Dong, Liang Zheng, and Yi Yang. Patchshuffle regularization, 2017.

[20] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[21] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[23] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer, 2014.

[24] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[25] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[26] Andreas Lugmayr, Martin Danelljan, Andrés Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 11451–11461. IEEE, 2022.

[27] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

[28] Kevin Meng, David Bau, Alex J Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[29] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, Feb. 2015.

[30] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[31] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021.

[32] Xi Peng, Zhiqiang Tang, Fei Yang, Rogerio S. Feris, and Dimitris Metaxas. Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[33] Luis Perez and Jason Wang. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*, 2017.

[34] Luis Perez and Jason Wang. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*, 2017.

[35] Luis Perez and Jason Wang. The effectiveness of data augmentation in image classification using deep learning, 2017.

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,

Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 2021.

[37] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

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

[39] Ali Razavi, Aäron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with VQ-VAE-2. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 14837–14847, 2019.

[40] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE, 2022.

[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[42] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022.

[43] Chitwan Saharia, William Chan, Huiwen Chang, Chris A. Lee, Jonathan Ho, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In Munkhtsetseg Nandigjav, Niloy J. Mitra, and Aaron Hertzmann, editors, *SIGGRAPH '22: Special Interest Group on Computer Graphics and Interactive Techniques Conference, Vancouver, BC, Canada, August 7 - 11, 2022*, pages 15:1–15:10. ACM, 2022.

[44] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

[45] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022.

[46] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022.

[47] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

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

[49] Patrice Y. Simard, Yann A. LeCun, John S. Denker, and Bernard Victorri. *Transformation Invariance in Pattern Recognition — Tangent Distance and Tangent Propagation*, pages 239–274. Springer Berlin Heidelberg, Berlin, Heidelberg, 1998.

[50] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis R. Bach and David M. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 2256–2265. JMLR.org, 2015.

[51] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[52] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research (JMLR)*, 15(1):1929–1958, 2014.

[53] David Stutz, Matthias Hein, and Bernt Schiele. Disentangling adversarial robustness and generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6976–6987, 2019.

[54] Fabio Henrique Kiyoiti dos Santos Tanaka and Claus Aranha. Data augmentation using gans, 2019.

[55] Zhiqiang Tang, Yunhe Gao, Leonid Karlinsky, Prasanna Sattigeri, Rogerio Feris, and Dimitris Metaxas. Onlineaugment: Online data augmentation with less domain knowledge. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 313–329. Springer, 2020.

[56] Ngoc-Trung Tran, Viet-Hung Tran, Ngoc-Bao Nguyen, Trung-Kien Nguyen, and Ngai-Man Cheung. On data augmentation for GAN training. *arXiv preprint arXiv:2006.05338*, 2020.

[57] Sajila Wickramaratne and Md Shaad Mahmud. Conditional-gan based data augmentation for deep learning task classifier improvement using fnirs data. *Frontiers in Big Data*, 4:659146, 07 2021.

[58] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019.

[59] Wei Xiong, Yutong He, Yixuan Zhang, Wenhan Luo, Lin Ma, and Jiebo Luo. Fine-grained image-to-image transformation towards visual recognition. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020.

[60] Shin'ya Yamaguchi, Sekitoshi Kanai, and Takeharu Eda. Effective data augmentation with multi-domain learning gans. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 6566–6574. AAAI Press, 2020.

[61] Oguz Kaan Yüksel, Sebastian U Stich, Martin Jaggi, and Tatjana Chavdarova. Semantic perturbations with normalizing flows for improved generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6619–6629, 2021.

[62] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 6023–6032, 2019.

[63] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations (ICLR)*, 2018.

[64] Ruixiang Zhang, Tong Che, Zoubin Ghahramani, Yoshua Bengio, and Yangqiu Song. Metagan: An adversarial approach to few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31. Curran Associates, Inc., 2018.

[65] Yuxuan Zhang, Wenzheng Chen, Huan Ling, Jun Gao, Yinan Zhang, Antonio Torralba, and Sanja Fidler. Image gans meet differentiable rendering for inverse graphics and interpretable 3d neural rendering. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[66] Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. Datasetgan: Efficient labeled data factory with minimal human effort. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 10145–10155. Computer Vision Foundation / IEEE, 2021.