Data Augmentation for Image Classification using Generative AI

Fazle Rahat[1] M Shifat Hossain[1] Md Rubel Ahmed[1] Sumit Kumar Jha[2] Rickard Ewetz[3]
[1]University of Central Florida, Orlando, FL
[2]Florida International University, Miami, FL
[3]University of Florida, Gainesville, FL
{fazle.rahat,mshifat.hossain,mdrubel.ahmed}@ucf.edu jha@cs.fiu.edu rewetz@ufl.edu

Abstract

Scaling laws dictate that the performance of AI models is proportional to the amount of available data. Data augmentation is a promising solution to expanding the dataset size. Traditional approaches focused on augmentation using rotation, translation, and resizing. Recent approaches use generative AI models to improve dataset diversity. However, the generative methods struggle with issues such as subject corruption and the introduction of irrelevant artifacts. In this paper, we propose the Automated Generative Data Augmentation (AGA). The framework combines the utility of large language models (LLMs), diffusion models, and segmentation models to augment data. AGA preserves foreground authenticity while ensuring background diversity. Specific contributions include: i) segment and superclass based object extraction, ii) prompt diversity with combinatorial complexity using prompt decomposition, and iii) affine subject manipulation. We evaluate AGA against state-of-the-art (SOTA) techniques on three representative datasets, ImageNet, CUB and iWildCam. The experimental evaluation demonstrates an accuracy improvement of 15.6% *and* 23.5% *for in and out-of-distribution data compared to baseline models respectively. There is also* 64.3% *improvement in SIC score compared to the baselines.*

1. Introduction

Deep learning models often struggle with domain adaptation when exposed to new conditions, such as changes in weather, lighting, and geographic locations [1]. This issue is particularly evident in applications like rare bird or animal species identification, where insufficient training data can hinder the model's ability to generalize effectively [2]. Adding more training data from diverse domains can help alleviate this issue; however, collecting high-quality and relevant data is inherently costly [3].

Significant research efforts have been dedicated to traditional data augmentation approaches based on geometric modifications, including cropping, translations, and rotations [2]. The limitations of these techniques is that the subject features may be altered and the limited image diversity. On the other hand, the recent advancements within generative AI is providing new opportunities for data augmentation [2] using large language models (LLMs) [4], vision-language models (VLMs) [5, 6], image synthesis models [7, 8]. In particular, the ability to synthesize photo realistic images from natural language [9, 10, 11]. These models demonstrate exceptional performance on various tasks such as text-to-image generation [12, 13], image-to-image modification [14, 15], and image inpainting [16]. Recent work shows that large-scale diffusion models can be fine-tuned to generate augmented images for improving recognition tasks [13]. While fine-tuning image generation models for data augmentation is effective, their complexity and the need for replication across diverse datasets often make it impractical [17]. Methods for augmenting visually realistic images using text-guided techniques without model fine-tuning are proposed in [15, 17, 18]. However, our case study indicates that diffusion models struggle to augment fruitful training data from text prompts alone, often deviating from the intended subjects in the generated images.

In this paper, we propose the Automated Generative Data Augmentation framework called AGA to augment the training dataset to enhance fine-grained classification performance. Our method aims to alter the subjects minimally while introducing variability in the backgrounds during the augmentation process. AGA uses image segmentation to isolate subjects, a pre-trained LLM for varied background captions, Stable Diffusion for diverse background creation, and integrates subjects seamlessly with backgrounds. Automatic background image generation faces two main challenges. The first is creating diverse backgrounds without corrupting the foreground, a problem often overlooked by existing methods effectively addressed by the subject isolation technique of AGA . The other challenge is creating the right LLM prompt automatically. AGA solves this by including a prompt generation engine equipped with hierarchical instruction, spatial and temporal modality fixers.

Figure 1. Example augmentation using text-to-image, image-to-image, inpainting, and our approach on ImageNet10. Images generated by text-to-image and image-to-image significantly lose foreground information. Inpainting provides comparatively better results but corrupts the foreground with unnecessary modifications. AGA is able to produce diverse background images while keeping the foreground information grounded with original images.

This engine automatically produces a diverse set of text prompts for the LLM while using a small library of sample instructions, which ultimately ensures the diversity in the generated backgrounds. This paper makes the following key contributions:

- We introduce AGA , an innovative framework for data augmentation that focuses on diversifying backgrounds while preserving the subject of interest with various affine transformations, leading to robust and explainable classifiers.
- AGA leverages large language and vision models to automatically create diverse training images, eliminating the need for manual user input or fine-tuning.
- AGA improves the accuracy of fine-grained classification from 78.4% to 93.6% on a reduced version of the ImageNet dataset.

The rest of the paper is organized as follows: Section 2 discusses related work, Section 3 motivates our work with a case study, Section 4 presents our methodology. The experimental results are discussed in Section 5 and finally, Section 6 concludes the paper with potential future work.

2. Related Works

Image augmentation is a pivotal method for improving the performance and generalization ability of deep learning models. Early works often resort to geometric transformations such as flipping, cropping, and rotation, color space transformations, kernel filters [2]. Beyond simple manipulation, advanced techniques like Mixup [19] and CutMix [20] introduce advanced techniques such as mixing images to create new training examples and encourage the model to learn more robust representations. Additionally, automated augmentation methods such as RandAugment [21] randomly select and apply a sequence of transformations with varying magnitudes, eliminating the need for manual tuning of augmentation hyperparameters. However, these techniques often generate images which are not only visually unnatural [15] but also loses subject information.

With the advent of generative AI models, particularly diffusion models, image augmentation has witnessed a paradigm shift and these models are widely adopted in image generation [9, 12, 13, 15, 17, 22, 23]. Large-scale image-text datasets and models like CLIP [24] have enabled SOTA diffusion models to perform versatile tasks such as text-to-image generation, image-to-image transformation, and inpainting through text-guided prompts. Several studies have investigated how to enhance the classification accuracy using synthetic images generated by diffusion models [25, 26]. One study showed that it is possible to train a classifier for ImageNet solely using synthetic Data, leading to a performance improvement when applied to real-world tasks [27]. While another investigation demonstrates the effectiveness of fine-tuning Imagen [28] for data augmentation on ImageNet [13]. These fine-tuning based approaches face practical challenges due to complexity, cost, and dataset-specific requirements. Recent works utilize off-the-shelf diffusion models to diversify vision datasets without the need for fine-tuning [15, 17].

Recent works create synthetic images using either text-to-image [12] or image-to-image [14] methods, with text-guided image generation. Image-guided inpainting [16] also utilizes image modification to introduce diversity in the image data. However, these techniques significantly distort required subject information. To solve this issue and generate synthetic images without losing subject information, we propose AGA . This is an automatic segmentation-guided technique that utilizes recent object detection and segmentation models [29, 30] to augment data. AGA generates effective synthetic images while keeping foregrounds grounded with the original images.

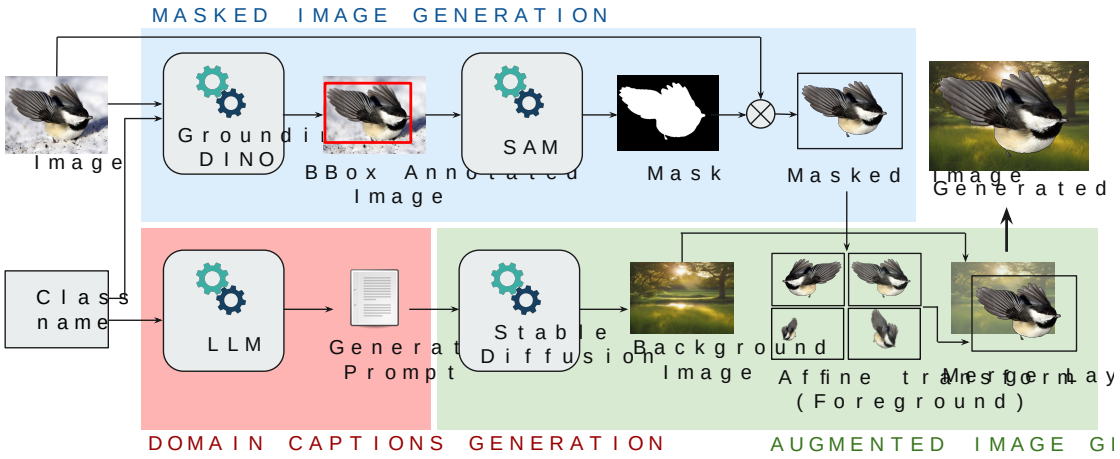


Figure 2. The methodology of the AGA framework. The inputs are an image and originals class name, while the outputs are corresponding augmented images. Subject isolation from input is performed by masked image generation. The domain captions generation engine generates diverse background prompts, which are utilized by stable diffusion to generate background images. Finally, these background images and isolated subjects are combined to generate augmented images.

3. A Motivating Case Study

Text-to-image, image-to-image, and inpainting are three key image augmentation techniques extensively utilized in recent image augmentation works. We conducted a case study of these methods for several datasets like ImageNet [31], CUB [32], and iWildCam [33], to understand their advantages and shortcomings. We discuss our observations using a representative image of a bird from the CUB data set, as illustrated in Figure 1.

Text-to-image: It can be observed that text-to-image, while capable of generating a high diversity of images, often produces samples in which the subject is so drastically altered that even human observers struggle to identify it. We find in the figure that the identifying mark of the bird, the red ring around the neck, is missing in the augmented images, which would translate into failures for downstream tasks.

Image-to-image: This type of augmentation frequently results in a significant loss of subject detail, akin to the image-to-text method. We see that the bird is very hard to spot in the augmented image, which might make the downstream object detection task tougher. The prairie chicken in the example appears to have been transformed into a parrot.

Inpainting: This method modifies the image within a masked area based on a text prompt but can inadvertently corrupt the subject's appearance. We see that the orange/red identifying ring is misplaced in one of the augmented images, thus this method can corrupt the subject.

In contrast, our proposed approach does not add any artifacts to the subject image. It successfully generates augmented images with diverse backgrounds while preserving the authenticity of the subject as seen in the augmented images by AGA . This form of data augmentation has the potential to translate into better performance in terms of fine-grained classification, generalizability, and explainability. Our proposed approach is presented in the next section.

4. Automated Generative Data Augmentation

In this section, we present the methodology of the AGA framework. The input to the framework is an image and the corresponding class name. The output is an augmented image based on the provided inputs. The framework augments an image in three main steps: i) subject isolation through *masked image generation*, ii) the generation of *domain-specific captions* for diverse backgrounds, and iii) augmented image editing for combining the foreground and background. An overview of the AGA framework is shown in Figure 2.

4.1. Masked Image Generation

This step deals with isolating the subject of an input image from its background. In general, such subject masks are not readily available beside the image and class name. Therefore, dense mask estimation models can be used to correctly generate pixel-level masks for subjects using the image and text (class name) only.

AGA includes Segment Anything Model (SAM) [29], one of the SOTA image segmentation tool, for this purpose. SAM is capable of segmenting the subject from an image based on some guiding inputs such as single or multiple point locations on an object, or the object's bounding box, to create precise segmentation masks. As the training dataset usually does not include the point locations or bounding box for the subject, object detection models can be utilized to generate the boxes in this regard. Bounding

boxes provide approximate spatial locations of objects of our interest in the image.

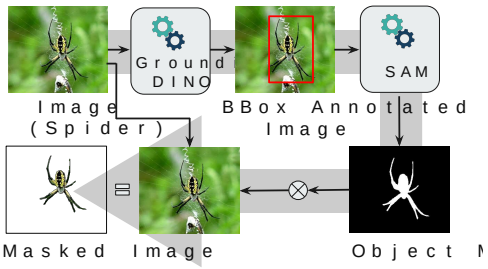


Figure 3. Masked image generation process diagram

There are several SOTA object detection models available, like YOLO [34], GoundingDINO [30]. In AGA workflow, GroundingDINO model is used to generate bounding box due to its superior performance. Empirical analysis shows that for fine-grained text prompts, GroundingDINO often fails to provide optimal bounding box results. For instance, when attempting to locate the bounding box for a specific bird class such as `water ouzel`, the hierarchical naming of the class text `bird` proves more effective than the fine-grained class name. Therefore, AGA utilizes superclasses as text prompts to provide clearer input instructions to the GroundingDINO for object bounding box creation.

The details of proposed mask generation process is shown in Figure 3. GroundingDINO generates the bounding box, indicated in red, for the subject of interest, which in this case is the spider. This bounding box is then feed to the SAM to guide itself to produce the segmentation mask. Once the mask is obtained, the masked image of the subject is created by combining the original image with the segmentation mask.

4.2. Domain Caption Generation

In the AGA pipeline, the generation of domain captions is a crucial task, as it directly influences the diversity of the background images produced. These captions are automatically generated through a two-step process using a prompt generation engine. Initially, the engine samples from three predefined sets: the instruction set (Ins), the background set (Bgr), and the temporal modality set ($Temp$) as the prompt fixers. The instruction set ensures the prompt begins with an appropriate command, the background set introduces spatial diversity, and the temporal set enriches the prompt with times of the day and seasons.

A SOTA LLM, Llama, is employed to transform these engineered prompts into detailed captions that guide the vision diffusion model in generating the background images. Furthermore, a list of words to avoid is incorporated to refine the output, ensuring the prompts remain focused and relevant. The words to avoid include the class names or subject of the image dataset to be augmented, as those subject might corrupt the background prompt. This structured approach ensures each dataset receives tailored prompts, enhancing the resulting image diversity, which is described in Figure 4. Moreover, each part of the prompt results in a combinatorial increase in diversity. This reduces the number of prompt samples that are required to be provided for each category of instructions.

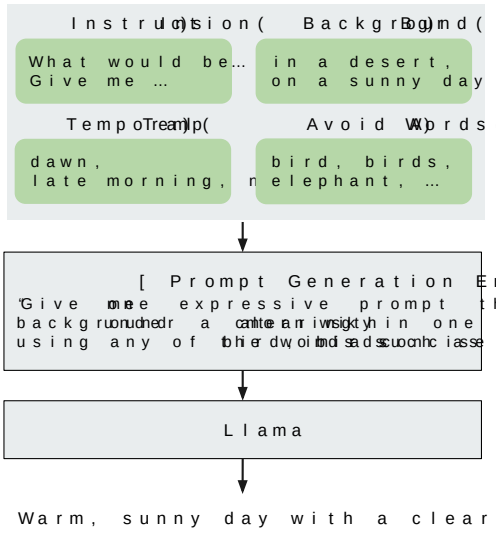


Figure 4. Prompt generation for background diversity.

4.3. Augmented Image Generation

In this step, utilizing the masked image obtained from Section 4.1 and the background caption prompt from Section 4.2, AGA generates a new image with an altered background. The caption prompt serves as the input for a large vision model, which is responsible for creating the background image. Among several text-to-image generation models available, such as DALL-E [35], Imagen [28], and Stable Diffusion [12], AGA employs the Stable Diffusion model for this purpose. Once we have both the masked image and the background image produced by the vision model, AGA proceeds to create the new augmented image. The merging technique used ensures that the background image fills all areas with the masked image, except where the subject is located. This method allows the subject to remain prominent against the newly generated background.

Additionally, to enhance diversity without altering the semantic content of the image, AGA applies traditional affine transformations to the masked image prior to merging. These transformations include flipping (f), rotating (θ), and scaling (s) the subject. Figure 5 illustrates these image editing processes and the respective transformations, showcasing how they contribute to the diversity of the final augmented image while preserving its original meaning.

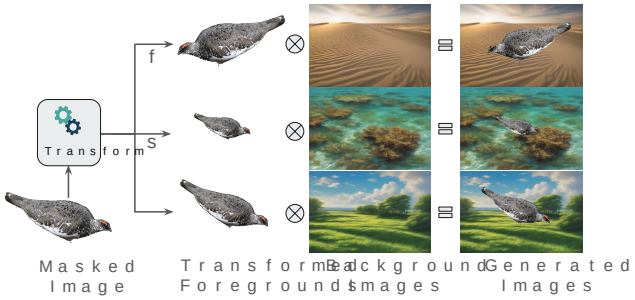


Figure 5. Merging image mask with the generated backgrounds while utilizing affine transformations.

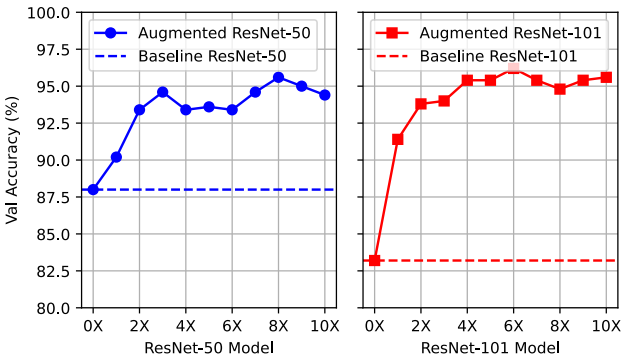


Figure 6. The figure shows the top-1 accuracy with respect to the degree of data augmentation (AGA) for the ResNet-50 and ResNet-101 models on ImageNet10. It can be observed that the accuracy rapidly improves in the beginning while showing an overall upward trend until 10 times augmentation.

5. Experimental Evaluation

We implement the AGA framework in Python, utilizing open source APIs for machine learning models. The implementation runs on a machine equipped with an NVIDIA A100 graphics card. The following sections provide detailed descriptions of the dataset preparation, evaluation setup, and key findings.

Setup. We created a subset of the ImageNet [31] dataset, named ImageNet10, by randomly selecting 10 classes. This subset comprises $13,046$ and 500 training and validation images across the following classes. We refer to this trainset as the original dataset and generate synthetic images from it using the AGA methodology. In subsequent discussions, models described as trained with augmented data refer to those trained using both the original and augmented datasets. Beside this ImageNet10 dataset, we utilize iWildCam [33] dataset, which contains a large collection of global camera trap images. Similarly, we extend our experiments to the CUB [32] dataset, a fine-grained classification set of 200 bird species from Flickr. We maintain the same data distribution ratio as in the previous work'[15] for train and test set to ensure a fair comparison. Detailed dataset descriptions are included in the supplementary materials.

We evaluate the AGA method across two main categories. First, we evaluate our pipeline on in-distribution data using the ImageNet validation set. Second, we assess the robustness of the augmentation method on out-of-distribution ImageNet samples. For this we use the ImageNet variations: ImageNet-Sketch [36] and ImageNet-V2 [37] where ImageNet-Sketch is the sketch version and ImageNet-V2 is the reproduced version of ImageNet respectively. The CUB and iWildCam datasets are used to conduct a comparison study with the previous work [15]. We consider two types of models for our experiments: those trained with original image data and those trained with augmented image data. For comparison, we maintain baseline hyperparameters while augmenting the original training data with augmented data at various scales.

We also compare with other augmentation techniques from recent times: (1) MixUp [19], a data augmentation technique improves deep learning model generalization by creating virtual training examples through convex combinations of original data points and labels, to enhance model robustness. (2) CutMix [20], an approach that creates mixed samples by randomly slicing and combining patches from multiple training images. (3) RandAugment [21], simplifies data augmentation by reducing the search space for augmentation strategies, automating the selection of operations. (4) ALIA [15] analyzes training images to identify diverse background captions, then uses this information to create variations of the images with different backgrounds and contexts. (5) DA-Fusion [23] use diffusion models to generate diverse, high-quality samples, improving model performance by synthesizing realistic variations of training data. (6) Beyond Generation [22] utilize text-to-image models for object detection and segmentation, enhancing visual understanding tasks. In addition to these experiments, we include the results of explainability enhancements for machine learning models using AGA augmented data.

Implementation. We employ ResNet variants 18, 50, 101, 152 as the classification models for training. We train these CNN models from scratch using PyTorch's standard training script [38] which includes PyTorch's default hyperparameter set [39]. AGA utilize a Llama-2-13B-GPTQ from Hugging-Face [40] to create background image caption prompts. These prompts are generated for each image using our prompt engineering method outlined in Section 4.2. Background images are then generated using the Stable Diffusion XL [7] text-to-image model from Hugging-Face, with default hyperparameters. The prompt generation engine operates with three distinct modality sets: an instruction, spatial, and temporal modality set size of 3, 18, and 13 respectively. Supplementary materials include additional training and hyperparameter descriptions.

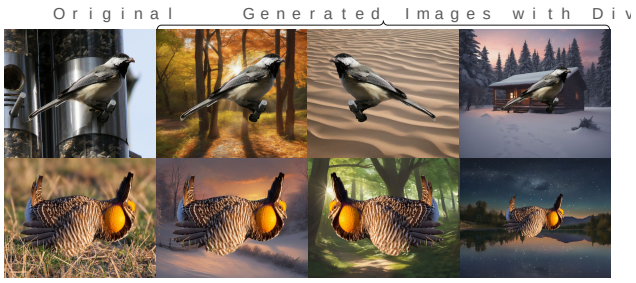


Figure 7. The figure displays the original image samples from ImageNet10 and the generated images using AGA . The generated images effectively preserve the authenticity of the subjects while exhibiting diverse backgrounds. Additional synthetic images are shown in supplementary element.

The remainder of this section is organized, as follows: We first evaluate the effectiveness of AGA on fine-grained image classification in Section 5.1. Next, the generalizability is evaluated in Section 5.2. Lastly, the impact on the explainability is evaluated in Section 5.3. Supplementary materials contain additional evaluation results.

5.1. Fine-grained Image Classification

Accuracy vs. Degree of Data Augmentation: We first evaluate the improvement in classification accuracy with respect to the amount of data augmentation in Figure 6. The figure shows the classification accuracy of a ResNet-50 and a ResNet-101 model on the ImageNet10 dataset. It can be observed that the classification accuracy rapidly improves for data augmentation in the range of 1X to 3X. After that, there are still average improvements but not as significant. In contrast to prior work by Azizi et al. [13], which reported performance degradation in ResNet-50 classification accuracy when the size of augmented data exceeded four times the original dataset. The results indicate that with AGA, the performance of both ResNet-50 and ResNet-101 models increases with the scale of synthetic data augmentation. Specifically, we scale up to ten times the original size of the ImageNet10 dataset, which contains approximately 13,000 images. As illustrated in Figure 6, validation accuracy trends upwards as the dataset size increases, without the performance degradation observed in the prior study. This suggests that AGA does not compromise baseline performance, even at high augmentation scales.

Comparison with SOTA: We now turn our attention to comparing the performance of AGA with previous approaches to data augmentation on the ImageNet10, iWild, and CUB data sets, which is shown in Figure 8. The figure shows the performance of AGA with data augmentation up to 2X, MixUP [19], CutMix [20], RandAug [21], DA-Fusion [23], Beyond Generation [22], and ALIA [15]. Recall that former three methods are traditional data augmentation methods while the latter three are based on generative AI. We only show results of RandAug and ALIA on iWild and CUB because the source code cannot easily be executed on ImageNet10 dataset.

Figure 7 displays samples of images generated by AGA for ImageNet10. The figure shows that AGA successfully augments the input image with diverse backgrounds while preserving the properties of the foreground subject. We compare four ResNet models validation accuracy when the models are trained with (1) real images and (2) augmented dataset at various scales. Figure 8 reports validation results of the respective models for AGA, along with CutMix and MixUp on the ImageNet10 chart. While CutMix, MixUp, and the baseline rely solely on original ImageNet10 images, our study extends to include augmented data up to two times the original dataset size (denoted as 1X to 2X). All models are evaluated using the same original ImageNet validation dataset. Our findings indicate that AGA consistently outperforms both the baseline and other augmentation techniques across all tested scales.

Following this, we compare the performance of AGA with other augmentation techniques such as CutMix, MixUp, Beyond Generation, DA-Fusion, and ALIA on the CUB and iWild datasets. Using AGA , we generate syn-

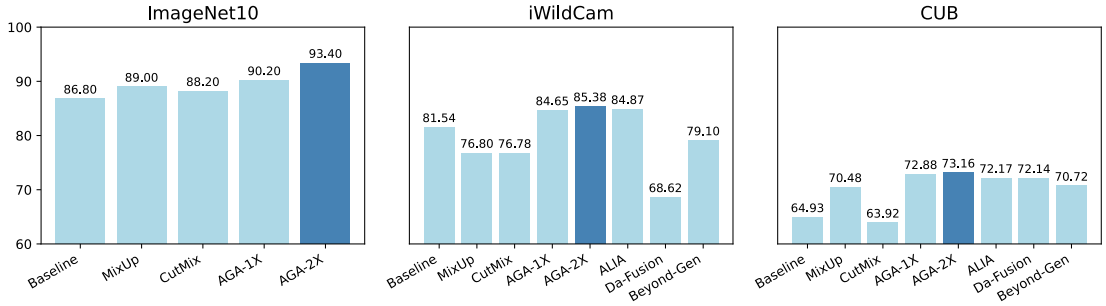


Figure 8. This figure presents bar charts comparing validation performance across the ImageNet10, iWildCam, and CUB datasets. For ImageNet10, the ResNet-50 model is trained from scratch with both original data and augmented data in various scales, shows AGA significantly outperforming other methods. In contrast, for iWildCam and CUB datasets, we employ a pretrained ResNet-50 model as used in ALIA. AGA consistently exceeds the performance of the baseline, CutMix, RandAug, DA-Fusion, and Beyond Generation at both one-time (1X) and two-time (2X) augmentation levels, and surpasses ALIA in the 2X augmentation scenario for iWildCam. For the CUB dataset, AGA again outperforms all competitors for both 1X and 2X augmentations.

Table 1. **Top-1 Accuracy on ImageNet10 and its out-of-distribution variations.** Our approach involves training models with both real ImageNet10 training images and synthetic images generated through our pipeline. We perform an in-distribution evaluation on ImageNet10-val and out-of-distribution evaluation on ImageNet-V2 and ImageNet-Sketch datasets. We directly applied the trained models to these out-of-distribution evaluation datasets without further fine-tuning. The margin of improvement (Δ) over baseline models trained with only real images.

Model	In-Distribution			Out-of-Distribution					
	ImageNet10-Val			ImageNet-V2			ImageNet-S		
	Baseline	AGA	Δ	Baseline	AGA	Δ	Baseline	AGA	Δ
ResNet-18	88.80%	93.40%	4.60	78.43%	89.21%	10.78	33.46%	47.75%	14.29
ResNet-50	86.80%	94.60%	7.80	81.37%	85.29%	3.92	27.78%	50.09%	**22.31**
ResNet-101	78.40%	93.60%	**15.60**	65.69%	89.22%	**23.53**	28.77%	46.77%	18.00
ResNet-152	81.60%	93.80%	12.20	72.55%	88.24%	15.69	27.98%	46.57%	18.59

thetic images at multiple scales and follow the training methodology outlined in ALIA's scripts to enable direct comparison. Figure 8 displays the validation performances on iWildCam, with our method surpassing all others at twice the augmentation scale. In addition to that, Figure 8 also depicts AGA outperforming competing approaches in CUB for both +1X and +2X augmentation scales.

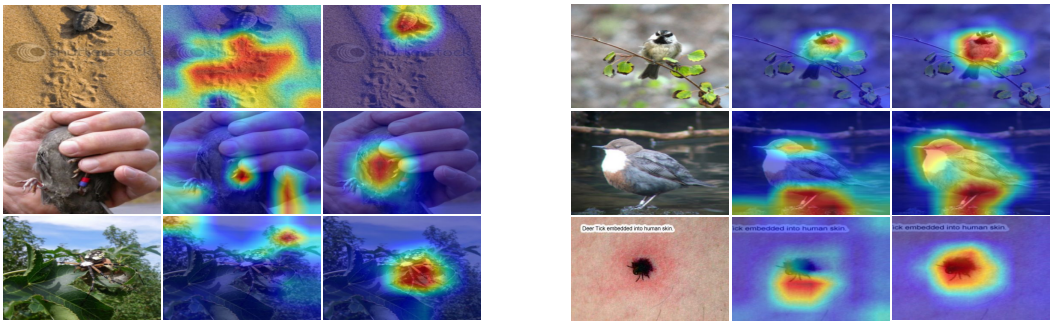
5.2. Evaluation of Impact on Generalizability

Machine learning models typically struggle with out-of-distribution data, but models trained with AGA-augmented data show commendable performance in such cases. We assess our image augmentation method on ImageNet-Sketch and ImageNet-V2 datasets, training the CNN models with both original images and a combination of original and augmented images. For evaluation, we adhere to the same validation dataset across all models. The ImageNet-Val dataset is used for in-distribution testing, while validation data from ImageNet-Sketch and ImageNet-V2 are used for out-of-distribution testing.

Our results are summarized in Table 1, which includes performance metrics for four ResNet models on both in-distribution and out-of-distribution data, with the specific gains over baseline models quantified as Δ in the table. The table highlights up to 15.6% improvements in accuracy for the ResNet-101 model for **ImageNet10-Val** when trained with AGA augmented data. We also see significant performance improvements for out-of-distributions, proving the fact that AGA augmented data increase generalizability of fine-grained classification models.

5.3. Evaluation of Impact on Explainability

Explainability is an increasingly critical aspect of AI, particularly in understanding how machine learning models make decisions. Our study explores the impact of subject-oriented data augmentation provided by AGA on model explainability. By enhancing image diversity through augmentation, we aim to develop more robust and interpretable classifiers. We train models on both the baseline ImageNet10 dataset and augmented data to compare performance. For visualizing how models focus on relevant areas within images, we employ GradCam [41], a tool that highlights significant regions influencing model decisions. In our findings, as shown in Figure 9, we compare models at the 85^{th} epoch, trained solely on real ImageNet10 data and those trained on ImageNet10 augmented data. The model trained only on real data incorrectly classifies three specific images (Figure 9a), whereas the model trained with AGA-



(a) Baseline Model fails to identify samples correctly (b) Both baseline and augmented models correctly identified the sample

Figure 9. This figure shows the impact of the data augmentation on explainability using feature attributions computed using GradCam [41]. (a) images only correctly classified by the classifier trained using data augmentation. (b) images correctly classified by both the original model and the model trained with data augmentation. The model trained with only the original real data fails to identify the bird correctly and focuses on the scatter pixel region. However, it can be observed that even when both models provide the correct classification, the augmented model provides better attributions of the object. More visualization results are presented in supplementary material.

Table 2. Comparison of baseline and AGA-augmented training for four ResNet model variants (18, 50, 101, and 152) using the AIC, SIC, and insertion tests. We perform an in-distribution evaluation on ImageNet10-val and an out-of-distribution evaluation on ImageNet-V2 and ImageNet-Sketch datasets. We directly apply the trained models to these out-of-distribution evaluation datasets without further fine-tuning.

Metric	Model	In-Distribution		Out-of-Distribution			
		ImageNet10-Val		ImageNet-V2		ImageNet-S	
		Baseline	AGA	Baseline	AGA	Baseline	AGA
AIC(↑)	ResNet-18	0.717	**0.764**	0.694	**0.824**	**0.576**	0.532
	ResNet-50	0.792	**0.815**	0.779	**0.799**	**0.670**	0.539
	ResNet-101	0.619	**0.843**	0.682	**0.798**	0.524	**0.598**
	ResNet-152	0.599	**0.878**	0.560	**0.865**	0.521	**0.614**
SIC(↑)	ResNet-18	0.745	**0.816**	0.811	**0.880**	**0.681**	0.640
	ResNet-50	0.816	**0.882**	0.792	**0.877**	**0.679**	0.639
	ResNet-101	0.527	**0.866**	0.579	**0.795**	0.437	**0.654**
	ResNet-152	0.557	**0.872**	0.541	**0.860**	0.488	**0.644**
Insertion(↑)	ResNet-18	0.220	**0.299**	0.202	**0.271**	0.173	**0.211**
	ResNet-50	0.235	**0.320**	0.215	**0.290**	0.190	**0.248**
	ResNet-101	0.115	**0.443**	0.115	**0.386**	0.120	**0.291**
	ResNet-152	0.128	**0.446**	0.120	**0.406**	0.131	**0.294**

augmented data correctly identifies these images. GradCam visualizations reveal that the baseline model often focuses on irrelevant pixels, whereas the AGA-trained model more accurately targets pixels within the subject area. This explains that the augmented data helps the model to learn correctly. Further comparisons using images correctly classified by both models (Figure 9b) show that the AGA-augmented model more consistently identifies correct subject areas, underscoring the benefits of diverse training data for improved model accuracy and explainability.

Additionally, we conduct quantitative explainability analysis utilizing performance information curves (PICs) [42], which include two components: the softmax information curve (SIC) and the accuracy information curve (AIC). The PICs serve as a metric to assess model performance relative to the informational content (entropy) present in the input data. The SIC reflects the softmax value for the input's original class, contributes to the model explainability assessment. Moreover, an insertion test [43] was also conducted to gauge model training performance across different methods. Table 2 presents the AIC, SIC, and insertion test outcomes for various ResNet model variants (18, 50, 101, and 152). The AGA-augmented models generally exhibited improved performance across most cases, barring two instances in both AIC and SIC evaluations. This discrepancy can be attributed to the out-of-distribution nature of the ImageNet-V2 and ImageNet-Sketch datasets relative to the models trained on the ImageNet10 dataset and its augmented variant. Augmenting with additional data caused smaller models to struggle with class confusion, reducing performance in certain scenarios.

6. Conclusion and Future Work

We introduce AGA, a novel data augmentation method designed to address data scarcity in fine-grained image recognition. Our approach integrates image segmentation, automated background caption generation, and diffusion-based image synthesis to diversify backgrounds while maintaining the subject's integrity, thus enhancing training datasets for improved fine-grained classification performance, especially in low-data situations. AGA reveals that additional generated data assists the deep learning model in concentrating on the expected subject regions, as evidenced by the Grad-CAM attribution method. The framework also demonstrates strong generalization on out-of-distribution data. AGA experiences compatibility issues concerning proper subjects and backgrounds. It occasionally generates visually inconsistent synthetic images by combining subjects with contextually inappropriate backgrounds. This limitation underscores the potential for future research to explore new methods for generating images that maintain subject integrity while ensuring compatibility with backgrounds. Our code is publicly available via https://github.com/Fazle045/AGA.git.

Acknowledgement

This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, under Award Numbers DE-SC0024428 and DE-SC0023494, and by the Defense Advanced Research Projects Agency (DARPA) under Co-operative Agreement #FA8750-23-2-0501. This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof makes any warranty or responsibility for the accuracy, completeness of any information. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

References

[1] Sheng-Wei Huang, Che-Tsung Lin, Shu-Ping Chen, Yen-Yi Wu, Po-Hao Hsu, and Shang-Hong Lai. Auggan: Cross domain adaptation with gan-based data augmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 1

[2] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019. 1, 2

[3] Yuji Roh, Geon Heo, and Steven Euijong Whang. A survey on data collection for machine learning: a big data-ai integration perspective. *IEEE Transactions on Knowledge and Data Engineering*, 33(4):1328–1347, 2019. 1

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[5] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 1

[6] Yixuan Su, Tian Lan, Yahui Liu, Fangyu Liu, Dani Yogatama, Yan Wang, Lingpeng Kong, and Nigel Collier. Language models can see: Plugging visual controls in text generation. *arXiv preprint arXiv:2205.02655*, 2022. 1

[7] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 1, 5

[8] Axel Sauer, Kashyap Chitta, Jens Müller, and Andreas Geiger. Projected gans converge faster. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 17480–17492. Curran Associates, Inc., 2021. 1

[9] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794. Curran Associates, Inc., 2021. 1, 2

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020. 1

[11] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 1

[12] 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. 1, 2, 4

[13] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J Fleet. Synthetic data from diffusion models improves imagenet classification. *arXiv preprint arXiv:2304.08466*, 2023. 1, 2, 6

[14] 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. *arXiv preprint arXiv:2108.01073*, 2021. 1, 2

[15] Lisa Dunlap, Alyssa Umino, Han Zhang, Jiezhi Yang, Joseph E Gonzalez, and Trevor Darrell. Diversify your vision datasets with automatic diffusion-based augmentation. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2, 5, 6

[16] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11461–11471, 2022. 1, 2

[17] Zhuoran Yu, Chenchen Zhu, Sean Culatana, Raghuraman Krishnamoorthi, Fanyi Xiao, and Yong Jae Lee. Diversify, don't fine-tune: Scaling up visual recognition training with synthetic images. *arXiv preprint arXiv:2312.02253*, 2023. 1, 2

[18] Hritik Bansal and Aditya Grover. Leaving reality to imagination: Robust classification via generated datasets. In *ICLR 2023 Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models*, 2023. 1

[19] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 2, 5, 6

[20] 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/CVF International Conference on Computer Vision (ICCV)*, October 2019. 2, 5, 6

[21] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 2, 5, 6

[22] Yunhao Ge, Jiashu Xu, Brian Nlong Zhao, Neel Joshi, Laurent Itti, and Vibhav Vineet. Beyond generation: Harnessing text to image models for object detection and segmentation, 2023. 2, 5, 6

[23] Brandon Trabucco, Kyle Doherty, Max Gurinas, and Ruslan Salakhutdinov. Effective data augmentation with diffusion models, 2023. 2, 5, 6

[24] 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*, pages 8748–8763. PMLR, 2021. 2

[25] 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? *arXiv preprint arXiv:2210.07574*, 2022. 2

[26] Shaobo Lin, Kun Wang, Xingyu Zeng, and Rui Zhao. Explore the power of synthetic data on few-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 638–647, 2023. 2

[27] Mert Bülent Sarıyıldız, Karteek Alahari, Diane Larlus, and Yannis Kalantidis. Fake it till you make it: Learning transferable representations from synthetic imagenet clones. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8011–8021, 2023. 2

[28] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 2, 4

[29] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2, 3

[30] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 2, 4

[31] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3, 5

[32] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 3, 5

[33] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pages 5637–5664. PMLR, 2021. 3, 5

[34] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 4

[35] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR, 18–24 Jul 2021. 4

[36] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019. 5

[37] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019. 5

[38] TorchVision maintainers and contributors. Torchvision: Pytorch's computer vision library. https://github.com/pytorch/vision, 2016. 5

[39] How to Train State-Of-The-Art Models Using TorchVision's Latest Primitives — pytorch.org. https://pytorch.org/blog/how-to-train-state-of-the-art-models-using-torchvision-latest-primitives/#baseline. [Accessed 24-04-2024]. 5

[40] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022. 5

[41] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 7

[42] A. Kapishnikov, T. Bolukbasi, F. Viegas, and M. Terry. Xrai: Better attributions through regions. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4947–4956, Los Alamitos, CA, USA, nov 2019. IEEE Computer Society. 8

[43] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018. 8