Domain-Compatible Synthetic Data Generation for Infrequent Objects Detection

Negin Sokhandan, Ninad Kulkarni, Yash Shah, and Suchitra Sathyanarayana

Amazon Web Services AI
{ngnsl, ninadkul, syash, suchisat}@amazon.com

Abstract. The recent advances in generative models have resulted in massive progress in the quality of the generated images to the point that in many cases they cannot be easily distinguished from real images. Despite this quality improvement, using AI generated images for the purpose of training robust down-steam computer vision models for real-world applications has proven to be very challenging. The AI generated images usually lack the required diversity and scene complexity that is crucial for many real-world applications, specifically the ones with safety concerns. The difficulty of this challenge grows significantly when the underlying application involves detection of some specific objects that appear with critically low frequency in the available real datasets. This paper studies a new approach for generating diverse, complex and domain-compatible synthetic images for detecting infrequent objects by employing a diffusion-based generative model pretrained on a generic dataset. More specifically, the impact of using the generated synthetic images with the proposed approach in solving the real world problem of detecting emergency vehicles in road scenes is investigated. Furthermore, the challenges of generating synthetic datasets with the proposed approach will be thoroughly discussed.

1 Introduction

Detection of some domain specific and infrequent objects can be a crucial part of many computer vision based systems. An example of such scenarios is the detection of emergency vehicles for an autonomous driving car application. Since the number of images containing the specific objects of interest in the available datasets is critically limited, generating supplementary synthetic images is a viable solution for training robust downstream object detection models. Employing deep generative models to generate synthetic images for training downstream models in a real-world application imposes some key challenges listed as follows:

Insufficient training samples for the generative model A deep generative model relies on a large training dataset covering different varieties of the object of interest to be able to generate realistic images. In the case of infrequent objects, the lack of sufficient training images is the reason synthetic images are required in the first place.

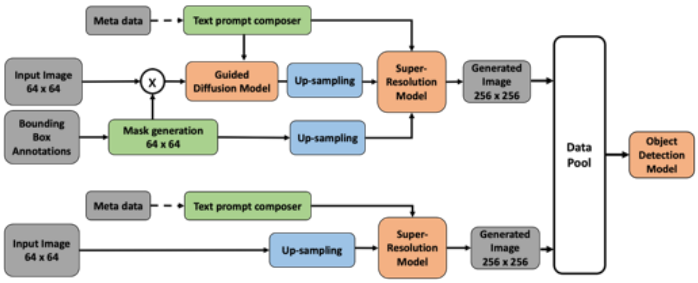

Fig. 1. Block diagram of the architecture of the proposed approaches.

Insufficient diversity and scene complexity The majority of of recent advancements in improving the performance of generative models have been focused on enhancing the quality of the generated images and making them more photo-realistic. The AI-generated images usually lack the required scene complexity and diversity which is essential for training robust downstream models [1].For the same reason there is normally a distribution shift between the generated images and the real ones in terms of complexity and diversity [9].

Generated images may require labeling As opposed to synthetic images generated by rendering engines, AI-generated images may require an annotation process to be ready for a real application.

In order to tackle the above challenges and generate synthetic images that can be effectively used in real-world applications, in this paper we investigate three different approaches of using a generative model that has been only trained on a generic dataset. The proposed approaches can be used to generate a large, complex and widely diverse dataset from a small relevant real dataset. We use a diffusion-based model [8][10][4] that can be conditioned on different information and be partially masked during the generative process to make carefully controlled changes in the real images in a systematic way. This allows the generation of a sufficiently large domain-compatible dataset that covers the required variety and complexity for training a robust downstream model. Since the proposed approach uses real images as the basis to create the synthetic images, there is no domain-shift between the generated images and the real dataset. Conditioning the generative process on a set of guiding text prompts as well as partially masking specific parts of the image during the process allows imposing a customized level of diversity while maintaining the domain characteristics and scene complexity of the real images. The proposed approach also allows either preserving the available annotations or automatically generating new annotations for the synthetically generated objects. We run several experiments to extensively assess the performance enhancement that the generated images provide to the final downstream object detection models.

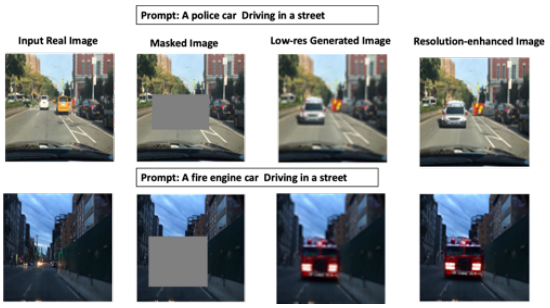

Fig. 2. A few examples of input and output endpoints for Approach 1 .

2 Related Work

One of the most commonly used approaches to generate synthetic image data is through the use of photo-realistic 3D physics engines[13] [3]. These engines can be used to render images from 3D computer-aided design (CAD) models of the target objects. The photo-realism achieved through these image rendering engines has reached a point where synthetic images can be hardly distinguished from real ones [7]. However, there are some drawbacks to these synthetic data generation approaches that make them unsuitable for many practical applications. These include, but are not limited to, requiring 3D asset development, challenges in tuning design parameters (e.g. brightness) and lack of the required diversity and complexity in the image background. Deep generative models including generative adversarial networks (GANs) have been vastly studied for synthetic image generation and synthetic augmentation [15][5]. In the field of medical imaging, GAN-based data augmentation has particularly been used to improve sensitivity and specificity of models tried on small medical imaging datasets by 5-7% [2][5]. Class imbalance has been addressed by generating additional examples of infrequent samples through adversarial autoencorders, a GAN variant [11].More over, deep learning based style transfer has shown 2% improvements in classification accuracy over traditional augmentation strategies [16]. Style transfer, in particular, is capable of preserving image content while copying the style of a separate, unrelated image [6]. Denoising diffusion models were initially introduced by [14]. Recent work has demonstrated the ability of diffusion models to compete and potentially outperform traditional generative adversarial networks in realistic image generation and producing synthetic results indistinguishable from real images to human evaluators in some cases [4][17].

3 Methodology

In the proposed methodology for syntehtic image generation, first a pretrained diffusion model (Dhariwal and Nichol 2021) (Nichol et al. 2021) is fine-tuned on

Fig. 3. An examples of the steps of Approach 2 .

a generic dataset which does not necessarily include the infrequent target objects (we used a generic driving dataset (Yu et al. 2020)). In order to condition the diffusion process on text, we use a CLIP model (Radford et al. 2021) that perturbs the denoising process mean with the gradient of the dot product of the image and text encoding with respect to the image. Next, we explore three different image manipulation approaches with this model that allows generating synthetic images that contain a large variety of infrequent objects of interest. These synthetic images are then used for training downstream object detection models as shown in Figure 1. Finally, a text-conditioned super-resolution diffusion model is cascaded with the generative model in the pipeline to increase the resolution of the generated images. The proposed approaches are based on the assumption that a very small but domain-relevant real dataset is available and synthetic images are generated by manipulating those real images. In fact, using this small real data as the basis is essential in keeping the generated images in the target domain. In this section, the three proposed image manipulation approaches will be explained in detail.

3.1 Approach 1: Synthetic Infrequent Objects in a Real Background

The idea behind this approach which is depicted in the upper part of Figure 1, is to generate instances of the infrequent objects of interest inside a background sampled from the real dataset to maintain the generated images in the same domain as the real dataset. The importance of this approach is that it can be employed to generate a sufficiently large synthetic dataset even if the real dataset does not include any images containing the infrequent target objects. The architecture of this approach consists of four main components: A mask generator block, a text prompt composer unit, a text guided diffusion generative model and a super-resolution model. The input image serving as background and the corresponding annotations are first fed to a mask generator block which proposes a mask based on the current bounding boxes in the image. The generated mask is then applied to the original image and the resulted masked image is fed to the text conditioned diffusion model. The diffusion model iteratively manipulates the masked part of the image following the input text prompt guidance until it generates an instance of the target object inside the masked section which is well blended with the background. The output of this model is then fed to a

Fig. 4. Changing the weather condition in the real image using Approach 3 .

diffusion-based super-resolution model (Nichol and Dhariwal 2021) to enhance its resolution. The super-resolution model can also be conditioned on the text prompt for improved enhancement. Figure 2 illustrates a few examples of the inputs and output endpoints of the pipeline of this approach. In the rest of this subsection, the mask generator and prompt composer blocks are described.

Mask generator block This block proposes a region for masking the input image based on the available bounding boxes in the annotations. In order to find a proper area for the placement of the target object, one or more adjacent bounding boxes are randomly picked and merged together to make a target bounding box while the following rules are met:

- The proposed bounding box should not cut any of the other bounding boxes to avoid unrealistic coincidences between the generated objects and the ones in the background.
- If needed, the orientation of the bounding box should be compatible with the required object alignment. Usually the orientation of the bounding box dictates the orientation of the generated object and can be used as an additional factor for randomization.

Other customized rules can be integrated depending on the target application.

Text prompt composer unit This block composes a text prompt to guide the diffusion process toward generating the desired target image. Each composed prompt consists of five main components as follows:

Subject In approach 1, subject is randomly sampled from the list of infrequent target objects.

Verb Verb is randomly sampled from a list of possible actions relevant to the target object. For example for a driving scene dataset, the possible verbs can be driving, crossing, parking, etc.

Location Represents the location of the target object in the image and it can be either extracted from meta data (approach 1) or randomly sampled from possible options (approach 2).

Condition This field describes a global condition for the image. For example for a road scene dataset this field can describe the weather condition, e.g. rainy, snowy, foggy, etc.

Time Optionally describes the time of day, e.g. morning, night, sunset, etc.

Fig. 5. Examples of demonstrating challenges with text condition image generation.

3.2 Approach 2: Real Infrequent Objects in Synthetic Background

This approach can be also represented by the top part of 1. However instead of generating target objects in a real background, it generates a synthetic background for a real target object. The target object is first cropped from a real image and after random resizing is placed in a random position in a blank (all zeros) background. The resulted combinations is then fed to the diffusion model. There are two important differences between this approach and approach 1:

1. As opposed to approach 1, in this approach the mask only covers the real object and leaves everywhere else in the image available for the diffusion model's generative manipulation. This results in the generation of a background that follows the text prompt guidance and blends well with the real object.
2. In this approach, the prompt composer unit randomly samples all of the background-related fields such as verb, location, condition and time from the the corresponding lists that are provided to the module based on the target application. The only field that will be extracted from the annotation is the type of target object that has been cropped from the real image.

Figure 3 illustrates the steps of this approach in an example.

3.3 Approach 3: Real Images Globally Altered

The third approach is represented by the bottom part of the block diagram in Figure 1. In this approach, certain aspects of the real images are altered as they are converted from low to high resolution by conditioning the super-resolution model to text prompts that guide the diffusion process toward those modifications. As suggested by the diagram, in this approach no masking is required as the entire input image is subject to the model's subtle modifications. In order to propose suitable text prompts for randomized modifications to input images, the text composer unit randomly samples the condition field from a list of application-relevant conditions while rest of the fields are extracted form the annotations or meta-data if it is available. For example, multiple altered

versions of an input real image can be generated synthetically by randomizing on weather condition or the time of the day. Figure 4 shows some examples of these modifications along with their corresponding text prompts.

4 Dataset

In this section, we introduce the real dataset that was used as a base for generating the synthetic images in all of our experiments. The LISA-Amazon Vehicle and Scene Attributes (LAVA) dataset [12] has been collected as a part of a collaboration between the Amazon Machine Learning Solutions Lab with the Laboratory of Intelligent and Safe Automobiles at the University of California, San Diego (UCSD) to build a large and richly annotated driving dataset with fine-grained vehicle, pedestrian, and scene attributes. The LAVA dataset is annotated for all types of vehicles, traffic signs, traffic lights and pedestrians with 2D bounding boxes, class labels and some meta data. A subset of the LAVA dataset that covers all the images with emergency vehicles in them (in addition to other vehicles) was separated and used for generating synthetic images and training the downstream object detection models. We refer to this subset as LAVA-emergency dataset. Table 1 shows the class distribution of the train and test splits of the LAVA-emergency dataset. It is essential to reserve a reasonable portion of the real dataset for testing to be able to reliably evaluate the impact of synthetic data generation approaches.

Table 1. Distribution of images and bounding boxes for real and synthetic datasets.

Dataset	Num. images	Medical	Fire	Police
Real-Train	215	47	42	126
Real-Test	539	270	68	215
Type-1	1876	447	569	939
Type-2	1875	620	306	949

5 Experiments

5.1 Experimental Setup

For all of the experiments in this section, the LAVA-emergency dataset is used as a base for generating synthetic images using the approaches in section 3. The downstream task in our experiments is the detection of emergency vehicles including medical vehicles (ambulances), fire engines and police cars. These emergency vehicles appear with a critically low frequency in the road-scene datasets. For better understanding of the evaluation results, we group the synthetic data generation techniques into three general types. Type-1 (S1), represents the approaches wherein the emergency vehicles themselves are synthetically generated

Table 2. Downstream object detection performance for each dataset.

Model and backbone	Dataset	Num. train images	mAP@0.50:0.95	mAR@0.50:0.95
SSD ResNet101 V1 FPN	R	215	0	0.028
SSD ResNet101 V1 FPN	R, S1	2091	0.147	0.441
SSD ResNet101 V1 FPN	R, S2	2090	0.396	0.59
SSD ResNet101 V1 FPN	R, S1, S2	3966	0.372	0.586
SSD MobileNet V1 FPN	R	215	0	0.095
SSD MobileNet V1 FPN	R, S1	2091	0.129	0.331
SSD MobileNet V1 FPN	R, S2	2090	0.475	0.637
SSD MobileNet V1 FPN	R, S1, S2	3966	0.357	0.583
EfficientDet D1	R	215	0.053	0.439
EfficientDet D1	R, S1	2091	0.136	0.523
EfficientDet D1	R, S2	2090	0.368	0.594
EfficientDet D1	R, S1, S2	3966	0.458	0.641
Faster RCNN Inception ResNet V2	R	215	0.173	0.451
Faster RCNN Inception ResNet V2	R, S1	2091	0.454	0.723
Faster RCNN Inception ResNet V2	R, S2	2090	0.521	0.695
Faster RCNN Inception ResNet V2	R, S1, S2	3966	0.494	0.714

(only Approach 1). Type-2 (S2) represents all the approaches wherein the emergency vehicles are real but they have been placed in a synthetically generated or modified background (approach 2 and approach 3). Table 1 shows the distribution of generated data over different emergency vehicles categories. In these experiments, for composing the text prompts, the weather condition is randomly and uniformly sampled from a list of 5 weather conditions namely, sunny, rainy, snowy , foggy and cloudy. The location of the vehicle is randomly sampled from one of four options: street, road (each with a probability of 0.35), parking (with a probability of 0.25) and bridge (with a probability of 0.05). Each synthetic image is generated by applying 100 diffusion steps to the masked real input image (in Approach 1 and 2). The resolution of the generated images is then enhanced by applying 30 addition diffusion steps through the super-resolution model. Each experiment uses either only real data (R) or a combination of it with one or more types of synthetic images. The objective of these experiments is to evaluate how each of the synthetic data generation approaches improves the performance of the downstream object detection models when combined with the real data.

5.2 Results

Table 2 shows the performance of various object detection algorithms trained on difference combinations of real and synthetic images on the emergency-LAVA test set. As shown in this table the single-stage detectors such as different flavors of SSD and EfficientDet are barely able to learn anything from the small real training set. However, incrementally adding synthetic images to augment the real training images remarkably improves the detector's performance on the real test set. The EfficientDet D1 model has monotonically increasing mAP and mAR as more synthetic data is added. For SSD ResNet101, SSD MobileNet and Faster R-CNN models there is a considerable performance improvement when trained on R, S1 or R, S2 compared to when they are only trained on R. However, for

these models, there is a slight drop in performance when they are trained on R, S1, S2 compared to when they are trained on R, S2. As mentioned in section 3.1, in synthetic Type-1 images the emergency vehicles themselves are generated by the model and the generator model has been trained on a generic dataset which contains vehicles from a variety of different countries in the world. The LAVA-emergency test set however contains only emergency vehicles from Southern California, and thus the discrepancy in performance when involving S1 in training along with S2 can be explained by the change in emergency vehicles characteristics from different geo-locations. However, in S2 images, the emergency vehicles have been directly adopted from emergency-LAVA training set and they are compatible with the emergency vehicles in the test set. Therefore, increasing the number of Type-2 images always improves the performance of all of the object detection models.

5.3 Practical Challenges

Although the synthetically generated images by the proposed approaches are realistic and diverse, there are a few challenges that need to be considered depending on the target application as follows:

Relative size of the objects When an image generation process is conditioned on text, sometimes the relative sizes of the generated objects can be slightly out of proportionate with respect to the background objects, regardless of the type of the generative model. While some downstream vision tasks such as object detection are not negatively impacted by this, some others may be impacted. The top row of Figure 1 shows a few examples with slightly disproportionate objects.

The number of the objects One of the concepts that normally do not transfer properly between language and vision spaces is the exact quantity of objects. Similar to the previous case, the exact number of objects does not impact many of the vision tasks (e.g. object detection).

The relative position of the objects Similar to relative sizes of objects, their relative positions with respect to each other can sometimes be unrealistic when the generative process is conditioned on text. The bottom row of Figure 5 shows a few examples impacted by this effect.

6 Conclusions

In this work, a new approach for generating synthetic data for training downstream models in a critically low data regime was studied. The experimental results showed that employing the synthetic images generated by the proposed approach significantly improved the performance of all of the investigated object detection models. Employing approaches similar to the proposed approach to augment insufficiently small real datasets used in training the downstream computer vision models is specifically crucial for applications with safety concerns.

References

1. Block, D., Teliban, I., Greiner, F., Piel, A.: Prospects and limitations of conditional averaging. Physica Scripta **2006**(T122), 25 (2006)
2. Bowles, C., Chen, L., Guerrero, R., Bentley, P., Gunn, R., Hammers, A., Dickie, D.A., Hernández, M.V., Wardlaw, J., Rueckert, D.: Gan augmentation: Augmenting training data using generative adversarial networks. arXiv preprint arXiv:1810.10863 (2018)
3. Cisse, M., Adi, Y., Neverova, N., Keshet, J.: Houdini: Fooling deep structured prediction models. arXiv preprint arXiv:1707.05373 (2017)
4. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in Neural Information Processing Systems **34** (2021)
5. Frid-Adar, M., Klang, E., Amitai, M., Goldberger, J., Greenspan, H.: Synthetic data augmentation using gan for improved liver lesion classification. In: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018). pp. 289–293. IEEE (2018)
6. Gatys, L.A., Ecker, A.S., Bethge, M.: A neural algorithm of artistic style. CoRR **abs/1508.06576** (2015), http://arxiv.org/abs/1508.06576
7. Hamesse, C., Lahouli, R., Fréville, T., Pairet, B., Haelterman, R.: Training machine learning algorithms for computer vision tasks in difficult conditions: 3d engines to the rescue (2019)
8. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020)
9. Joshi, C.: Generative adversarial networks (gans) for synthetic dataset generation with binary classes (2019)
10. Kim, G., Ye, J.C.: Diffusionclip: Text-guided image manipulation using diffusion models (2021)
11. Lim, S.K., Loo, Y., Tran, N., Cheung, N., Roig, G., Elovici, Y.: DOPING: generative data augmentation for unsupervised anomaly detection with GAN. CoRR **abs/1808.07632** (2018), http://arxiv.org/abs/1808.07632
12. Ninad, K., Akshay, R., Jonathan, B., Jeremy, F., Mohan, T., Nachiket, D., Greer, R., Saman, S., Suchitra, S.: Create a large-scale video driving dataset with detailed attributes using amazon sagemaker ground truth (2021)
13. Pollok, T., Junglas, L., Ruf, B., Schumann, A.: Unrealgt: using unreal engine to generate ground truth datasets. In: International Symposium on Visual Computing. pp. 670–682. Springer (2019)
14. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning. pp. 2256–2265. PMLR (2015)
15. Vega-Márquez, B., Rubio-Escudero, C., Riquelme, J.C., Nepomuceno-Chamorro, I.: Creation of synthetic data with conditional generative adversarial networks. In: International Workshop on Soft Computing Models in Industrial and Environmental Applications. pp. 231–240. Springer (2019)
16. Zheng, X., Chalasani, T., Ghosal, K., Lutz, S., Smolic, A.: Stada: Style transfer as data augmentation. CoRR **abs/1909.01056** (2019), http://arxiv.org/abs/1909.01056
17. Zhou, S., Gordon, M.L., Krishna, R., Narcomey, A., Morina, D., Bernstein, M.S.: HYPE: human eye perceptual evaluation of generative models. CoRR **abs/1904.01121** (2019), http://arxiv.org/abs/1904.01121