Domain-Compatible Synthetic Data Generation for Emergency Vehicle Detection

Dipika Khullar  
Amazon Web Services AI  
dikhulla@amazon.com

Negin Sokhandan  
Amazon Web Services AI  
ngnsl@amazon.com

Ninad Kulkarni  
Amazon Web Services AI  
ninadkul@amazon.com

Yash Shah  
Amazon Web Services AI  
syash@amazon.com

Suchitra Sathyarayanan  
Amazon Web Services AI  
suchisat@amazon.com

Abstract—Recent advancements in generative models have led to significant improvements in the quality of generated images, making them virtually indistinguishable from real ones. However, using AI generated images for training robust computer vision models for real-world applications, especially object detection in road scene perception, is still a challenge. AI generated images usually lack the required diversity and scene complexity where specific objects appear with critically low frequency in the available real datasets.

An example of such applications is the detection of emergency vehicles like police cars, fire trucks, and ambulances in road scenes. These vehicles appear with drastically low frequencies in available datasets. Successfully generating synthetic images of road scenes that include these types of vehicles and using them in training downstream models would prove useful for autonomous driving vehicles, mitigating safety concerns on the road.

To address this, this paper proposes a new approach for synthetically generating diverse, complex, and domain-compatible images of emergency vehicles in road scenes by employing a diffusion-based generative model pretrained on a generic dataset. We investigate the impact of using generated synthetic images in the performance of downstream object detection models. Finally, we thoroughly discuss challenges of generating synthetic datasets with the proposed approach.

I. INTRODUCTION

Detecting specific and infrequent objects, such as emergency vehicles in autonomous driving, is crucial for computer vision systems. With limited real images of these objects, generating synthetic images is an effective solution to train object detection models. However, using deep generative models to generate synthetic images for real-world applications faces some challenges, including:

1) 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. If there’s not enough data for rare objects, synthetic images must be used to fill the gap [17].

2) Insufficient Diversity and Scene Complexity. The majority of recent advancements in improving the performance of generative models have been focused on enhancing the quality of generated images and making them more photo-realistic. The AI-generated images usually lack the required scene complexity and diversity essential for training robust downstream models [2]. For the same reason there is normally a distribution shift between generated images and the real ones in terms of complexity and diversity [11].

3) 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 [17].

This paper presents three methods to generate synthetic images using a generative model trained only on a generic dataset, to overcome the challenges faced in using synthetic images for real-world applications. The proposed approaches can be used to generate a diverse and extensive dataset from a limited real dataset relevant to the task.

We use a diffusion-based model [10] [12] [5] that can be conditioned on a set of guiding text prompts as well as partially masking specific parts of the image during the generative process to make carefully controlled changes to 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 generated images provide to the final downstream object detection models.
II. RELATED WORK

One of the most commonly used approaches to generate synthetic image data is through use of photo-realistic 3D physics engines [18] [4]. 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 [8]. 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 [23] [6]. 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% [3] [6].

Class imbalance has been addressed by generating additional examples of infrequent samples through adversarial autoencoders, a GAN variant [13]. Moreover, deep learning based style transfer has shown 2% improvements in classification accuracy over traditional augmentation strategies [25]. Style transfer, in particular, is capable of preserving image
content while copying the style of a separate, unrelated image [7].

Denoising diffusion models were initially introduced by [21]. 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 [5] [26].

III. Methodology

First, a pretrained diffusion model [5] [15] is fine-tuned on a generic dataset which does not necessarily include the infrequent target objects (we used a generic driving dataset [24]). In order to condition the diffusion process on text, we use a CLIP model [19] 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.

A. Approach 1: Synthetic Infrequent Objects in a Real Background

Approach 1, depicted in the upper part of Figure 1, generates instances of infrequent objects of interest inside a background sampled from real data to maintain the generated images in the same domain as the real dataset. This approach 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 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 resulting 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 diffusion-based super-resolution model [15] 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.

1) 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.
Fig. 4. Approach 3 modifies real images during their conversion from low to high resolution. The approach modifies real images by conditioning a super-resolution model with text prompts guiding the diffusion process. No masking is needed as the entire image is subject to modifications. Text prompts for modifications are randomly selected from an application-relevant list and other fields are extracted from annotations or meta-data. Altered versions of the image can be generated by randomizing based on conditions like weather and time of day.

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 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.

C. 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 from 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.

IV. Dataset

In this section, we outline the data used for experimentation. The real train dataset (R) is used for generating the Synthetic Type-1 (S1) and Synthetic Type-2 (S2) images. The augmented dataset (AUG) is created from the real emergency vehicle data using standard augmentation methods and serves as a baseline. Table 1 shows the class distribution of the train and test subsets of the real data as well as the augmented dataset.

A. Real Data (Real-Train, Real-Test)

The LISA-Amazon Vehicle and Scene Attributes (LAVA) dataset [16] has been collected as a part of a collaboration between the Amazon Machine Learning Solutions Lab with
Fig. 5. Examples of practical challenges with text condition image generation. We read the images left to right in each row. Image 1 (row 1, image 1): the vehicle is too large compared to its surroundings. Image 2 (row 1, image 2): ambulance is perpendicular to the street and not parked, this would not happen in real life. Image 3 (row 1, image 3): the fire truck is too small. Image 4 (row 1, image 4): the white police car in the front is smaller than the black police car in the back, making the white car look like a toy. Image 5 (row 2 image 1): the van is in the pedestrian walkway. Image 6 (row 2, image 3): the truck in the front is way bigger than the ambulance behind it. Image 7 (row 2 image 3): the two generated police vehicles look too close together. Image 8: we observe a flying car.

### Table I

Datasets Used for Experimentation

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Num. Images</th>
<th>Medical</th>
<th>Fire</th>
<th>Police</th>
<th>Normal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real-Train (R)</td>
<td>5215</td>
<td>47</td>
<td>32</td>
<td>126</td>
<td>5000</td>
</tr>
<tr>
<td>Synthetic Type-1 (S1)</td>
<td>1876</td>
<td>487</td>
<td>576</td>
<td>1028</td>
<td>0</td>
</tr>
<tr>
<td>Synthetic Type-2 (S2)</td>
<td>1875</td>
<td>642</td>
<td>366</td>
<td>1081</td>
<td>0</td>
</tr>
<tr>
<td>Augmented (AUG)</td>
<td>1876</td>
<td>383</td>
<td>372</td>
<td>1121</td>
<td>0</td>
</tr>
<tr>
<td>Real-Test (R-Test)</td>
<td>1539</td>
<td>268</td>
<td>68</td>
<td>203</td>
<td>1000</td>
</tr>
</tbody>
</table>

TABLE II

Augmentation Types Used for Augmentations (AUG) Dataset

<table>
<thead>
<tr>
<th>Augmentation Type</th>
<th>Num. images</th>
</tr>
</thead>
<tbody>
<tr>
<td>Horizontal Flip</td>
<td>157</td>
</tr>
<tr>
<td>Random Brightness Contrast</td>
<td>157</td>
</tr>
<tr>
<td>Random Shadow</td>
<td>157</td>
</tr>
<tr>
<td>MudSpatter</td>
<td>157</td>
</tr>
<tr>
<td>ISONoise</td>
<td>157</td>
</tr>
<tr>
<td>ToSepia</td>
<td>157</td>
</tr>
<tr>
<td>HorizontalFlipSunFlare</td>
<td>157</td>
</tr>
<tr>
<td>PixelDropout</td>
<td>157</td>
</tr>
<tr>
<td>RainSpatter</td>
<td>157</td>
</tr>
<tr>
<td>RandomToneCurve</td>
<td>157</td>
</tr>
<tr>
<td>Equalization</td>
<td>157</td>
</tr>
<tr>
<td>Blur</td>
<td>149</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>1876</strong></td>
</tr>
</tbody>
</table>

V. Experiments

### A. Experimental Setup

Each experiment uses real data (R), combined with one or more types of synthetic images (S1, S2, or S1+S2) to detect medical, fire, and police emergency vehicles. We benchmark our solution against standard augmentation techniques, as outlined in Table II. The purpose of these experiments is to show how each of the synthetic data generation approaches improves performance of the downstream object detection models when combined with the real data.

For better understanding of the evaluation results, we group the synthetic data generation techniques into three general

Comparing synthetic data with augmented real data provides a baseline for the performance of models trained on synthetic data. Table II show the transformations used in augmented dataset (AUG)
TABLE III
Downstream Object Detection Performance for Each Dataset.

<table>
<thead>
<tr>
<th>Model</th>
<th>Dataset</th>
<th>Num. Train Images</th>
<th>mAP@0.5:0.95 Emergency Vehicles</th>
<th>mAP@0.5:0.95 Normal Vehicles</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSD ResNet101 V1 FPN</td>
<td>R</td>
<td>5215</td>
<td>0.075</td>
<td>0.368</td>
</tr>
<tr>
<td>SSD ResNet101 V1 FPN</td>
<td>R+AUG</td>
<td>7091</td>
<td>0.205</td>
<td>0.589</td>
</tr>
<tr>
<td>SSD ResNet101 V1 FPN</td>
<td>R+S1</td>
<td>7091</td>
<td>0.177</td>
<td>0.604</td>
</tr>
<tr>
<td>SSD ResNet101 V1 FPN</td>
<td>R+S2</td>
<td>7090</td>
<td>0.297</td>
<td>0.609</td>
</tr>
<tr>
<td>SSD ResNet101 V1 FPN</td>
<td>R+S1+S2</td>
<td>8966</td>
<td><strong>0.368</strong></td>
<td><strong>0.639</strong></td>
</tr>
<tr>
<td>EfficientDet D1</td>
<td>R</td>
<td>5215</td>
<td>0.056</td>
<td>0.356</td>
</tr>
<tr>
<td>EfficientDet D1</td>
<td>R+AUG</td>
<td>7091</td>
<td>0.142</td>
<td>0.480</td>
</tr>
<tr>
<td>EfficientDet D1</td>
<td>R+S1</td>
<td>7091</td>
<td>0.180</td>
<td>0.503</td>
</tr>
<tr>
<td>EfficientDet D1</td>
<td>R+S2</td>
<td>7090</td>
<td>0.203</td>
<td>0.518</td>
</tr>
<tr>
<td>EfficientDet D1</td>
<td>R+S1+S2</td>
<td>8966</td>
<td><strong>0.287</strong></td>
<td><strong>0.508</strong></td>
</tr>
<tr>
<td>Faster RCNN Inception ResNet V2</td>
<td>R+AUG</td>
<td>7091</td>
<td>0.256</td>
<td>0.613</td>
</tr>
<tr>
<td>Faster RCNN Inception ResNet V2</td>
<td>R+S1</td>
<td>7091</td>
<td>0.189</td>
<td>0.578</td>
</tr>
<tr>
<td>Faster RCNN Inception ResNet V2</td>
<td>R+S2</td>
<td>7090</td>
<td>0.419</td>
<td>0.577</td>
</tr>
<tr>
<td>Faster RCNN Inception ResNet V2</td>
<td>R+S1+S2</td>
<td>3966</td>
<td><strong>0.434</strong></td>
<td>0.611</td>
</tr>
<tr>
<td>YOLOX V2</td>
<td>R</td>
<td>5215</td>
<td>0.112</td>
<td>0.337</td>
</tr>
<tr>
<td>YOLOX V2</td>
<td>R+AUG</td>
<td>7091</td>
<td>0.157</td>
<td>0.340</td>
</tr>
<tr>
<td>YOLOX V2</td>
<td>R+S1</td>
<td>7091</td>
<td>0.229</td>
<td>0.439</td>
</tr>
<tr>
<td>YOLOX V2</td>
<td>R+S2</td>
<td>7090</td>
<td>0.333</td>
<td>0.449</td>
</tr>
<tr>
<td>YOLOX V2</td>
<td>R+S1+S2</td>
<td>8966</td>
<td><strong>0.418</strong></td>
<td><strong>0.428</strong></td>
</tr>
<tr>
<td>Deformable DETR</td>
<td>R</td>
<td>5215</td>
<td>0.427</td>
<td>0.444</td>
</tr>
<tr>
<td>Deformable DETR</td>
<td>R+AUG</td>
<td>7091</td>
<td>0.440</td>
<td>0.574</td>
</tr>
<tr>
<td>Deformable DETR</td>
<td>R+S1</td>
<td>7091</td>
<td>0.515</td>
<td>0.557</td>
</tr>
<tr>
<td>Deformable DETR</td>
<td>R+S2</td>
<td>7090</td>
<td>0.642</td>
<td>0.553</td>
</tr>
<tr>
<td>Deformable DETR</td>
<td>R+S1+S2</td>
<td>8966</td>
<td><strong>0.662</strong></td>
<td><strong>0.681</strong></td>
</tr>
</tbody>
</table>

types. Type-1 (S1), represents the approaches where the emergency vehicles themselves are synthetically generated (only Approach 1). Type-2 (S2) represents all the approaches where the emergency vehicles are real but they have been placed in a synthetically generated or modified background (Approach 2 and approach 3). Table I 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 additional diffusion steps through the super-resolution model.

B. Results

Table III shows how our synthetic data generation technique improves the performance of object detectors in comparison to conventionally augmented datasets. More precisely, the results show that adding the combined synthetic data (R+S1+S2) results in 16% to 20% improvement in the mAP values compared to the conventionally augmented dataset(R+AUG).

The addition of synthetic data improves the mAP on emergency vehicles and maintains the performance on normal vehicles. The mAP of all models on R-Test emergency vehicles improves as more synthetic data is added, and for some models such as SSD ResNet101 [14] and EfficientDet D1 [22], the normal vehicle performance also improves. All combinations of synthetic data including S1, S2 and S1+S2 outperform the conventionally augmented data for EfficientDet D1, YOLOX V2 [20], and Deformable DETR [27] models. For the SSD ResNet101 and Faster R-CNN [9] models R+AUG dataset performs slightly better than R+S1. This can be attributed to the geographical differences in emergency vehicles between the generic dataset used to train the synthetic Type-1 dataset and the LAVA-emergency test set, R-Test.

As mentioned in section 3.1, the synthetic Type-1 (S1) emergency vehicles are generated by the generative model trained on a generic dataset containing vehicles from a variety of different countries in the world. The LAVA-emergency test set, R-Test, however contains only emergency vehicles from Southern California, and thus the discrepancy in performance when involving S1 in training compared to R+AUG can be explained by the change in emergency vehicles characteristics from different geographies.

Increasing the number of synthetic Type-2 (S2) images always improves the performance of all of the object detection models. Experimental comparison of R+S1 and R+S2 training shows consistently higher performance for models trained with S2 for all models and backbones.

C. 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 when generating a dataset for a specific application using these approaches. The most common challenges can be listed as follows:

1) Relative Size of Objects

When an image generation process is conditioned on
text, sometimes the relative sizes of the generated objects can be slightly out of proportion 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.

2) Number of 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).

3) Relative Position of 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.

VI. CONCLUSION
In this work, we present a new method for generating synthetic data to train computer vision models in cases of limited real data. Our experiments show that the synthetic data generated through our approach improves downstream object detection for infrequent objects while maintaining performance of the majority class. The synthetic data generation solution in this paper is a practical approach to improving infrequent object performance and is particularly crucial for safety-sensitive applications where real data is limited.

REFERENCES