Synthetic Data Generation for Scarce Road Scene Detection Scenarios

Dipika Khullar
Amazon Titan Labs
dikhulla@amazon.com

Negin Sokhandan
Amazon Generative AI Incubator Center

Ninad Kulkarni
Amazon Titan Labs
ninadkul@amazon.com

Yash Shah
Amazon Generative AI Incubator Center
syash@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.

1 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 [18].
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].

NeurIPS 2023 Workshop on Synthetic Data Generation with Generative AI.

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 [18].

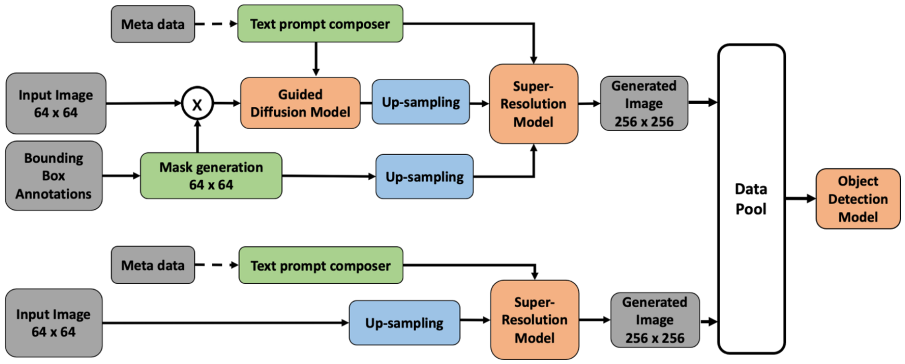

Figure 1: The block architecture diagram shows proposed solutions 1,2, and 3 of the proposed approaches. The top path depicts Approach 1 and 2, masking portions of the image and using the model to fill those in. Approach 1 generates images inside a background sampled from the real data to maintain same domain as the real dataset. Appraoch 2 generates a synthetic background for a real target object. The bottom path, Approach 3, alters parts of the real images 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. No masking is required as the entire input image is subject to the model's subtle modifications

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][15] that can be conditioned on different information and be partially masked 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.

2 Related Work

One of the most commonly used approaches to generate synthetic image data is through use of photo-realistic 3D physics engines[19] [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 [24][6]. In the field of medical imaging,

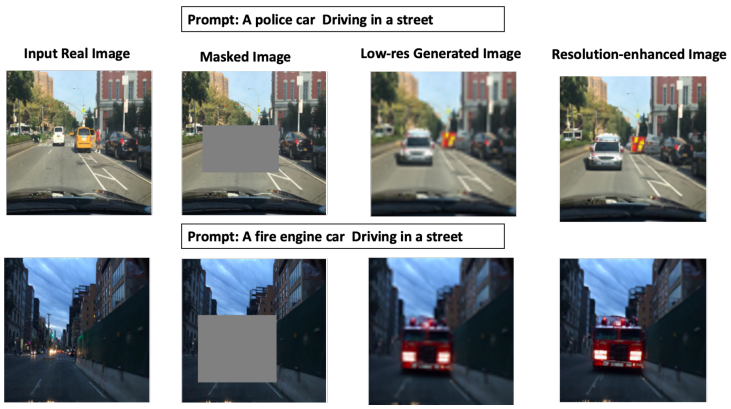

Figure 2: Approach 1 example inputs and outputs. The image and annotations are fed to a mask generator that creates a mask based on the bounding boxes. The masked image is then fed to a text-conditioned diffusion model generating a target object inside the mask that blends with the background. This output is then enhanced by a text-prompt-conditioned diffusion-based super-resolution model.

Figure 3: Approach 2. As opposed to approach 1, the mask here only covers the real object and leaves everywhere else in the image available for the diffusion model's generative manipulation. The prompt composer unit randomly samples all of the background-related fields such as verb, location, condition and time.

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 autoencorders, a GAN variant [13]. Moreover, deep learning based style transfer has shown 2% improvements in classification accuracy over traditional augmentation strategies [26]. 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 [22]. 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][27][15].

3 Methodology

First, a pretrained diffusion model [5] [16][15] is fine-tuned on a generic dataset which does not necessarily include the infrequent target objects (we used a generic driving dataset [25]). In order to condition the diffusion process on text, we use a CLIP model [20] 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

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 [16] 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.

3.1.1 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.

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

3.1.2 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, possible verbs can be "driving", "crossing", "parking".

Figure 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.

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.

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

3.2 Approach 2: Real Infrequent Objects in Synthetic Background

This approach can be also represented by the top part of Figure 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 resulting 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 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.

Figure 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 it's surroundings. Image 2 (row 1, image 2): ambulance is perpendicular to the street and not parked, this would not happen 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: observe a flying car.

4 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.

Table 1: Datasets Used for Experimentation

Dataset	Num. images	Medical	Fire	Police	Normal
Real-Train (R)	5215	47	42	126	5000
Synthetic Type-1 (S1)	1876	487	576	1028	0
Synthetic Type-2 (S2)	1875	642	366	1081	0
Augmented (AUG)	1876	383	372	1121	0
Real-Test (R-Test)	1539	268	68	203	1000

4.1 Real Data (Real-Train, Real-Test)

The LISA-Amazon Vehicle and Scene Attributes (LAVA) dataset [17] 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 containing all the images with emergency vehicles. was separated and used for generating synthetic images and training the downstream object detection models. We refer to this subset as the LAVA-emergency dataset R-Train.

Using the same LAVA dataset, we sampled 5000 frames for the normal vehicle class. These are any non-emergency vehicles that appear frequently on the road. Since emergency vehicles like fire, medical, and police are rare occurrences on the road, we ensure 5-30% of the data in our experiments are emergency vehicles.

4.2 Augmentated Data

To benchmark our synthetic data generation approaches, we perform classic augmentation techniques using the Albumentations library [1] on the real emergency vehicles in Real-Train. Comparing synthetic data with augmented real data provides a baseline for the performance of models trained on synthetic data. Table 2 show the transformations used in augmented dataset (AUG)

Table 2: Augmentation Types Used for Augmentations (AUG) Dataset

Augmentation Type	**Num. images**
Horizontal Flip	157
Random Brightness Contrast	157
Random Shadow	157
MudSpatter	157
ISONoise	157
ToSepia	157
HorizontalFlipSunFlare	157
PixelDropout	157
RainSpatter	157
RandomToneCurve	157
Equalization	157
Blur	149
total	1876

5 Experiments

5.1 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 2. 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 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 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 additional diffusion steps through the super-resolution model.

5.2 Results

Table 3 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 [23], the normal vehicle performance also improves.

Table 3: Downstream Object Detection Performance for Each Dataset.

Model	Dataset	Num. Train Images	mAP@0.50:0.95 Emergency Vehicles	mAP@0.50:0.95 Normal Vehicles
SSD ResNet101 V1 FPN	R	5215	0.075	0.487
SSD ResNet101 V1 FPN	R+AUG	7091	0.205	0.589
SSD ResNet101 V1 FPN	R+S1	7091	0.177	0.604
SSD ResNet101 V1 FPN	R+S2	7090	0.297	0.609
SSD ResNet101 V1 FPN	R+S1+S2	8966	**0.368**	0.639
EfficientDet D1	R	5215	0.056	0.306
EfficientDet D1	R+AUG	7091	0.142	0.480
EfficientDet D1	R+S1	7091	0.180	0.503
EfficientDet D1	R+S2	7090	0.203	0.518
EfficientDet D1	R+S1+S2	8966	**0.287**	0.508
Faster RCNN Inception ResNet V2	R	5215	0.177	0.598
Faster RCNN Inception ResNet V2	R+AUG	7091	0.256	0.613
Faster RCNN Inception ResNet V2	R+S1	7091	0.189	0.578
Faster RCNN Inception ResNet V2	R+S2	7090	0.419	0.577
Faster RCNN Inception ResNet V2	R+S1+S2	3966	**0.434**	0.611
YOLOX V2	R	5215	0.112	0.337
YOLOX V2	R+AUG	7091	0.157	0.340
YOLOX V2	R+S1	7091	0.229	0.439
YOLOX V2	R+S2	7090	0.333	0.449
YOLOX V2	R+S1+S2	8966	**0.418**	0.428
Deformable DETR	R	5215	0.427	0.444
Deformable DETR	R+AUG	7091	0.440	0.574
Deformable DETR	R+S1	7091	0.515	0.557
Deformable DETR	R+S2	7090	0.642	0.553
Deformable DETR	R+S1+S2	8966	**0.662**	0.681

All combinations of synthetic data including S1, S2 and S1+S2 outperform the conventionally augmented data for EfficientDet D1, YOLOX V2 [21], and Deformable DETR [28] 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.

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 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 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.

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.

6 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

[1] E. Khvedchenya V. I. Iglovikov A. Buslaev, A. Parinov and A. A. Kalinin. Albumentations: fast and flexible image augmentations. *ArXiv e-prints*, 2018.

[2] D Block, I Teliban, F Greiner, and A Piel. Prospects and limitations of conditional averaging. *Physica Scripta*, 2006(T122):25, 2006.

[3] Christopher Bowles, Liang Chen, Ricardo Guerrero, Paul Bentley, Roger Gunn, Alexander Hammers, David Alexander Dickie, Maria Valdés Hernández, Joanna Wardlaw, and Daniel Rueckert. Gan augmentation: Augmenting training data using generative adversarial networks. *arXiv preprint arXiv:1810.10863*, 2018.

[4] Moustapha Cisse, Yossi Adi, Natalia Neverova, and Joseph Keshet. Houdini: Fooling deep structured prediction models. *arXiv preprint arXiv:1707.05373*, 2017.

[5] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34, 2021.

[6] Maayan Frid-Adar, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Synthetic data augmentation using gan for improved liver lesion classification. In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 289–293. IEEE, 2018.

[7] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. A neural algorithm of artistic style. *CoRR*, abs/1508.06576, 2015. URL `http://arxiv.org/abs/1508.06576`.

[8] Charles Hamesse, Rihab Lahouli, Timothée Fréville, Benoît Pairet, and Rob Haelterman. Training machine learning algorithms for computer vision tasks in difficult conditions: 3d engines to the rescue. 2019.

[9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[11] C Joshi. Generative adversarial networks (gans) for synthetic dataset generation with binary classes, 2019.

[12] Gwanghyun Kim and Jong Chul Ye. Diffusionclip: Text-guided image manipulation using diffusion models. 2021.

[13] Swee Kiat Lim, Yi Loo, Ngoc-Trung Tran, Ngai-Man Cheung, Gemma Roig, and Yuval Elovici. DOPING: generative data augmentation for unsupervised anomaly detection with GAN. *CoRR*, abs/1808.07632, 2018. URL `http://arxiv.org/abs/1808.07632`.

[14] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European Conference on Computer Vision (ECCV)*, 2016.

[15] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 34, 2021.

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

[17] Kulkarni Ninad, Rangesh Akshay, Buck Jonathan, Feltracco Jeremy, Trivedi Mohan, Deo Nachiket, Ross Greer, Sarraf Saman, and Sathyanarayana Suchitra. Create a large-scale video driving dataset with detailed attributes using amazon sagemaker ground truth. 2021.

[18] Augustus Odena, Chris Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *International Conference on Learning Representations*, 2016.

[19] Thomas Pollok, Lorenz Junglas, Boitumelo Ruf, and Arne Schumann. Unrealgt: using unreal engine to generate ground truth datasets. In *International Symposium on Visual Computing*, pages 670–682. Springer, 2019.

[20] 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.

[21] Joseph Redmon and Ali Farhadi. Yolo9000: Better, faster, stronger. *arXiv preprint arXiv:1612.08242*, 2016.

[22] 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.

[23] Mingxing Tan and Quoc V Le. Efficientdet: Scalable and efficient object detection. *arXiv preprint arXiv:1911.09070*, 2019.

[24] Belén Vega-Márquez, Cristina Rubio-Escudero, José C Riquelme, and Isabel Nepomuceno-Chamorro. Creation of synthetic data with conditional generative adversarial networks. In *International Workshop on Soft Computing Models in Industrial and Environmental Applications*, pages 231–240. Springer, 2019.

[25] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020.

[26] Xu Zheng, Tejo Chalasani, Koustav Ghosal, Sebastian Lutz, and Aljosa Smolic. Stada: Style transfer as data augmentation. *CoRR*, abs/1909.01056, 2019. URL `http://arxiv.org/abs/1909.01056`.

[27] Sharon Zhou, Mitchell L. Gordon, Ranjay Krishna, Austin Narcomey, Durim Morina, and Michael S. Bernstein. HYPE: human eye perceptual evaluation of generative models. *CoRR*, abs/1904.01121, 2019. URL `http://arxiv.org/abs/1904.01121`.

[28] Xingyi Zhu, Yanwei Tu, Jian Sun, Shuo Gong, Kai Chen, Kaiming He, and Ross Girshick. Deformable detr. In *European Conference on Computer Vision (ECCV)*, 2020.