DALL-E for Detection: Language-driven Context Image Synthesis for Object Detection

Yunhao Ge\textsuperscript{1}, Jiashu Xu\textsuperscript{1}, Brian Nlong Zhao\textsuperscript{1}, Laurent Itti\textsuperscript{1}, and Vibhav Vineet\textsuperscript{2}

\textsuperscript{1} University of Southern California \{yunhaoge, jiashuxu, briannlz, itti\}@usc.edu
\textsuperscript{2} Microsoft Research Vibhav.Vineet@microsoft.com

Abstract. Object cut-and-paste has become a promising approach to efficiently generate large sets of labeled training data. It involves compositing foreground object masks onto background images. The background images, when congruent with the objects, provide helpful context information for training object recognition models. While the approach can easily generate large labeled data, finding congruent context images for downstream tasks has remained an elusive problem. In this work, we propose a new paradigm for automatic context image generation at scale. At the core of our approach lies utilizing an interplay between language description of context and language-driven image generation. Language description of a context is provided by applying an image captioning method on a small set of images representing the context. These language descriptions are then used to generate diverse sets of context images using the language-based DALL-E image generation framework. These are then composited with objects to provide an augmented training set for a classifier. We demonstrate the advantages of our approach over the prior context image generation approaches on four object detection datasets. Furthermore, we also highlight the compositional nature of our data generation approach on out-of-distribution and zero-shot data generation scenarios.

Keywords: DALL-E, Synthetic data, Language-driven, Context image, Object detection

1 Introduction

Training modern deep learning models require large labeled datasets \cite{43,26,48}. Obtaining such datasets is both expensive and time-consuming due to human effort. An alternative approach for data collection has been to synthesize images that can easily generate large accurate labeled data for training computer vision models. One promising direction for synthetic data generation has been based on object cut and paste \cite{17}. It involves compositing foreground objects onto background images. While the approach has been applied to generate data
Table 1: Desired quality of context generation method: Generation of a high-quality and diverse set of images (scalable), no(less) human involvement in context image generation, automatic generalization of the images for any new environment, scalable, explainable, privacy-preserving, and compositional.

| Method               | Quality | No Human | Adapt | Scalable | Explainable | Privacy | Compositional |
|----------------------|---------|----------|-------|----------|-------------|---------|---------------|
| Human capture        | ✓       | ✗        | ✗     | ✓        | ✗           | ✓       | ✗             |
| Web image            | ✓       | ✓        | ✗     | ✓        | ✗           | ✓       | ✗             |
| Public dataset       | ✓       | ❌        | ✗     | ✓        | ✗           | ✗       | ❌             |
| Generative models    | ✓       | ✓        | ✗     | ✓        | ✗           | ✓       | ✗             |
| Ours                 | ✓       | ✓        | ✓     | ✓        | ✓           | ✓       | ✓             |

for many computer vision tasks [21, 59, 38, 60], the question of how to select the background images that provide necessary context information is still open.

Context plays a very important role in learning a good object recognition model. Divalla et al. [14] provide empirical evidence to support this claim. Dvornik [16] showed that finding congruent context helped improve accuracy on object detection tasks. For example, placing airplanes and boats in their natural context helped to improve accuracy, e.g., airplanes are generally found in the sky and boats are on the water. Furthermore, in the object cut-and-paste work, Yun et al. [60] have also studied the effect of background images on image segmentation tasks. They observed that selecting images from the UW dataset [31] provided better accuracy than the COCO dataset [35]. This suggests that, for the above downstream task, images from UW data provide better context information.

This raises a question: how can context images be generated that will maximize accuracy on the downstream task? We first hypothesize that any approach that generates context images should satisfy these qualities (Fig. 1): no human involvement in context image generation, automatic generalization of the images for any new environment, scalable, generation of high quality and diverse set of images, explainable, compositional, and privacy-preserving.

Several approaches can be used to generate context images. The first approach could involve selecting web images. Large sets of background images can be easily gathered, but the presence of noisy and out-of-context background images can affect accuracy [60]. Next, background context images can be selected by humans by capturing images in the test environment. Human involvement in training data generation can severely affect scalability. Finally, background images could also come from a similar-looking dataset. For example, if the test environment involves kitchen scenes, context images can be taken from different kitchen datasets. This approach suffers from scalability and diversity issues. For example, it is not easy to find a similar-looking dataset without also having the object of interest present in those images. Furthermore, all the above approaches may not be privacy-preserving. While these approaches involve selecting real-world images as context images, generative models like GANs [22, 6]
Fig. 1: Motivation flow of our pipeline. How the desired properties described above are satisfied by our language-driven compositional context synthesis.

and VAEs [12,27] can also be used to generate context images. These images could consist of noisy and out-of-context images that may severely affect overall accuracy.

In this work, we propose a new paradigm for automatic and scalable context image generation that satisfies all the above-desired properties. At the core of our approach lies utilizing an interplay between language description of context and language-driven image generation. Given a small number of images that represent the context environment, we use image captioning to generate a high-level language description of the context automatically. The language description of the context is used within a text-to-image generation pipeline to generate a diverse set of images. Methods like DALL-E [42], RU-DALLE [1], CogView [13] have been trained on large vision-language data to generate high-quality images from textual descriptions. These diverse sets of generated images are used as context images. Foreground objects are pasted on these context images to provide high-quality accurate ground truth labels for training downstream computer vision tasks.

The proposed pipeline satisfies all the desired properties of context image generation (Fig. 1). The image captioning component helps to easily and succinctly generate the context of any new changing environment. Language helps to provide an explainable and compositional foundation to context description and data generation. Adding or removing objects or settings can be easily done in the language domain. For example, a description as an environment with table and chair can be easily modified to kitchen environment by utilizing the compositional properties of language as a kitchen environment with table and chair. Table. 1 shows the benefit of our approach in generating context images over other approaches.

| High quality with diverse set | Scalable | Privacy & Safety | Explainability | Flexible & Compositional | No / less human involve | Automatic adaptation & generalizable to new tasks |
|------------------------------|----------|------------------|----------------|-------------------------|------------------------|-----------------------------------------------|
| **Generative models**        |          |                  |                |                         |                        |                                               |
| DALL-E                       |          | Image understanding |               |                         |                        |                                               |
| **Language**                 |          |                  |                |                         |                        |                                               |
| **Image caption**             |          |                  |                |                         |                        |                                               |

We have conducted extensive experiments on four publicly available benchmark object detection datasets and compared them against different ways to generate context images. We demonstrate that our approach can achieve much better accuracy compared to the prior approaches. Furthermore, we also demon-
strate the benefit of our approaches in out-of-distribution context and zero-shot data generation scenarios that utilize the compositional nature of our method. We summarize the main contributions of our approach as follows:

– Propose a language-driven context image generation approach to automatically generate large-scale datasets.
– Demonstrate benefit over prior approaches in context image generation.
– Highlight compositional nature by generating context images from out-of-context images and zero-shot data generation scenarios.

2 Related work

Object cut and paste A series of works on using synthetic data for training computer vision problems have been proposed. Some of them include using graphics pipeline or computer games to generate high quality labelled data\cite{43,44,45,46,52,29}. Generally using graphics pipeline requires having 3D models of both objects and environment, that may limit the scalability of these approaches.

The idea of pasting foreground objects on background images has emerged as a easy and scalable approach for large scale data generation. Dosovitskiy et.al.\cite{15} created large data for optical flow problem by creating chair object and pasting them onto background images. Dwibedi et.al.\cite{17} showed this idea can be used to generate high quality labelled data for object instance segmentation task. The idea has been used to solve other problems like object detection and pose estimation problems\cite{50,51,53,28,52,40}. Yun et.al.\cite{60} showed that cut-and-paste strategy can be used for solving domain adaptation problem. In addition, Ghiasi et.al.\cite{21} also showed that copy-paste augmentation can provide benefit to semi-supervised learning approaches. While these works showed that object cut-and-paste can be used to generate large labelled synthetic data easily, they needed careful congruent background images selection for compositing foreground objects onto the background images. This effect has been also observed by\cite{16}. In our work, we have proposed an automatic approach to select congruent background images that can provide useful context information for object recognition task.

Context Several prior works have studied importance of context in recognition problem. Divvala et.al.\cite{14} studied role of context on Pascal Object detection problem\cite{18} and observed that the context helped to reduce overall object detection error. Mottaghi et.al.\cite{39} studied role of context in object detection and semantic segmentation tasks and developed a novel deformable parts based model to combine local and global context information. Similarly, Dvornik et.al.\cite{16} and Zhang et.al.\cite{61} showed that modeling context correctly improved object detection performance and semantic segmentation settings. Similarly Lee et.al.\cite{32} proposed a context aware foreground object placement in background images that helped to improve accuracy. In the domain of object cut-and-paste
problem, Yun et.al. [60] showed that selecting right background images for cut-and-paste helped improve accuracy. We are motivated by these observations to learn congruent background images for solving downstream object recognition tasks.

Language for object recognition. Language has been used to solve computer vision tasks. Vision-language based models have been developed for image captioning task [55,44,24], visual question answering tasks [5,34,9] and others [87,38,34]. In recent years, vision and language based multi-modal models have been developed for self-supervised training. Recent work CLIP [31] showed how training a model on large image-text pair dataset can generalize to several benchmark image classification datasets where current image based models performed very poorly. Language information been used along with vision information to solve other computer vision tasks like object detection [23,30] and semantic image segmentation task [33]. These works have also demonstrated benefits of leveraging language information in solving computer vision tasks.

Another line of work involves leveraging large image-text pairs for text to image generation tasks. These models learnt to generate realistic environment including out-of-context settings which may not be present during the training phase. Some of the examples include DALL-E [42], RU-DALLE [1], CogView [13]. We are motivated by generation quality of these text to image generation methods in this work.

3 Method

The goal of the paper is to efficiently generate a large set of labeled data for the training object detection model. The problem is framed as compositing foreground object masks onto background images following object cut-and-paste strategy [17]. Congruent background images provide essential context information for training object recognition. While the approach can easily generate large labeled data, the problem of finding congruent background images that provide necessary context information has remained an elusive problem. Finding the right context is the focus of our work.

We start by describing the object cut-and-paste cut approach, then we provide an overview of the pipeline before going into details of each individual step.

3.1 Object cut and paste.

We start by describing the details of the object cut-and-paste work of Dwibedi et.al. [17], which proposed an easy method to synthesize large labeled data. It consists of three steps.

Foreground mask segmentation. The first step involves generating a set of object masks. These object masks can be segmented from images consisting of those objects. The images come from the BigBird dataset [49], each paired with a depth image. However, it is shown [17] that simply thresholding depth images
can not produce a high quality ground truth object masks, especially for the objects that have transparent surfaces such as Coca Cola. Therefore, in order to generate better foreground masks, we train a 2-class FCN [36] to classify each pixel as foreground and background objects. The ground truth mask to train FCN is provided by non-transparent objects in BigBird dataset.

Background image selection. The second step involves the selection of background images that provides necessary contextual information for the object recognition tasks. A general approach is to simply take random web images or use images from any popular dataset to create background images. In particular, Dwibedi et.al. [17] used images from the UW dataset [31] to provide background images for evaluation on the GM Kitchen dataset [20]. Similarly, ImageNet [10], COCO [35] and Pascal [18] can be a good sources of background images.

Mask compositing. The final step involves compositing foreground object masks onto the background images. At each step a group of foreground object masks are selected and pasted into a sampled background images, and such procedure is repeated until all foreground object masks are pasted. Composition involves applying random 2d data augmentation, such as rotation and scaling, and then pasting on random background images at random locations. As a final step, different blendings are applied to alleviate boundary artifacts caused by pasting.

3.2 Language-driven Context generation

In this paper, we propose a new paradigm to generate context images automatically at a large scale. The proposed approach consists of three steps (a visualization of our pipeline is provided in Fig.2). The first step involves generating a diverse set of natural language descriptions of context by applying image captioning methods on a few images taken from the context environment. Text to image generation pipeline synthesizes a large set of context images from the natural language descriptions. These context images are then composited with foreground object masks for large-scale training data generation process. Furthermore, the proposed approach also allows for compositional and explainable data generation as well. We describe each of these steps in detail below.

Context description images (CDI). We first assume that we have been given a small set of images that describes the context environment, we call them context description images (CDI). This small set of images can come from an environment that contextually looks similar to the test environment. For example, if the test scenario includes a kitchen environment, the small set of initial kitchen images can be taken from any public dataset or from web images. It should be noted that these context images can be as small as one image.

Image caption. Next step involves succinctly describing context information from the given CDIs. Language can be used to provide a concise description of the context information. In addition, language is both compositional and explainable in nature. Given these advantages, we use language descriptions to represent context information. This raises a question: how can these language descriptions be generated?
Fig. 2: Overview of our pipeline. The user will need to provide a set of CDIs, and note that the user can provide as little as one image. (1), we leverage SCST [44] image captioning models to generate captions for the user provided CDIs. (2), we feed the captions to DALL-E to create images that align with input captions. (3), we combine BigBird foreground objects to obtain synthetic images using cut and paste. (4), we feed the synthetic images to object detection models to train on.

In order to automatically generate such a description of the context, we use image captioning methods. These captioning approaches generate a set of diverse textual captions for input images. Over the years, many image captioning methods have been developed [44, 24]. We use self-critique sequence training (SCST) for image captioning work developed by Rennie et.al. [44]. The method utilizes test time inference to provide rewards during the image to caption training stage. REINFORCE [56] is used to optimize the test time reward. Such an optimization framework allows SCST to achieve very good accuracy on different caption generation datasets including COCO caption generation challenge. Further, it should be noted that our method does not rely on any specific image to caption generation method. Any popular or new SOTA method can be used in place of the SCST method.

For each CDI, \( K \) natural language descriptions are generated using the SCST method. If there are \( N \) CDIs, the caption generation steps provide \( N \times K \) natural language descriptions.
**Image generation.** Next step involves generating a diverse set of $M$ images for each language description of the context information. Recent time has seen a remarkable breakthrough in the text to image generation approach. Generally, these approaches involve training large models using a hugely diverse set of text-image data pairs in a self-supervised fashion. Following the success of transformer models in the language domain the [7,11], text to image generation model also utilizes transformer architectures. The training procedure involves concatenating image and text features into a single stream and training the transformer model in an autoregressive manner. Some popular frameworks for text to image data generation are DALL-E [42], CogView [13], RU-DALL-E [1]. The DALL-E model has been trained on 200 million image-text pairs.

The text-to-image generation model has several benefits. First, it is a compact version of web-scale image-text pair data. That makes it both portable and scalable. Also, being a generative model, the generation pipeline could create new scenarios that were not present in the training data. Furthermore, the synthetic nature of the data generation procedure allows our method to be privacy-preserving.

For each text description, we use DALL-E to generate $M$ images. This approach helps us to generate a total of $N \times K \times M$ images from $N$ CDIs. Generally, in our experiments, $N$ is of order 10, $K$ is of order 10 and $M$ is of order 100. So, even for single context images, we are able to automatically generate a large set of new context images. We will show examples of some input CDIs and output context images from our approach in section 4.1 and 4.5.

**Labeled generation.** Given a set of context images and foreground mask images, we generate labelled ground truth data by pasting foreground mask onto the context images. In order to synthesize the composite image, we randomly select context images and paste the foreground mask on random location in the image. In addition, we follow [17] for compositing foreground on the background image. Note that this method involves applying a Gaussian blur on the object boundary with blur kernel as $\sigma$. This allows us to synthesize a large set of training images with accurate ground truth bounding box labels.

**Compositional data generation.** The natural language description allows compositional data generation. We can intervene in the language description to add or remove a key word (Fig. 3). For example, the word *kitchen* can be added to generate context. This allows us to generate new images with new context information. Similarly, we can remove some unwanted objects. For example, if initial context description involves people, we can remove people from generated images by simply not mentioning the word *people*. The text to image generation pipeline can then generate large set of images without people.

4 Experiments

In this section, we demonstrate that our method can generate good context images for downstream object detection tasks on three popular benchmark datasets. We consider GMU-Kitchen dataset [20] (Sec. 4.1), Active Vision dataset [3] (Sec. 4.2), YCB-video dataset [57] (Sec. 4.3) and PASCAL VOC [19] (Sec. 4.4).
Fig. 3: We highlight compositional and explainable properties of our method. Specifically, when the provided CDI cannot perfectly describe the real test scenario, the compositional property of language can help to correct context description by remove/add/style change. For instance, if the initial description contains noisy information "man and a woman", we can directly intervene and remove the noise information to generate congruent context description. Note that, all the 4 example test scenarios are from GMU kitchen dataset.

More details about these datasets are provided in the coming sections. Furthermore, we also provide results highlighting compositional nature of our data generation process (Sec. 4.5).

Training procedure and evaluation criterion. We use faster RCNN [43] with ResNet-50 [26] as the backbone to train object detection network. Models have been trained till convergence for both the baselines and our approaches. In our experiments we set learning rate as 0.001 with a weight decay 0.0005. Furthermore, we use standard mean average precision (mAP) for the evaluation of object detection results.

4.1 GMU-Kitchen dataset

In order to conduct experiments on the GMU-Kitchen dataset, we follow the evaluation protocol described in the Object Cut-and-Paste [17] paper. **Dataset.** GMU-kitchen dataset [20] is created by placing object instances in 9 kitchen scenes. 11 objects that overlap with 33 objects from the Big Bird dataset [49] are selected for evaluation. The dataset consists of 6,728 images. When we use GMU kitchen as test set, similar as Cut-and-Paste paper, 3-fold cross-validation was used.

**Foreground mask.** In order to create foreground masks for the GMU-kitchen objects, we first take images from the Big Bird dataset that consists of six hundred images taken from six RGB-D calibrated cameras for each of the 33 objects. We follow the strategy described in Sec. 3.1 to get the foreground masks. **Baselines.** We evaluate our approach with several other standard approaches for selecting context images. The main baseline consists of a comparison with...
Fig. 4: Training images generated by our pipeline: pasting foreground objects on DALL-E synthesized images with CDI form UW dataset.

Results. Quantitative results are shown in the Table 2. We observe that our approach is able to get almost 2.1 percentage points improvement over the Object Cut-and-Paste baseline. This highlights that our approach generates a better diverse set of context images compared to the images from the UW dataset. Furthermore, we also observe 37% points, 63.2%, and 16% points of improvements over black, CDIs, and COCO images respectively. Next, we conduct experiments to demonstrate that caption descriptions are important to generate congruent context images. To this end, we use random language descriptions to generate images from DALL-E. We observe almost 12% points improvement of our approach over the random images from DALL-E. Finally, the use of real-world GMU training data combined with our synthetic data helps to achieve the best performance on the GMU test set, i.e., almost 5% points improvement over training with real GMU training data only. Compared to all the baselines, our approach is able to achieve better performance. This highlights the benefit of using our language-driven context image generation approach to synthesize congruent context images. Some qualitative images are shown in Fig. 4 and Fig. 5 that provides a glimpse into the kinds of context images generated by our approach.
Table 2: Quantitative results on GMU-Kitchen dataset. We compare our approach with several prior approaches. Our approach achieve highest accuracy over the baselines. Combining GMU kitchen real training samples with our synthetic data yields the best results on this dataset. Real: real GMU training data as training set. Top row terms are: \#CDI: the number of context description images, CC: Coca Cola, CM: Coffee mate, HB: honey bunches, HS: hunt’s sauce, MR: mahatma rice, NV1: nature V1, NV2: nature V2, PO: palmolive orange, PS: pop secret, Pbbq: pringles bbg, RB: red bull.

| Dataset               | \#CDI | CC   | CM   | HB   | HS   | MR   | NV1 | NV2 | PO | PS | Pbbq | RB | mAP |
|-----------------------|-------|------|------|------|------|------|-----|-----|----|----|------|----|-----|
| Real GMU train        | -     | 81.9 | 95.3 | 92.0 | 87.3 | 90.5 | 88.9| 80.5| 92.3| 88.9| 58.6 | 86.3|     |
| Black                 | 1500  | 42.3 | 62.4 | 64.7 | 5.3  | 3.3  | 61.1| 56.5| 75.3| 1.6 | 26.7 | 33.9| 41.2 |
| CDI                   | 10    | 51.4 | 26.4 | 2.1  | 12.2 | 12.1 | 0.4 | 0.1 | 1.0 | 0.1 | 29.8 | 30.9| 15.0 |
| Random (COCO)         | 1500  | 50.7 | 80.1 | 77.5 | 15.3 | 32.2 | 81.7| 87.9| 71.7| 66.8| 59.0 | 68.5| 62.8 |
| Random (DALL-E)       | 1500  | 64.8 | 86.9 | 78.7 | 49.2 | 62.2 | 84.8| 83.8| 72.6| 70.9 | 57.2 | 24.1| 66.8 |
| UW-Kitchen            | 1500  | 75.7 | 91.1 | 87.7 | 51.6 | 66.5 | 91.5| 88.7| 76.2| 63.2 | 70.5 | 75.2| 76.1 |
| DALL-E (ours)         | 1500  | 79.0 | 92.9 | 90.4 | 44.9 | 77.0 | 92.1| 88.0| 77.5| 64.1 | 75.7 | 80.2| 78.3 |
| DALL-E (ours)         | 2400  | 79.5 | 93.4 | 88.5 | 59.0 | 71.5 | 91.4| 88.1| 76.1| 78.7 | 75.7 | 80.6| 80.1 |
| DALL-E (ours)+Real    | 1500  | 94.4 | 98.2 | 95.2 | 90.7 | 92.5 | 94.1| 93.0| 72.8| 98.3 | 98.7 | 79.8| 91.4 |

Table 3: Per class detection accuracy on Active Vision dataset. CC: Coca Cola, HB: honey bunches, HS: hunt’s sauce, MR: mahatma rice, NV2: nature V2, PO: palmolive orange, PS: pop secret, Pbbq: pringles bbg, RB: red bull.

| Dataset               | CC   | HB   | HS   | MR   | NV2 | RB   | mAP |
|-----------------------|------|------|------|------|-----|------|-----|
| UW-Kitchen            | 42.8 | 20.9 | 15.4 | 1.7  | 19.8| 34.8 | 22.6|
| DALL-E (ours)         | 45.0 | 21.9 | 19.3 | 4.8  | 22.4| 37.5 | 25.8|

4.2 Active Vision dataset

In order to conduct experiments on the Active-vision dataset [3], we follow the evaluation protocol described in the Object Cut-and-Paste paper [17].

**Dataset.** We consider six objects to evaluate performance on the Active-vision dataset. These objects overlap with the GMU-Kitchen dataset objects. There are 4,000 images that consists of 2,000 test images. These images are also captured in the kitchen environment. The Foreground masks are same as GMU kitchen dataset in Sec. 4.1.

**Results.** Quantitative results are shown in the Table 3. We observe that our approach is able to get 3.2 percentage points improvement over the object cut-and-paste baseline. This highlights that our approach generates better context images compared to the images from UW dataset.
4.3 YCB-video dataset

**Dataset.** YCB-video objects [57] is a popular object detection and pose estimation dataset in computer vision and robotics community. It consists of 21 object classes that are generally taken from our daily environments. The data consists of video sequences of YCB-objects present in different indoor environments. The test images are selected from these sequences.

**Foreground mask.** These objects are subset of YCB-object dataset [8]. The YCB-object dataset provides 600 images taken from six camera where objects are placed on white background. Masks are generated by the same strategies discussed in the Sec. 3.1.

**Results.** Table 4 shows quantitative numbers of both the baseline and our approaches. Note that our approach can get 7.2% points improvement over the Object Cut-and-Paste [17] approach with UW-kitchen real context images. This highlights the benefit that our language-driven DALL-E generated images provide congruent and diverse context images compared to using real world images from other public datasets.

4.4 PASCAL-VOC dataset

We also evaluate our method on PASCAL VOC 2012 object detection task [19]. In Table 5 we report accuracy using three commonly used PASCAL VOC object detection metrics: mAP@50, mAP@75 and mAP.
**Dataset.** The dataset has 20 foreground classes. The training and validation set consist of 1,464 images and 1,449 images, respectively with bounding box along with instance segmentation masks. We use the instance segmentation masks from the training set ground truth labels as our foreground masks for cut-and-paste operations.

**Experiments set up.** We conduct two groups of experiments with total of 8 experiments. All experiments use ResNet-101 as backbone. We fix the backbone and fine-tune the head modules only. The first group consists of 4 experiments (EXP). All of them use same foreground object masks that are present in the original 1,464 training images. Further these 1,464 training images are considered context description images (CDIs). EXP-1 is the baseline which involves training the object detection model on the original 1,464 training images without any augmentation of the images using cut-and-paste steps. EXP-2 applies cut-and-paste method into the baseline experiment. Specifically, we randomly paste the foreground object masks onto the training images to form 4,202 different images. For EXP-3, we first use information provided by the context description images to generate context images from DALL-E. We then use cut-and-paste to compose the same set of foreground object masks as in the EXP-2 on the synthesized DALL-E images. This results in a total of 8,438 images. EXP-4 combine the original real training images with DALL-E synthesized images as context image. The second group consists of 4 experiments which are similar as the 4 experiments in the first group. The only difference is the number of CDIs. The second group use only 200 training images as CDIs (10 images per class sampled from original 1,464 training set). The used foreground object masks and all other setting is the same as the first group.

**Results** Table. 5 shows the results of the two groups of 8 experiments. In first group of 4 experiments, we observe at most 6% points improvement on mAP@50 with the help of our method, which shows the effectiveness of our method on large scale and large variance context dataset. In the second group of 4 experiments with only 200 CDIs, we observe that the baseline is hard to converge which might be due to the limited number of training images. However, with our method, we achieve a significant performance improvement.

4.5 Compositional Model

Here we demonstrate the compositional nature of our approach and highlight how language, as a self-interpretable modality with compositionality property, can provide several benefits for synthetic data generation.

**Out of distribution CDIs.** We first consider scenarios where the context description images are out-of-distribution images. For example, suppose the task is to do evaluation in the real kitchen environment, but the context description images are sketch or cartoon images of the kitchen. Even in these scenarios our approach can generate very good context images. This is achieved because the image caption method still works on these out-of-distribution images. Some of these out-of-distribution images, their corresponding captions and context images generated by our approach are provided in the Fig. 3 and Fig. 6.
Table 5: Object detection results on PASCAL VOC dataset. mAP is computed as average of IoU ranging from 50 to 95 with step size 5.

| Dataset                        | #CDI | mAP@50 | mAP@75 | mAP |
|--------------------------------|------|--------|--------|-----|
| Real VOC train                | -    | 45.5   | 7.8    | 17.0|
| + cut paste                   | 1,464| 48.3   | 14.6   | 21.0|
| DALL-E (ours)                 | 1,464| 38.2   | 8.0    | 14.4|
| DALL-E (ours) + Real          | 1,464| 51.5   | **19.2**| **24.1**|
| Real VOC train subset         | -    | 6.4    | 0.6    | 9.4 |
| + cut paste                   | 200  | 38.8   | **8.6**| 19.6|
| DALL-E (ours)                 | 200  | 34.6   | 6.4    | 15.1|
| DALL-E (ours) + Real          | 200  | **39.6**| 6.7    | **20.8**|

Language intervention. Next we show the benefit of compositional nature of our language based context image generation. Suppose the scenario where the user can not provide perfect CDI. They may contain noisy information (distractors or out-of-distribution elements). For instance, the test set scenario are real world kitchen while the provided noisy CID are kitchen images with humans as distractors or only kitchen accessories objects images (microwave, sink). So, presence of distractor objects in the context images may hamper the overall accuracy. Our language-based text to image generation pipeline can handle such situation. In the first case, user provide both noisy CDIs and the knowledge of distractor object or noise. For example, the first example in Fig. 3 user provide people as distractor object. To remove the distractor, we automatically detect and remove the distractor word from the caption by word detection before using them within DALL-E framework. Similarly, We could also add without people to the captions to remove the distractors. Because the text to image generation pipeline DALL-E interprets the meaning of without and generates context images without the distractor objects.

Another scenarios is add scenario, where the CDIs miss some context information. For instance, in the second example in Fig. 3 User provides the noisy CDI and the missing context information kitchen. The noisy CDI focuses only on
Fig. 6: Visualization of context images generated by DALL-E for noisy and out-of-distribution CDIs. In the first row, inputs are cartoon kitchen and in the second row the CDIs are sketch kitchen. Both cartoon and sketch provide noise information about environment. Observe how our approach can handle highlighted words (distractor words) and generate high quality context images.

microwave but does not sufficiently describe kitchen. So we add user provided word *A kitchen of* as a prefix to the caption for CDI description. Captions before and after interventions are shown in the Fig. 3 and Fig. 7.

We use the generated context images with the original caption from the noisy DIS and the generated context images with modified caption to form two datasets and demonstrate the advantages of our compositional advantage in the Table 6. As seen, using context images after modification helps to improve performance by almost 6.7% points, 10.2% points, 5.3% points, and 6% points over generated images from non-modified caption on 4 out-of-distribution scenario (Cartoon kitchen, Skeleton kitchen, objects in Kitchen and Kitchen with human).

5 Conclusion

Object cut-and-paste has become a promising approach to efficiently generate large sets of labelled training data. It involves compositing foreground object masks onto background images. The background images, when congruent with the objects, provide helpful context information for training object recognition
models. However, finding right context images for a downstream task has remained an elusive problem. In this work, we propose a new paradigm for automatic context image generation at scale. At the core of our approach lies utilizing an interplay between language description of context and language-driven image generation. We demonstrate advantages of our approach over the prior context image generation approaches on three object detection datasets. Furthermore, we also highlight the compositional nature of our data generation approach on out-of-distribution and zero-shot data generation scenarios.
References

1. sberbank-ai/ru-dalle: Generate images from texts. in russian. https://github.com/sberbank-ai/ru-dalle (Accessed on 03/07/2022)

2. Agrawal, A., Batra, D., Parikh, D., Kembhavi, A.: Don’t just assume: look and answer: Overcoming priors for visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4971–4980 (2018)

3. Ammirato, P., Poirson, P., Park, E., Košecká, J., Berg, A.C.: A dataset for developing and benchmarking active vision. In: 2017 IEEE International Conference on Robotics and Automation (ICRA). pp. 1378–1385. IEEE (2017)

4. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6077–6086 (2018)

5. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 2425–2433 (2015)

6. Brock, A., Donahue, J., Simonyan, K.: Large scale GAN training for high fidelity natural image synthesis. In: International Conference on Learning Representations (2019).

7. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems 33, 1877–1901 (2020)

8. Calli, B., Singh, A., Walsman, A., Srinivasa, S., Abbeel, P., Dollar, A.M.: The ycb object and model set: Towards common benchmarks for manipulation research. In: 2015 international conference on advanced robotics (ICAR). pp. 510–517. IEEE (2015)

9. Das, A., Datta, S., Gkioxari, G., Lee, S., Parikh, D., Batra, D.: Embodied question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–10 (2018)

10. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)

11. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

12. Diederik Kingma, M.W.: Autoencoding variational bayes. In: ICLR (2014)

13. Ding, M., Yang, Z., Hong, W., Zheng, W., Zhou, C., Yin, D., Lin, J., Zou, X., Shao, Z., Yang, H., et al.: Cogview: Mastering text-to-image generation via transformers. Advances in Neural Information Processing Systems 34 (2021)

14. Divvala, S.K., Hoiem, D., Hays, J.H., Efros, A.A., Hebert, M.: An empirical study of context in object detection. In: 2009 IEEE Conference on computer vision and Pattern Recognition. pp. 1271–1278. IEEE (2009)

15. Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., van der Smagt, P., Cremers, D., Brox, T.: Flownet: Learning optical flow with convolutional networks. In: ICCV (December 2015)

16. Dvornik, N., Mairal, J., Schmid, C.: Modeling visual context is key to augmenting object detection datasets. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 364–380 (2018)
17. Dwibedi, D., Misra, I., Hebert, M.: Cut, paste and learn: Surprisingly easy synthesis for instance detection. In: Proceedings of the IEEE international conference on computer vision. pp. 1301–1310 (2017)
18. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International journal of computer vision 111(1), 98–136 (2015)
19. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision 88(2), 303–338 (2010)
20. Georgakis, G., Reza, M.A., Mousavian, A., Le, P.H., Košecká, J.: Multiview rgb-d dataset for object instance detection. In: 2016 Fourth International Conference on 3D Vision (3DV). pp. 426–434. IEEE (2016)
21. Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T.Y., Cubuk, E.D., Le, Q.V., Zoph, B.: Simple copy-paste is a strong data augmentation method for instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2918–2928 (2021)
22. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672–2680 (2014)
23. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Zero-shot detection via vision and language knowledge distillation. arXiv e-prints pp. arXiv–2104 (2021)
24. Gurari, D., Zhao, Y., Zhang, M., Bhattacharya, N.: Captioning images taken by people who are blind. In: European Conference on Computer Vision. pp. 417–434. Springer (2020)
25. Handa, A., Patraucean, V., Badrinarayanan, V., Stent, S., Cipolla, R.: Scenenet: Understanding real world indoor scenes with synthetic data. CoRR abs/1511.07041 (2015), [http://arxiv.org/abs/1511.07041](http://arxiv.org/abs/1511.07041)
26. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
27. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A.: beta-vae: Learning basic visual concepts with a constrained variational framework. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France (2017)
28. Hinterstoisser, S., Lepetit, V., Wohlhart, P., Konolige, K.: On pre-trained image features and synthetic images for deep learning. arXiv preprint arXiv:1710.10710 (2017)
29. Hodan, T., Vineet, V., Gal, R., Shalev, E., Hanzelka, J., Connell, T., Urbina, P., Sinha, S., Guenter, B.: Photorealistic image synthesis for object instance detection. ICIP (2019)
30. Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., Carion, N.: Mdetr-modulated detection for end-to-end multi-modal understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1780–1790 (2021)
31. Lai, K., Bo, L., Ren, X., Fox, D.: A large-scale hierarchical multi-view rgb-d object dataset. In: 2011 IEEE international conference on robotics and automation. pp. 1817–1824. IEEE (2011)
32. Lee, D., Liu, S., Gu, J., Liu, M.Y., Yang, M.H., Kautz, J.: Context-aware synthesis and placement of object instances. Advances in neural information processing systems 31 (2018)
33. Li, B., Weinberger, K.Q., Belongie, S., Koltun, V., Ranftl, R.: Language-driven semantic segmentation. arXiv preprint arXiv:2201.03546 (2022)
34. Li, G., Duan, N., Fang, Y., Gong, M., Jiang, D.: Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 11336–11344 (2020)
35. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
36. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)
37. Lu, J., Goswami, V., Rohrbach, M., Parikh, D., Lee, S.: 12-in-1: Multi-task vision and language representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10437–10446 (2020)
38. Mayer, N., Ilg, E., Fischer, P., Hazirbas, C., Cremers, D., Dosovitskiy, A., Brox, T.: What makes good synthetic training data for learning disparity and optical flow estimation? International Journal of Computer Vision 126(9), 942–960 (2018)
39. Mottaghi, R., Chen, X., Liu, X., Cho, N.G., Lee, S.W., Fidler, S., Urtasun, R., Yuille, A.: The role of context for object detection and semantic segmentation in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 891–898 (2014)
40. Rad, M., Lepetit, V.: Bb8: A scalable, accurate, robust to partial occlusion method for predicting the 3d poses of challenging objects without using depth. In: International Conference on Computer Vision. vol. 1, p. 5 (2017)
41. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
42. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International Conference on Machine Learning. pp. 8821–8831. PMLR (2021)
43. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis & Machine Intelligence (6), 1137–1149 (2017)
44. Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V.: Self-critical sequence training for image captioning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7008–7024 (2017)
45. Richter, S.R., Hayder, Z., Koltun, V.: Playing for benchmarks. In: ICCV (2017)
46. Richter, S.R., Vinet, V., Roth, S., Koltun, V.: Playing for data: Ground truth from computer games. In: European conference on computer vision. pp. 102–118. Springer (2016)
47. Ros, G., Sellart, L., Materzynska, J., Vázquez, D., López, A.M.: The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In: CVPR (2016)
48. Shelhamer, E., Long, J., Darrell, T.: Fully convolutional networks for semantic segmentation. PAMI (2017)
49. Singh, A., Sha, J., Narayan, K.S., Achim, T., Abbeel, P.: Bigbird: A large-scale 3d database of object instances. In: 2014 IEEE international conference on robotics and automation (ICRA). pp. 509–516. IEEE (2014)
50. Su, H., Qi, C.R., Li, Y., Guibas, L.J.: Render for CNN: viewpoint estimation in images using cnns trained with rendered 3d model views. In: 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015. pp. 2686–2694 (2015). https://doi.org/10.1109/ICCV.2015.308

51. Tekin, B., Sinha, S.N., Fua, P.: Real-time seamless single shot 6d object pose prediction. CoRR abs/1711.08848 (2017), http://arxiv.org/abs/1711.08848

52. Tremblay, J., To, T., Birchfield, S.: Falling things: A synthetic dataset for 3d object detection and pose estimation. In: CVPR (2018)

53. Tripathi, S., Chandra, S., Agrawal, A., Tyagi, A., Rehg, J.M., Chari, V.: Learning to generate synthetic data via compositing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 461–470 (2019)

54. Vedantam, R., Desai, K., Lee, S., Rohrbach, M., Batra, D., Parikh, D.: Probabilistic neural symbolic models for interpretable visual question answering. In: International Conference on Machine Learning, pp. 6428–6437. PMLR (2019)

55. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3156–3164 (2015)

56. Williams, R.J.: Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine Learning (1992)

57. Xiang, Y., Schmidt, T., Narayanan, V., Fox, D.: Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. arXiv preprint arXiv:1711.00199 (2017)

58. Yang, X., Zhang, H., Qi, G., Cai, J.: Causal attention for vision-language tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9847–9857 (2021)

59. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6023–6032 (2019)

60. Yoo, W.H., Kim, T., Lee, J., Kim, J., Kim, J.: Cut-and-paste dataset generation for balancing domain gaps in object instance detection. IEEE Access 9, 14319–14329 (2021)

61. Zhang, L., Wen, T., Min, J., Wang, J., Han, D., Shi, J.: Learning object placement by inpainting for compositional data augmentation. In: European Conference on Computer Vision. pp. 566–581. Springer (2020)
Appendix

A Varying real data

We first demonstrate the effect of incorporating different percentages of real-world training images together with our synthesized images for training object detection models. We conduct experiments on the GMU kitchen [20] test set and the real-world training images are from the GMU kitchen training dataset (100% set contains 3837 images). This experimental setup is similar to the one followed in the Object cut-and-paste paper [17]. All the results are provided in the Table 7. We highlight the mAP accuracy of training with all synthetic data plus 10%, 40%, 70% and 100% real data.

Observe how using only a subset of real-world data (70%) with our synthesized images achieves better performance than full (100%) real-world data only. This suggests the advantages of our data generation approach saving the amount of human efforts required in labeling the real-world data significantly. Further, we also observe that accuracy gradually improves from 78.3% to 91.4% as we increase the amount of real-world data.

| Dataset | CC  | CM  | HB  | HS  | MR  | NV1 | NV2 | PO  | FS  | Pbbq | RB  | mAP |
|---------|-----|-----|-----|-----|-----|-----|-----|-----|-----|------|-----|-----|
| DALL-E (ours) | 79.0 | 92.9 | 90.4 | 44.9 | 77.8 | 92.1 | 88.6 | 77.5 | 64.1 | 75.7 | 80.2 | 78.3 |
| 100% Real | 81.9 | 95.3 | 92.0 | 87.3 | 86.5 | 96.8 | 88.9 | 80.5 | 92.3 | 88.9 | 58.6 | 86.3 |
| DALL-E (ours) + 10% Real | 90.5 | 96.9 | 93.2 | 74.0 | 90.7 | 86.5 | 84.7 | 97.7 | 96.4 | 72.1 | 81.6 |
| DALL-E (ours) + 40% Real | 91.8 | 97.4 | 94.5 | 84.9 | 75.1 | 90.7 | 78.6 | 52.1 | 96.9 | 87.6 | 77.9 | 84.3 |
| DALL-E (ours) + 70% Real | 92.7 | 98.2 | 95.2 | 90.9 | 88.0 | 93.1 | 89.7 | 50.3 | 97.6 | 92.2 | 78.3 | 87.9 |
| DALL-E (ours) + 100% Real | 94.4 | 98.2 | 95.2 | 90.7 | 92.5 | 94.1 | 93.0 | 72.8 | 98.3 | 98.7 | 79.8 | 91.4 |

Table 7: We highlight that our synthesized data together with 70% amount of real data achieves better performance than full (100%) set of real data only. This highlights the benefit of our approach in reducing total human efforts. DALL-E (ours) means DALL-E synthesized 1500 diverse images (use UW as CDI). Top row terms are: CC: Coca Cola, CM: Coffee mate, HB: honey bunches, HS: Hunt’s sauce, MR: mahatma rice, NV1: nature V1, NV2: nature V2, PO: palmolive orange, PS: pop secret, Pbbq: pringles bbg, RB: red bull.

B Additional results from our pipeline

In Fig. 5 of the main paper, we presented one example of context images generated from just one given Context Description Images (CDI). To demonstrate that our model is very generic and can generate a large set of diverse context images from given input as little as one image, we include some other examples of generated images from our pipeline as shown in Fig. 8.
C Additional results for section 4.4

In this section, we demonstrate more results for the compositional experiments present in section 4.4 of the main paper.

In table 5 of the main paper, we evaluate the models with synthetic data generated before intervention and after an intervention. Here in table 8, we provide additional results of training models just on the CDIs. We observe that the model can not yield good results due to the insufficient amount of training instances and the large domain gap. However, if we apply our method without intervention, we can get a significant performance boost by providing diverse training instances. Furthermore, by applying intervention, we can narrow the domain gap and yield even more performance gain, reinforcing the effectiveness of our approach.

| Dataset              | only CDI | No Intervention | After Intervention |
|----------------------|----------|-----------------|-------------------|
| Cartoon Kitchen      | 11.2     | 70.0            | 76.7              |
| Skeleton Kitchen     | 10.3     | 64.6            | 74.8              |
| Objects in Kitchen   | 9.4      | 71.8            | 77.0              |
| Kitchens with Human  | 10.2     | 70.9            | 76.9              |

Table 8: Quantitative results on GMU kitchen dataset highlighting the compositional benefits of our method in handling complex scenarios for example if CDIs come from out of distribution domains. For example, for real-world kitchen test environment (GMU kitchen), CDIs are cartoon kitchen images.

Moreover, we include figures that exemplify our generated images in the compositional experiments. Fig. 9, fig. 10, fig. 11, and fig. 12 are generated results before and after intervention for the experiment Cartoon Kitchen, Skeleton Kitchen, Objects in Kitchen, and Kitchens with Human in the Table, respectively.

D Model predictions

We show the predictions of faster RCNN [43] models trained on synthetic data (with # CDI = 1500 from UW dataset [20]) on the GMU-Kitchen dataset in the Fig. [13]. Note that this model follows the same setup as section 4.1 of the main paper. Although the model is trained on synthetic data, the model is able to yield good predictions without being accessible to GMU Kitchen’s train data.
Fig. 8: Context-images generated from our pipeline.
Fig. 9: The results after we transform cartoon images into real images to narrow the domain gap.
Fig. 10: The results after we transform sketch images to real images to narrow the domain gap.
Fig. 11: The results after we place objects in the kitchen.
Fig. 12: The results after we remove humans from the images. The top row is the given CIDs, and the second row is how we intervene in generated captions. The third row is the images generated based on original captions and intervened captions.
Fig. 13: Qualitative detection results. The colorful box is the prediction of the model, together with the predicted class and confidence. We only show prediction with confidence $> 0.9$. The different color of the box indicates the different prediction of the box.