6 Supplementary Materials

This document provides supplementary materials for the paper "SemAug: Semantically Meaningful Image Augmentations for Object Detection Through Language Grounding" submitted to ECCV 2022, including additional comparisons to state-of-the-art methods, experimental ablation study results, implementation details and example augmented images. The materials are delivered in the following sections:

- 6.1 Code
- 6.2 Additional comparison to Context-DA
- 6.3 Additional ablation results
- 6.4 Additional implementation details
- 6.5 Additional visualizations

6.1 Code

To facilitate the reproducibility and easier usage of our method, we release the code implementation of SemAug at

https://developer.huaweicloud.com/develop/aigallery/notebook/detail?id=4d9fc5b8-7fda-4b95-91c2-27deaa2c8490.

Additionally, a ReadMe file on documentation of how to use the code is also included in this link.

6.2 Additional comparison to Context-DA

In the experiments section of the paper, we compared our SemAug method with several other state-of-the-art (SOTA) methods on the COCO dataset. Here, we additionally provide a comparison with a SOTA context-based method Context-DA [11] on the Pascal VOC object detection. We use the same experiment settings as [11] so we can directly compare with their provided results. Baseline is a blitznet [12] model as the base network with vanilla augmentations. The results are given in Table 8. As shown, we see a +1.9% mAP improvement over Context-Aug [11] and a +3.8% mAP improvement over baseline [12].

6.3 Additional ablation results

Ablation of object selection strategy

We include an additional baseline strategy in our ablation studies. This method is an accuracy based mechanism, where we aim to boost the performance for the object category with lowest per-object AP (Average Precision) from the top N most similar object categories. This promotes to push up the lowest AP, and in turn the mAP (mean Average Precision).

Table 9 shows the results of this experiment for variants of our context based method. The `MostSimilar` baseline selects the most similar object category based on the cosine similarity. The Instance method (our default) first narrows down the selection to the top 3 most similar object categories by cosine similarity, then selects the object category with the least amount of instances in the dataset in order to mitigate the effect of unbalanced datasets, while still allowing for semantic knowledge to be injected. The Baseline-mAP method first narrows down the selection to the top 3 most similar object categories by cosine similarity, then selects the object category with the lowest mAP when trained using vanilla augmentation in an attempt to boost low performing categories, while still allowing for semantic knowledge to be injected. Results in Table 9 show that these methods are comparable.

Ablation on averaging similarities
In this experiment, we study the impact of averaging the similarities across all objects present in the image. This was meant to encompass more of the scene as a whole as opposed to matching objects individually. As shown in Table 10, this method did not improve the results but rather degraded the performance. This may be due to the fact that not all objects in a scene are semantically related and therefore averaging the similarities does not aid in finding contextually meaningful objects to be pasted.

Ablation of word embedding size
We compare our method using various dimensions for the GloVe pre-trained word embeddings. This allows us to see how our results change depending on the size of the embedding. Large dimensions are needed to fully capture the essence of words for more complex NLP problems such as captioning images or answering questions, but Table 11 suggests that we can take advantage of the faster computation times of smaller dimensions as embedding the similarity of objects does not necessarily need larger dimensions.

Ablation of the number of categories used for top-N
We compare our method using various numbers of categories for the top-N calculation. Allowing the method to choose between more options than just the single most similar object allows for similar objects with smaller representation in the dataset to be chosen. This aids in generalization as seen by the increase in AP

Method	Aeroplane	Bicycle	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Dining table	Dog	Horse	Motorbike	Person	Potted plant	Sheep	Sofa	Train	TV Monitor	Average
Baseline	63.6	73.3	63.2	57.0	31.5	76.0	71.5	79.9	40.0	71.6	61.4	74.6	80.9	70.4	67.9	36.5	64.9	63.0	**79.3**	64.7	64.6
Context-DA	69.9	73.8	63.9	**62.6**	35.3	**78.3**	73.5	80.6	42.8	73.8	62.7	74.5	**81.1**	**73.2**	68.9	38.1	67.8	64.3	**79.3**	66.1	66.5
SemAug	**70.8**	**75.6**	**68.2**	59.3	**41.2**	78.1	**78.7**	**81.5**	**45.7**	**76.2**	**68.0**	**75.3**	**81.1**	71.8	**71.1**	**45.0**	**69.3**	**65.4**	79.2	**66.6**	**68.4**

Table 8. Comparison of detection accuracy on VOC07-test. The model is trained on all categories at the same time, by using the 1464 images from VOC12train-seg and Blitznet. The first column specifies the augmentation method used in the experiment. The numbers represent average precision per class in %.

Table 9. Effect of object selection strategy. Using Efficientdet-d0 on the Pascal VOC dataset.

Object Selection Method	AP50
`MostSimilar`	76.66
`Baseline-mAP`	76.86
`Instance`	**77.35**

Table 10. Effect of averaging similarities across all objects in the image. Performed on the COCO dataset using Mask-RCNN with a Resnet-101 backbone.

Method	APdet	APseg
Average similarities	41.7	37.7
No averaging	**42.7**	**38.5**

between N=1 and N=3 in Table 12. As we used COCO, which has 80 categories, the AP appears to plateau as the N is increased. However, this number should be chosen carefully, as a smaller dataset such as VOC with only 20 categories might be forced to incorporate dissimilar categories if the N was chosen to be half the dataset.

Ablation of the number of objects pasted in the image

We compare the use of our method to paste one or two objects into an image. This experiment was conducted using MMDET and the Pascal VOC dataset. Inserting an object can occlude other objects in the scene, and adding too many may remove context from the image. As observed in Table 13, inserting more objects starts to hurt the performance.

Ablation of the effect of blending techniques

In this experiment, we compare the use of different blending methods with SemAug. This experiment was conducted using MMDET and the Pascal VOC dataset. Both Gaussian and averaging filters used a [5,5] kernel. Blending objects in the scene can make them appear more realistic from a human perspective, but from the results in Table 14, it does not appear to improve the network's performance.

Word Embedding	APdet, IOU			APdet, Area			APseg, IOU			APseg, Area		
Dimension	0.5:0.95	0.50	0.75	Sma.	Med.	Lar.	0.5:0.95	0.50	0.75	Sma.	Med.	Lar.
100	**42.7**	**64.5**	46.8	**26.2**	47.0	**56.1**	**38.5**	61.1	**41.5**	**21.8**	42.2	53.0
200	42.3	64.1	46.2	25.8	46.4	56.0	38.2	60.7	40.9	21.6	41.9	52.9
300	**42.7**	**64.5**	**46.9**	25.6	**47.3**	**56.1**	**38.5**	**61.3**	41.1	21.7	**42.3**	**53.4**

Table 11. Effect of word embedding dimension. All experiments were done using our SemAug with MMdet and Mask RCNN with a Resnet 101 backbone on COCO. Glove pretrained embeddings on the Wikipedia 2014 + Gigaword 5 dataset were used with different dimensions.

N	APdet	APseg
1	41.6	37.6
2	42.3	38.1
3	**42.7**	**38.5**
4	42.4	38.3
5	42.6	38.4
10	**42.7**	**38.5**

Table 12. Effect of N for top-N on COCO dataset using Mask-RCNN with a Resnet-101 backbone on the dataset.

Number of objects	mAP
1	**80.7**
2	79.5

Table 13. Effect of single or multiple objects pasted into the scene on the Pascal VOC dataset using Faster-RCNN with a Resnet 50 backbone.

6.4 Additional implementation details

As mentioned in the paper, not all semantic labels provided with COCO and VOC were found in the GloVe dataset. As such, the most similar word in the GloVe dataset was found manually and the word embedding for that word was used instead. Table 15 lists these substitutions.

6.5 Future Works

This work lends itself to ideas not in the scope of this paper. For example, an interesting direction would be weakly-supervised detection, where supervision comes only from a pre-built bank of objects. Additionally, while COCO and Pascal VOC were used here, evaluation on highly imbalanced datasets and larger datasets such as OpenImages would also be merited. Lastly, it would be interesting to see how this image augmentation technique would fair on a non-CNN such as a visual transformer, or in the case of visual question answering such as in [17].

6.6 Additional visualizations

Figure 8-16 demonstrate additional examples of our semantic augmentation strategy. They show side-by-side comparisons of original versus semantically augmented examples. Moreover, Figure 17-24 show examples were different instances from a same object category are selected each time. Figure 25-29 also show the case were different categories are augmented into a same host image.

Blending Method	mAP
No blending	**80.7**
Gaussian	79.4
Averaging	79.9

Table 14. Effect of blending techniques for objects pasted into the scene on the Pascal VOC dataset using Faster-RCNN with a Resnet 50 backbone.

Table 15. Substitutions used for semantic labels which were not in the GloVe dataset.

Original Word	Substituted Word
baseball bat	baseball
baseball glove	baseball
dining table	table
fire hydrant	hydrant
parking meter	parking
playing field	field
potted plant	plant
tennis racket	racket
traffic light	stoplight
stop sign	stoplight
waterdrops	droplets

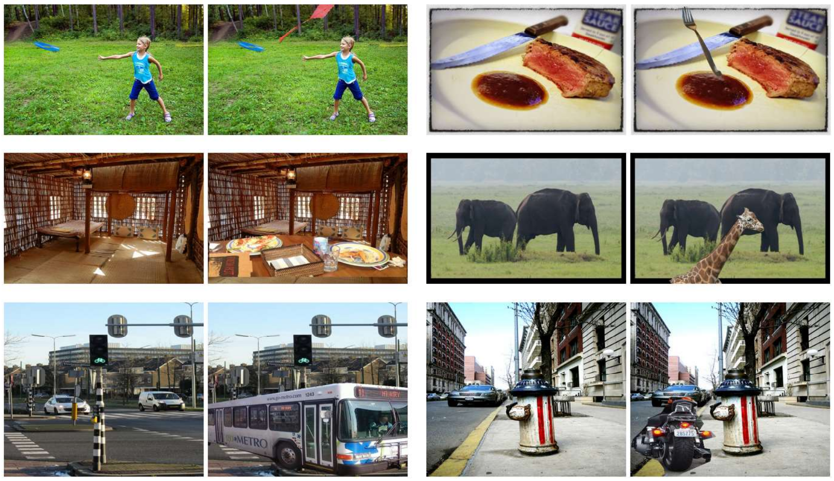

Fig. 8. Examples of original (left) vs semantically augmented (right) images.

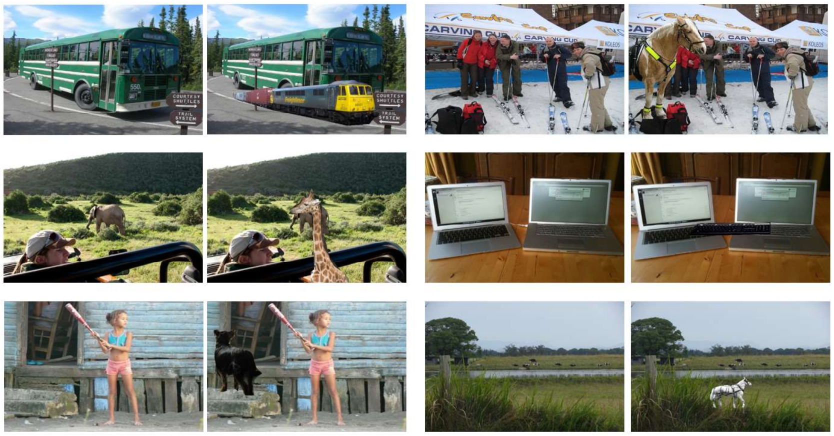

Fig. 9. Examples of original (left) vs semantically augmented (right) images.

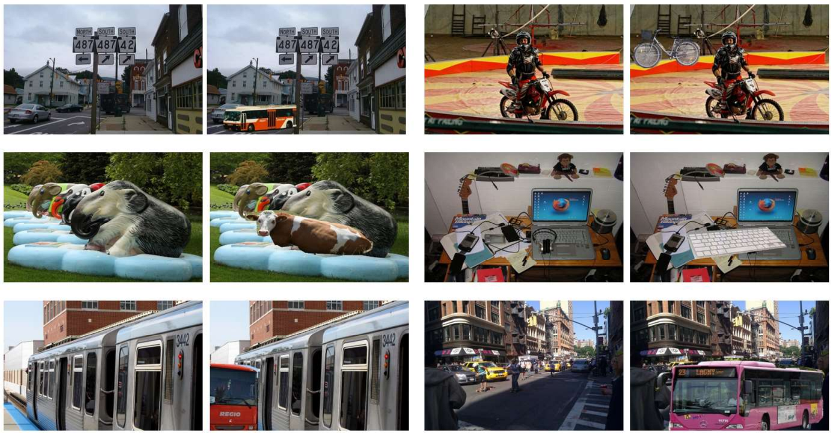

Fig. 10. Examples of original (left) vs semantically augmented (right) images.

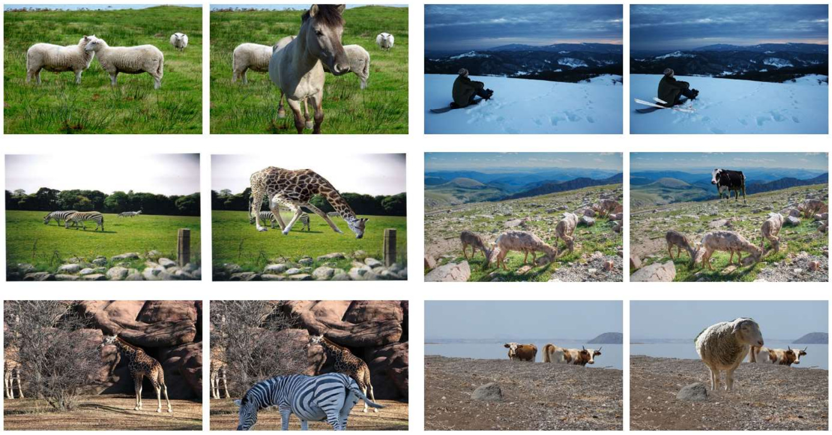

Fig. 11. Examples of original (left) vs semantically augmented (right) images.

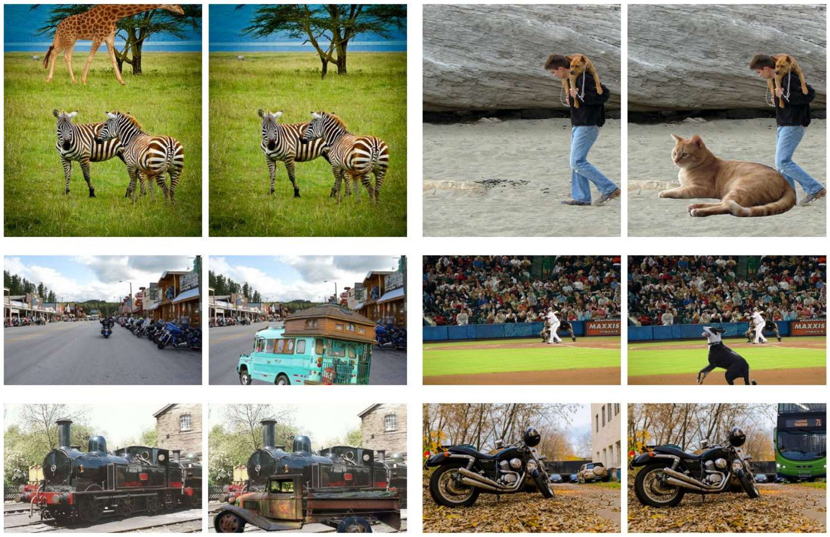

Fig. 12. Examples of original (left) vs semantically augmented (right) images.

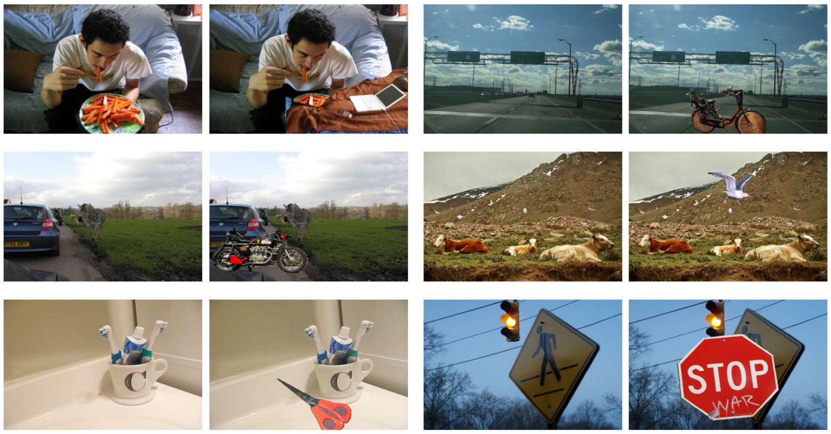

Fig. 13. Examples of original (left) vs semantically augmented (right) images.

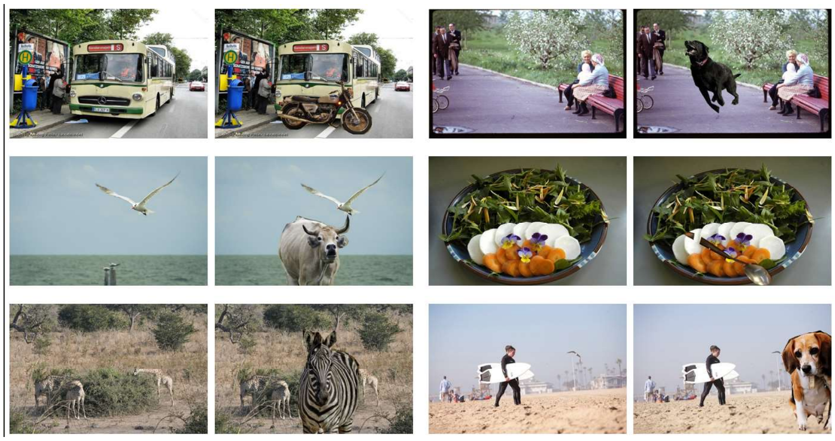

Fig. 14. Examples of original (left) vs semantically augmented (right) images.

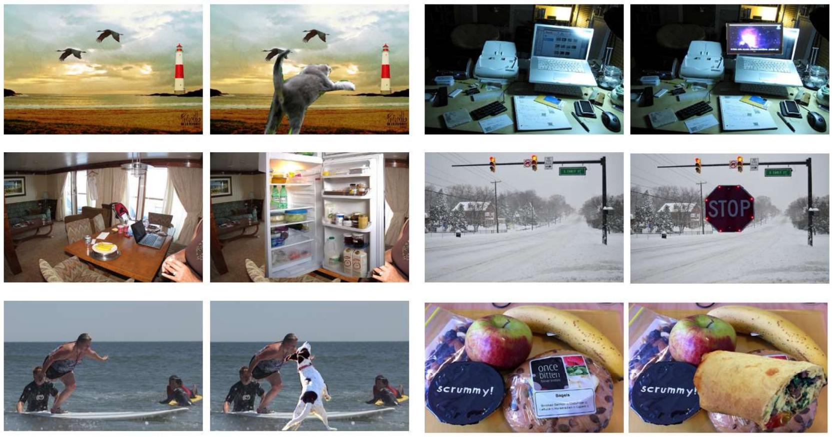

Fig. 15. Examples of original (left) vs semantically augmented (right) images.

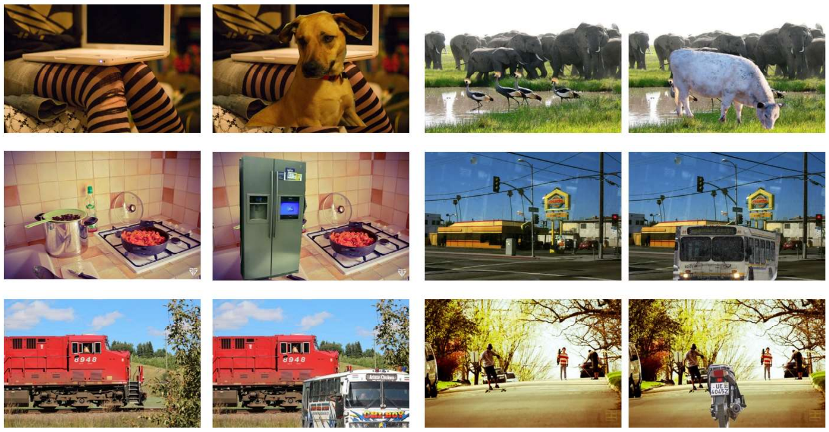

Fig. 16. Examples of original (left) vs semantically augmented (right) images.

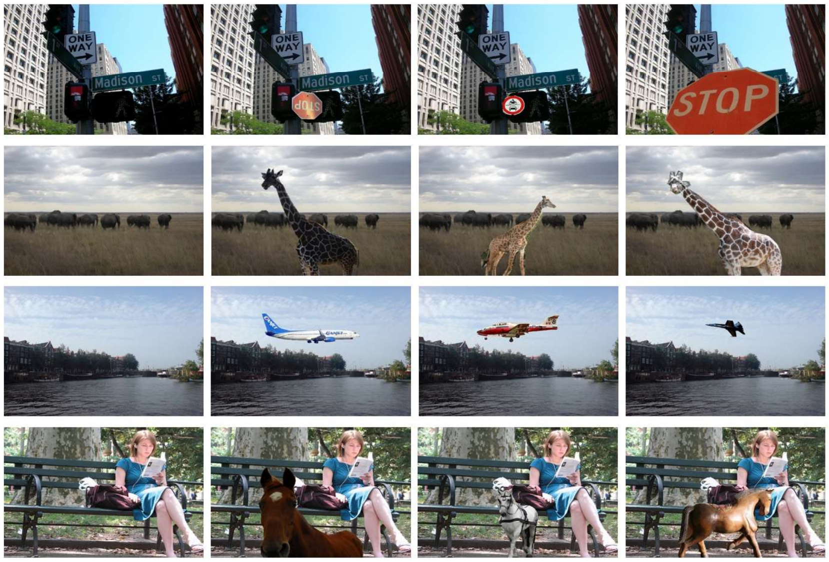

Fig. 17. Examples of original (left column) vs semantically augmented images. Different instances of a same object category are being inserted into the host image.

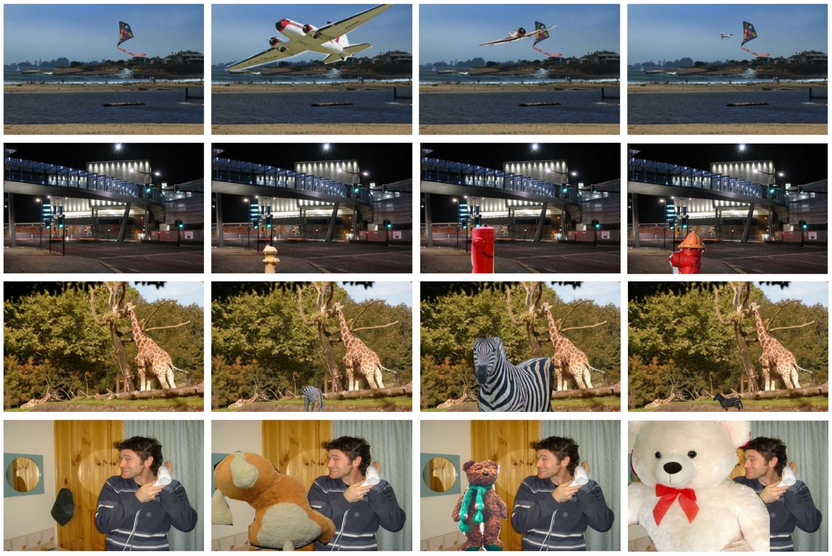

Fig. 18. Examples of original (left column) vs semantically augmented images. Different instances of a same object category are being inserted into the host image.

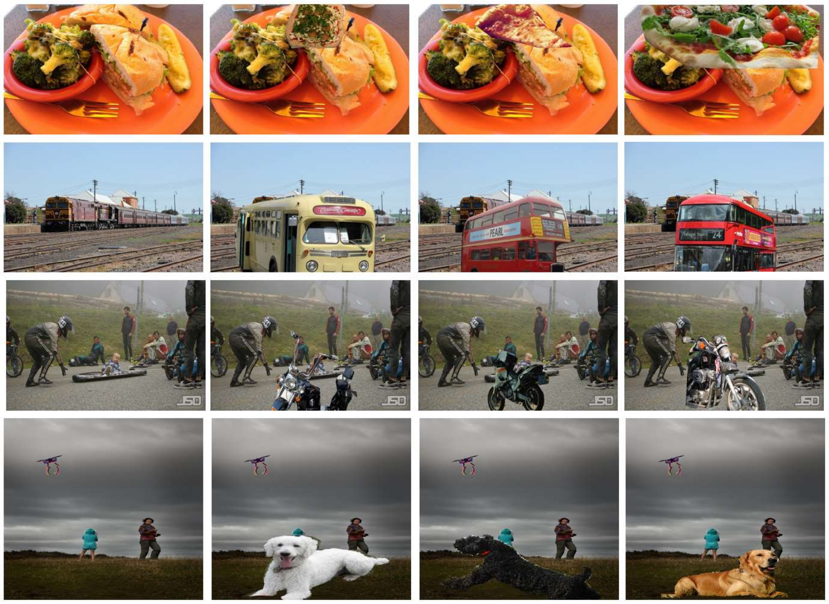

Fig. 19. Examples of original (left column) vs semantically augmented images. Different instances of a same object category are being inserted into the host image.

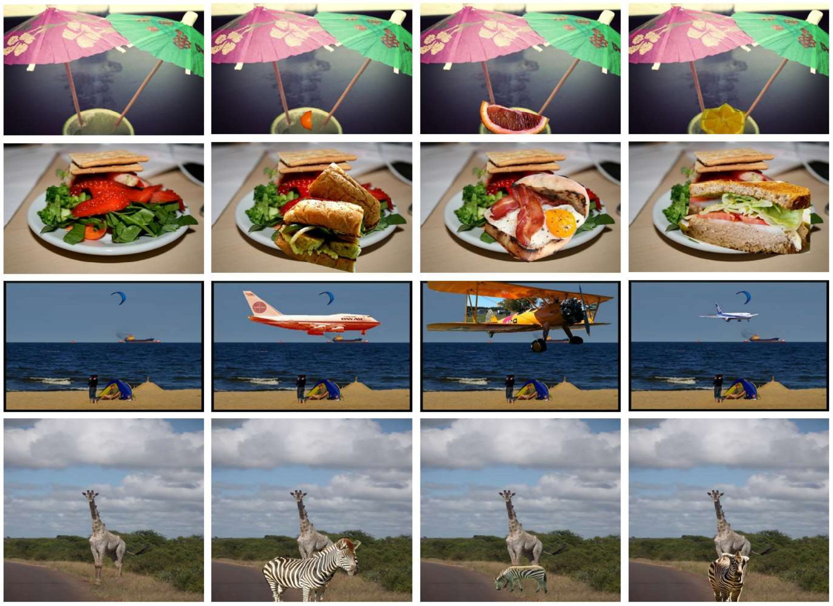

Fig. 20. Examples of original (left column) vs semantically augmented images. Different instances of a same object category are being inserted into the host image.

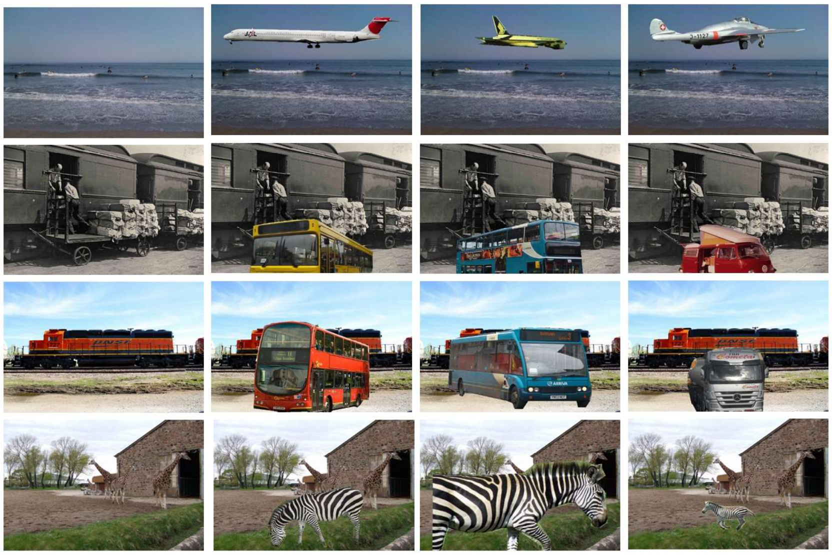

Fig. 21. Examples of original (left column) vs semantically augmented images. Different instances of a same object category are being inserted into the host image.

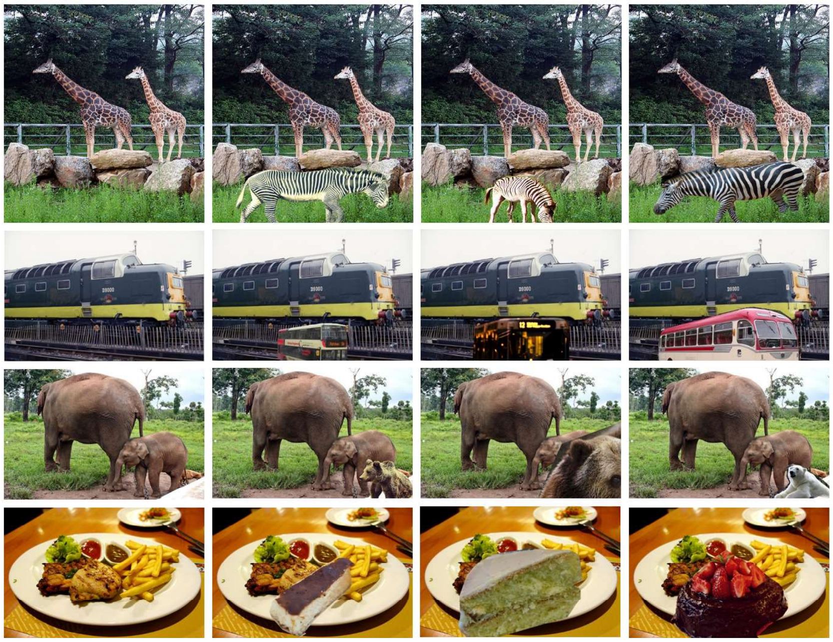

Fig. 22. Examples of original (left column) vs semantically augmented images. Different instances of a same object category are being inserted into the host image.

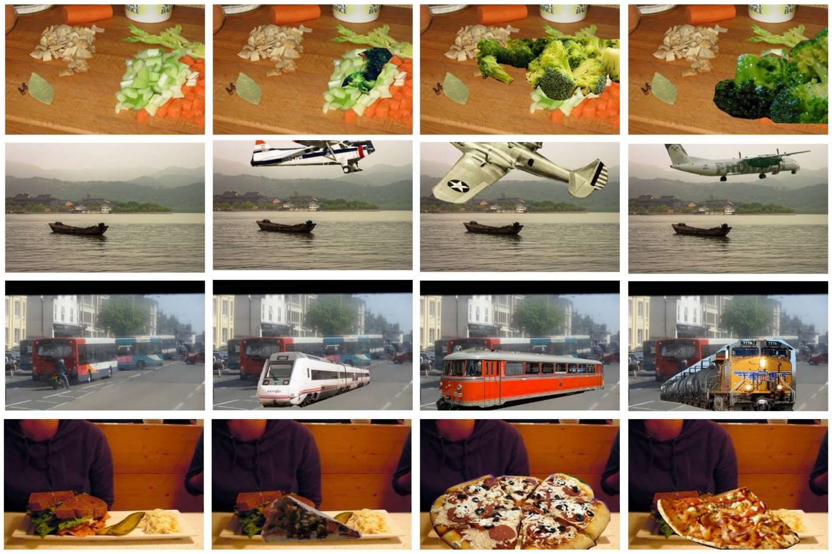

Fig. 23. Examples of original (left column) vs semantically augmented images. Different instances of a same object category are being inserted into the host image.

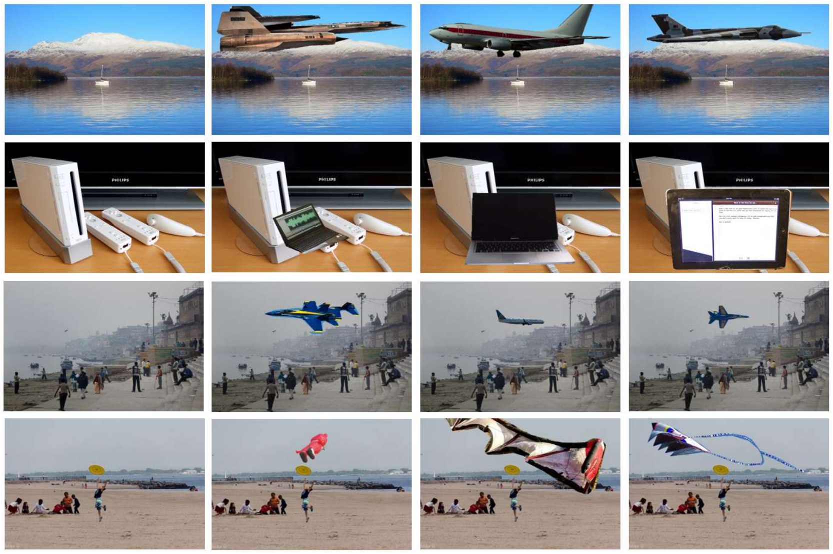

Fig. 24. Examples of original (left column) vs semantically augmented images. Different instances of a same object category are being inserted into the host image.

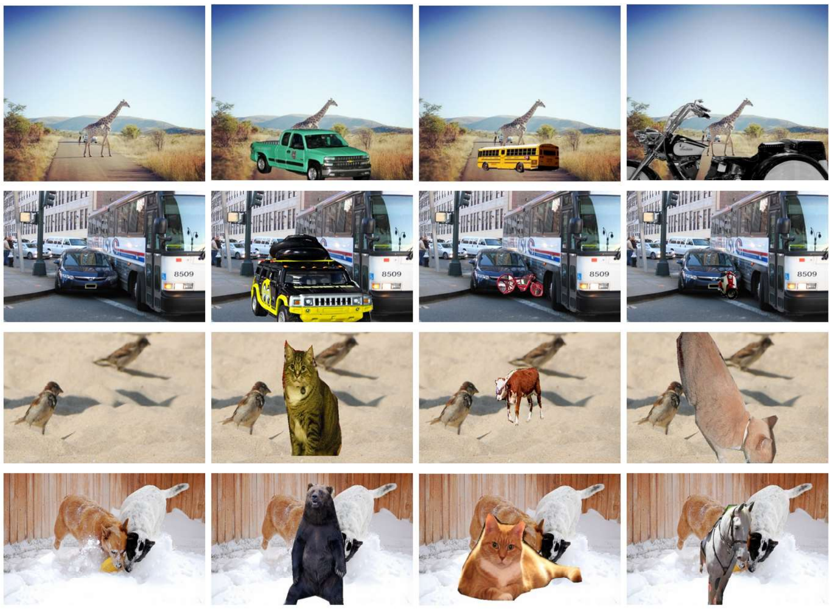

Fig. 25. Examples of original (left column) vs semantically augmented images. Different object categories are being inserted into the host image.

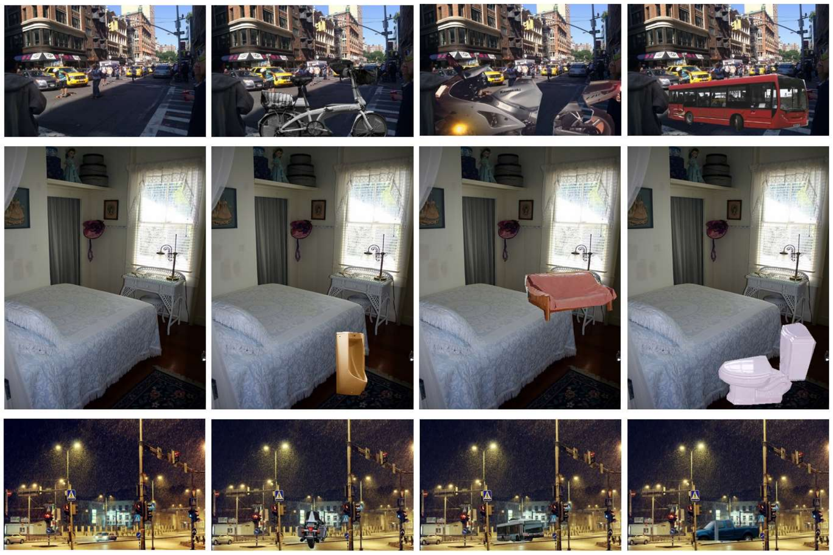

Fig. 26. Examples of original (left column) vs semantically augmented images. Different object categories are being inserted into the host image.

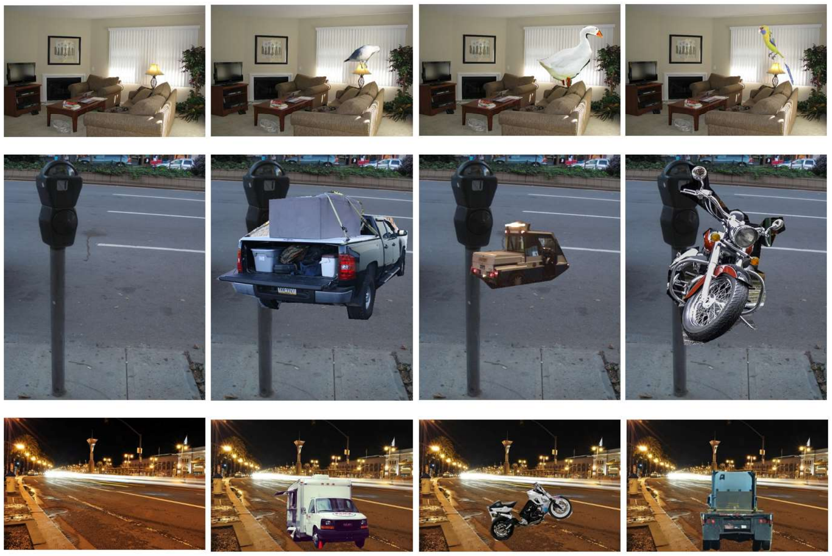

Fig. 27. Examples of original (left column) vs semantically augmented images. Different object categories are being inserted into the host image.

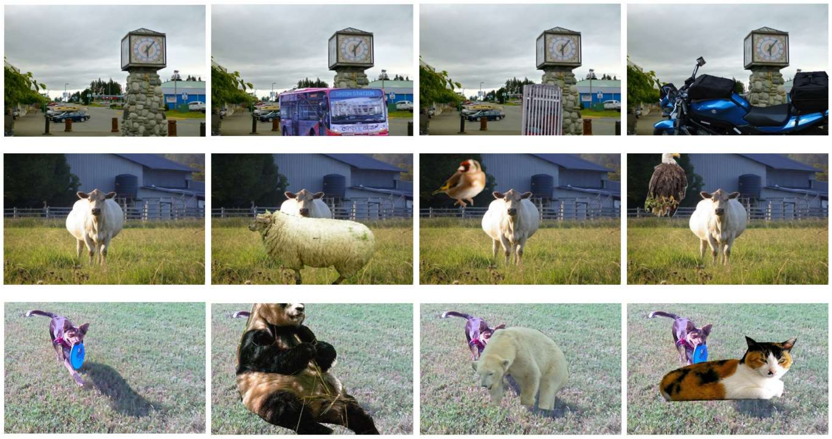

Fig. 28. Examples of original (left column) vs semantically augmented images. Different object categories are being inserted into the host image.

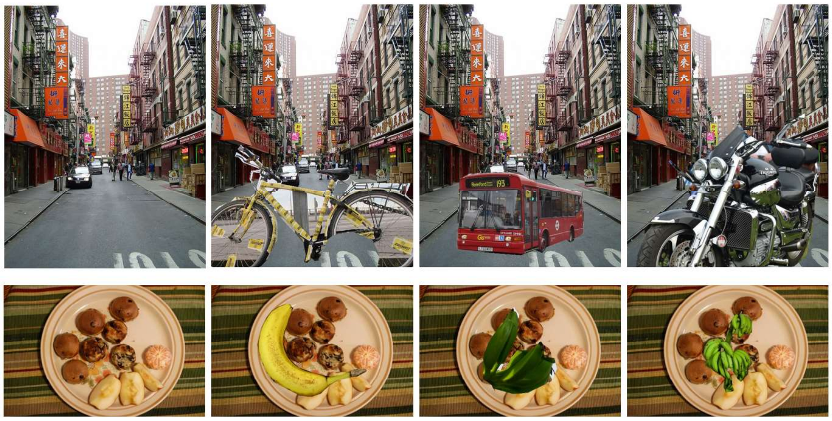

Fig. 29. Examples of original (left column) vs semantically augmented images. Different object categories are being inserted into the host image.