Supplementary Materials: Diverse Generation

Anonymous Authors

1 DIVERSE IMAGE GENERATION

We propose DAFT-GAN, an effective approach for integrating text and images. To validate the performance of our proposed model, we utilized the CUB-200-2011, Oxford-102, and MS-COCO datasets. Through both quantitative and qualitative evaluations, we verified that our proposed model outperforms other models. Furthermore, when conducting inference on a variety of classes and complex scenes, not limited to simple scenes or specific classes, we visually confirmed that our model largely produces natural and realistic images.

As shown in Fig.1, inference was performed on the CUB dataset, in Fig.2 on the Oxford-102 dataset, and in Fig.3 on the COCO dataset. While our model excels at generating natural images for CUB and Oxford-102, it particularly stands out in generating high-quality images across diverse scenarios, including challenging classes and complex scenes, within the COCO dataset. Previous inpainting models often struggled to produce high-quality results in challenging scenarios or complex scenes, such as generating images of people or intricate patterns. However, our model demonstrates good quality generation across diverse situations, including animals, humans, and complex indoor environments.

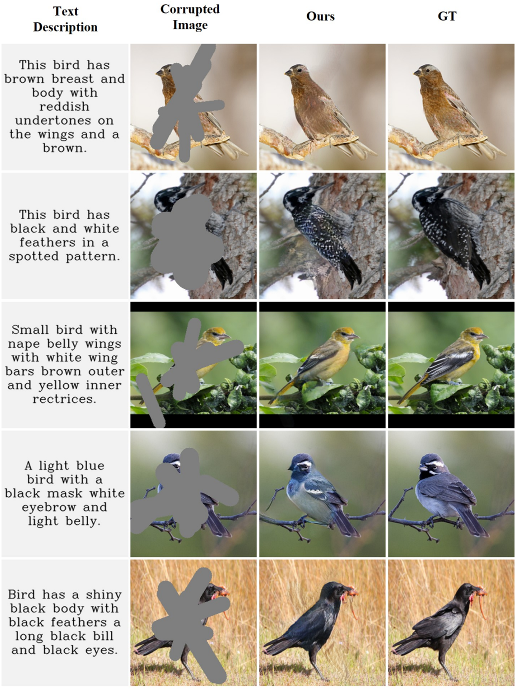

Text Description
Corrupted Image
Ours
GT
This bird has brown breast and body with reddish undertones on the wings and a brown.
This bird has black and white feathers in a spotted pattern.
Small bird with nape belly wings with white wing bars brown outer and yellow inner rectrices.
A light blue bird with a black mask white eyebrow and light belly.
Bird has a shiny black body with black feathers a long black bill and black eyes.

Figure 1: Results of our proposed model on CUB-200-2011 dataset. Corrupted (left), generated (middle), and ground-truth (right) images are presented.

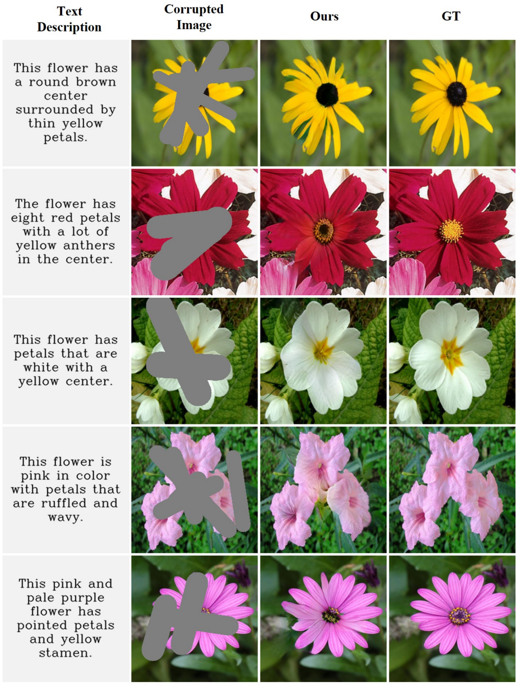

Text Description
Corrupted Image
Ours
GT
This flower has a round brown center surrounded by thin yellow petals.
The flower has eight red petals with a lot of yellow anthers in the center.
This flower has petals that are white with a yellow center.
This flower is pink in color with petals that are ruffled and wavy.
This pink and pale purple flower has pointed petals and yellow stamen.

Figure 2: Results of our proposed model on Oxford-102 dataset. Corrupted (left), generated (middle), and ground-truth (right) images are presented.

2 IMPORTANCE OF TRAINING MASK DIVERSITY FOR ENHANCING ROBUSTNESS

When conducting quatitative and qualitative evaluations of our proposed DAFT-GAN as well as other inpainting models, we trained all comparison models using diverse irregular masks with a wide range of masking ratios from around 10% to 70% to improve its robustness. This is also aimed at solely focuses on evaluating performance and effectiveness based on the inpainting method. In contrast, using only biased masks can make the model's robustness highly vulnerable, as shown in Fig.4. The previous MMFL trained the model using only 25% center masks, and the results were satisfactory when evaluated on the same 25% center masks, as shown in Fig.5. However, when the same model was evaluated with diverse irregular masks, the overall results became poor, and the areas outside the center mask that did not need to be recovered during training became completely degraded, essentially becoming noise-like. This demonstrates that when training with biased masks, the model's weights also become biased towards only being able to recover for those specific mask patterns. For the experiments in this paper, we directly trained and evaluated the other models, including MMFL, using the same diverse irregular masks, to enable an accurate comparative analysis.

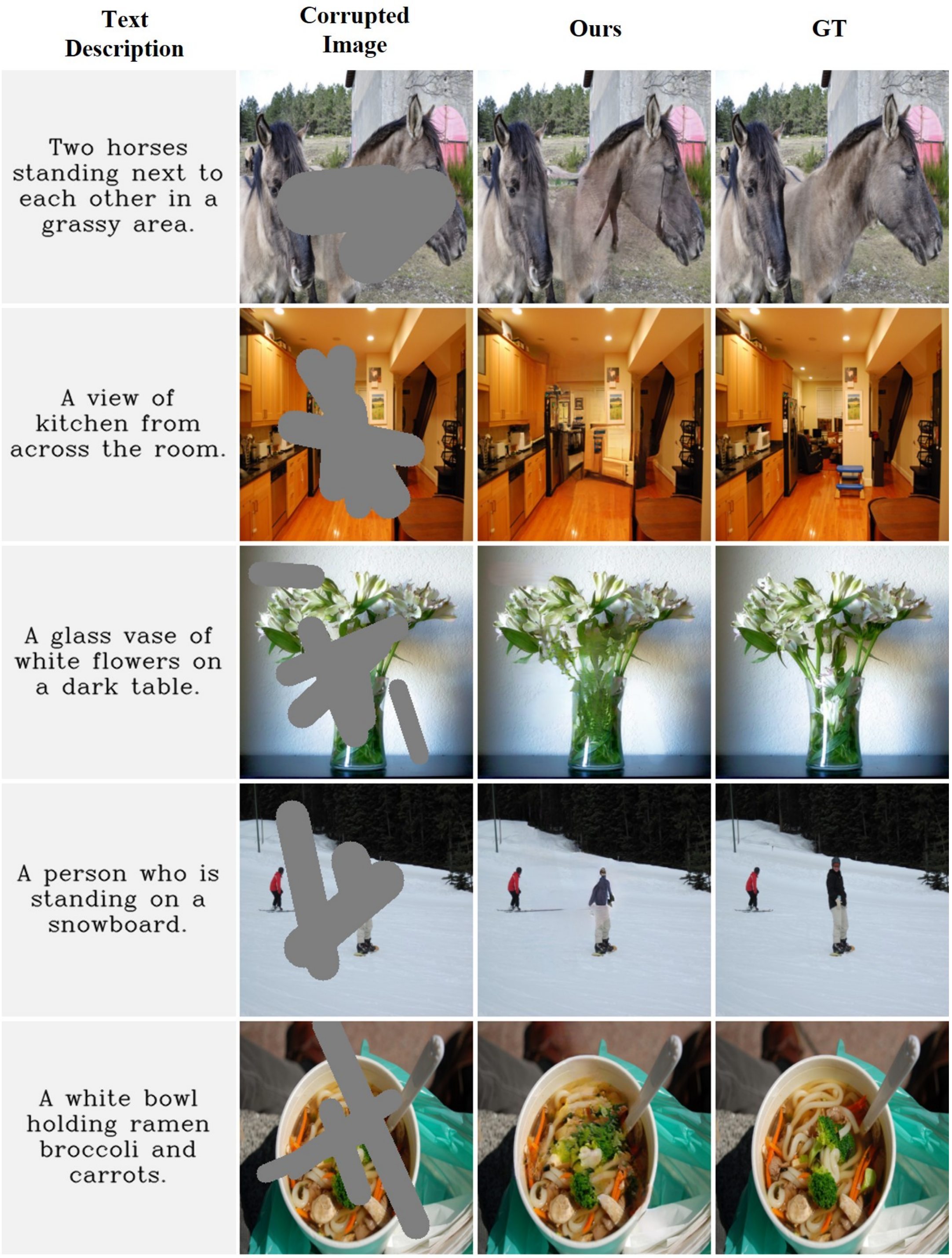
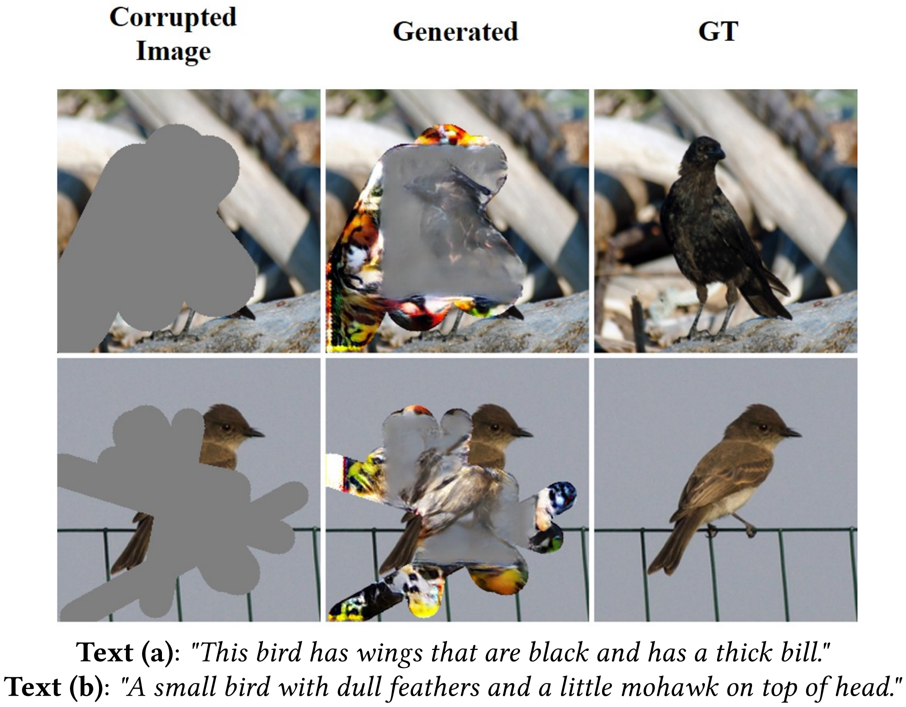
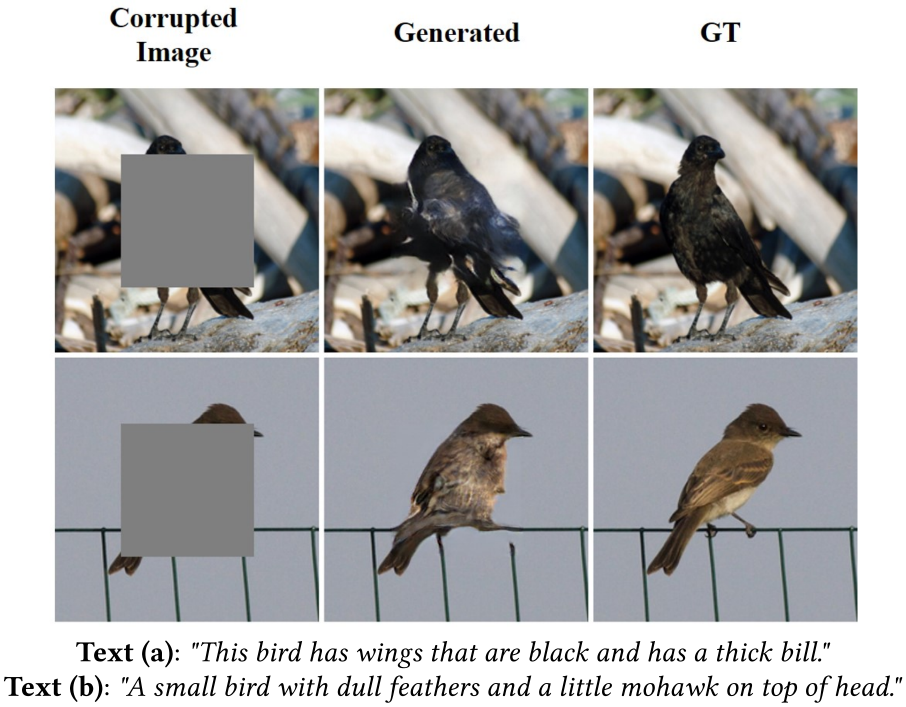

Text Description
Corrupted Image
Ours
GT
Two horses standing next to each other in a grassy area.
A view of kitchen from across the room.
A glass vase of white flowers on a dark table.
A person who is standing on a snowboard.
A white bowl holding ramen broccoli and carrots.

Figure 3: Results of our proposed model on MS-COCO dataset. Corrupted (left), generated (middle), and ground-truth (right) images are presented.



Text (a): *"This bird has wings that are black and has a thick bill."*
Text (b): *"A small bird with dull feathers and a little mohawk on top of head."*

Figure 4: Generatated images with diverse irregular masks, trained with center masks. Corrupted (left), generated (middle), and ground-truth (right) images are presented. Text(a) (first row), Text(b) (second row) are demonstrated.



Text (a): *"This bird has wings that are black and has a thick bill."*
Text (b): *"A small bird with dull feathers and a little mohawk on top of head."*

Figure 5: Generatated images with center masks, trained with center masks. Corrupted (left), generated (middle), and ground-truth (right) images are presented. Text(a) (first row), Text(b) (second row) are demonstrated.