DDB: Diffusion Driven Balancing to Address Spurious Correlations

Supplementary Material

1. Dataset Details

Waterbirds. This dataset is introduced by [6] and is a synthesized collection created by cropping birds from the CUB dataset [7] and pasting them onto backgrounds from the Places dataset [10]. In this dataset, seabirds and waterfowl are labeled as “Waterbirds,” while other birds are labeled as “Landbirds.” The backgrounds are categorized into two types: ocean and natural lake for the water background, and bamboo forest and broadleaf forest for the land background.

In the dataset, waterbirds with a water background and landbirds with a land background are considered the majority groups, while waterbirds with a land background and landbirds with a water background are considered the minority groups. The number of samples in each group is shown in Table 1.

CelebA. Following [6], we utilized the ”blond” and ”non-blond” attributes from celebrity images in [4] as labels. In this dataset, there is a spurious correlation between gender and label. The majority groups are typically considered to be blond females and non-blond males, while the minority groups are considered to be blond males and non-blond females. However, as mentioned in the paper, the number of non-blond females and males are almost equal, and thus, this dataset only have one minority group. We also used the class-balanced dataset for our approach, similar to [5]. The number of samples in each group is shown in Table 2.

Metashift. This dataset was introduced by [3] for a binary classification task with labels ”dog” and ”cat.” The dataset exhibits a diversity shift, as categorized by [9], where test images contain different spurious attributes compared to the training set. Following [8], cat images in the training set have spurious attributes, such as ”bed” or ”sofa” in the background, while dog images have spurious attributes, such as ”bench” and ”bike.” In contrast, the test images do not contain these attributes; instead, the spurious attribute in the test images is ”shelf.” The number of samples in each group is shown in Table 3.

Table 1. The number of samples in the training, validation, and test sets for the Waterbirds dataset

Split	Train	Validation	Test
(landbird, land background)	3,498	467	2,255
(landbird, water background)	184	466	2,255
(waterbird, land background)	56	133	642
(waterbird, water background)	1,018	133	642

Table 2. The number of samples in the training, validation, and test sets for the CelebA dataset

Split	Train	Validation	Test
(NonBlond, female)	12,426	8,535	9,767
(NonBlond, male)	11,841	8,276	7,535
(Blond, female)	22,880	2,874	2,480
(Blond, male)	1,387	182	180

Table 3. The number of samples in the training, validation, and test sets for the Metashift dataset

Split	Train	Validation	Test
(Cat, sofa)	231	0	0
(Cat, bed)	380	0	0
(Dog, bench)	145	0	0
(Dog, bike)	367	0	0
(Cat, shelf)	0	34	201
(Dog, shelf)	0	47	259

The time required to generate new samples is shown in Table 4 using an A100 GPU. This includes the masking process, generation using the diffusion model, and the computation of the Integrated Gradient (IG) score.

Table 4. Generation time in datasets.

Generation time	Metashift	CelebA	Waterbirds
(hours)	0.5	5	1

1.1. Details on CelebA and Spurious Features

In addition to spurious objects (such as backgrounds in the Waterbirds dataset or gender-based facial attributes in the CelebA dataset), spurious features can also exist within datasets. In the CelebA dataset, there is a correlation between hair color and other hair attributes, such as hair length or shape. For example, there is an imbalance between long blond hair and long non-blond hair, as well as between wavy blond hair and wavy non-blond hair [5]. Our approach makes the ERM model robust to both of these spurious correlations by balancing the dataset in both ways. In Figure 1, we show multiple images that are not wavy and have short blond hair, addressing this type of spurious correlation.

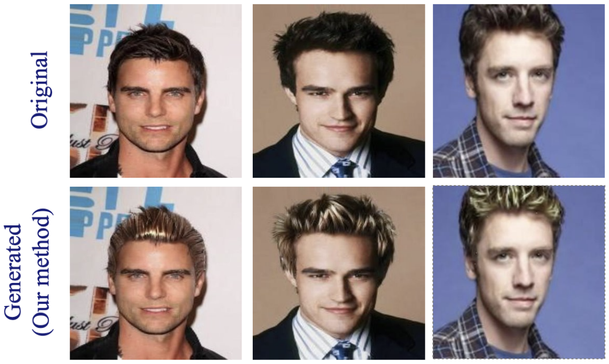

Figure 1. Generated samples in the CelebA dataset. The generated samples exhibit the spurious features of the non-blond class, as they are short and not wavy.

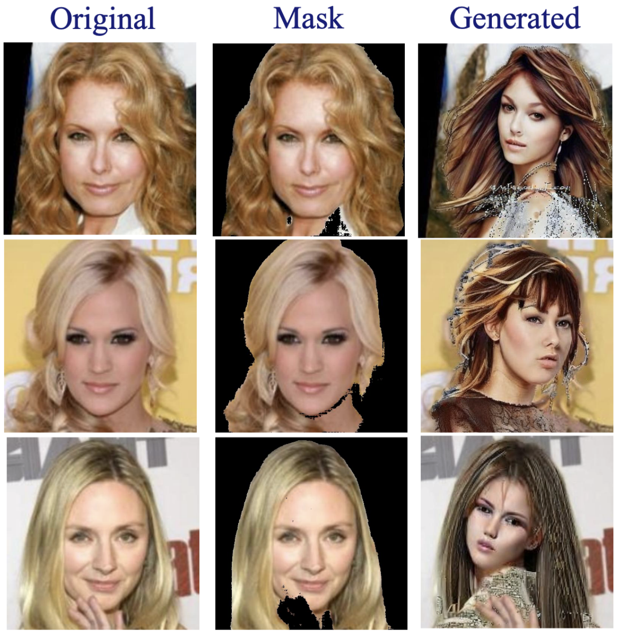

Figure 2. Generated non-blond females in cases where masking fails.

2. Experiments

2.1. DDB Training

ERM Model. We utilized the ERM model for the CelebA and Waterbirds datasets from [5], and the Metashift dataset from [8].

Hyperparameters. For textual inversion training, we utilized the Huggingface implementation. The detailed hyperparameters for our method on the three datasets are shown in Table 5.

Algorithm. A formal algorithm of our method is provided in Algorithm 1.

Algorithm 1: Diffusion-Driven-Balancing (DDB)

```
Input: Training Dataset $\mathcal{D}_{train}$, Class $i$,$i'$,
       Inpainting Model $SD$, Masking Model $m$,
       Hyperparametes $K_i, K_{i'}, \gamma_1, \gamma_2, \mathrm{P}_i, \mathrm{P}_{i'}$
$Textual\_Dset_i \leftarrow$
  Pick a few samples $(x, y) \in \mathcal{D}_{train}$;
$Textual\_Dset_{i'} \leftarrow$
  Pick a few samples $(x, y') \in \mathcal{D}_{train}$;
$C_i^* = TextualInversion(SD, Textual\_Dset_i)$;
$C_{i'}^* = TextualInversion(SD, Textual\_Dset_{i'})$;
$\mathcal{D}_i \leftarrow SaveLowLossSamples(\mathcal{D}_{train}, i, K_i)$;
$\mathcal{D}_{i'} \leftarrow SaveLowLossSamples(\mathcal{D}_{train}, i', K_{i'})$ ;
// Equation ??
$\Psi_i \leftarrow MeanSoftMax(\mathcal{D}_i, i)$ ;
$\Psi_i \leftarrow MeanSoftMax(\mathcal{D}_{i'}, i')$;
$\mathcal{D'}_{i'} =$
  $GenerateNewDataset(f, i, i', \Psi_i, \mathrm{P}_{i'}, m, C_{i'}^*, SD)$;
$\mathcal{D'}_i =$
  $GenerateNewDataset(f, i', i, \Psi'_i, \mathrm{P}_i, m, C_i^*, SD)$;
$\mathcal{D}_{train} \leftarrow \mathcal{D}_{train} \cup \mathcal{D'}_i \cup \mathcal{D'}_{i'}$ ;
for epoch = 1,2,3...,O do
    for batch $\mathcal{B} in \mathcal{D}_{tr}$ do
        $B'_i = Samples from \mathcal{B} \in \mathcal{D'}_i$;
        $B'_{i'} = Samples from \mathcal{B} \in \mathcal{D'}_{i'}$;
        $L_{CE} = \frac{1}{|B|} \sum_{(x,y) \in B} l(f_\theta(x), y)$;
        $L_{gen1} = \frac{1}{|B'_i|} \sum_{(x,y) \in B'_i} l(f_\theta(x), y)$;
        $L_{gen2} = \frac{1}{|B'_{i'}|} \sum_{(x,y) \in B'_{i'}} l(f_\theta(x), y')$;
        $L_{total} = L_{CE} + \gamma_1 L_{gen1} + \gamma_2 L_{gen2}$;
        $f \leftarrow UpdateWeights(L_{total})$
    end
end
```

2.2. Textual inversion

Figure 3 shows the impact of the textual inversion dataset size. As demonstrated, the optimal number of samples leads to the specified token better representing the causal parts, resulting in higher-quality images.

We selected minority group samples for the textual inversion dataset to ensure that the learnable token does not reconstruct the spurious feature s_{maj}, thereby preventing error propagation when other components fail. For instance, in the CelebA dataset, where the mask is intended to isolate the hair, if the mask encompasses both the face and the hair, Stable Diffusion will regenerate the face in accordance with the learned representation of the minority group samples, provided the token has been trained to capture such features. As shown in Figure 2, the mask includes the face region as well. However, the learned token generates a non-

Table 5. Hyperparameters. The batch size and learning rate are for the retraining phase of the ERM model.

Dataset	Batch size	Learning rate	γ_1	γ_2	P_1	P_2
CelebA	64	0.00001	2	7	2	0.3
Metashift	16	0.00005	1	4	0.3	2
Waterbirds	32	0.000005	6	10	1	1

blond woman, as the text embedding model was trained on a dataset of non-blond women.

2.3. Pruning

As mentioned in the paper, in the Metashift dataset, many samples in the dog class are too small compared to the image or other animals present, leading the masking model to mistakenly identify these non-relevant parts as the causal component. Some examples of these problematic samples are shown in Figure 4. In the CelebA dataset, identifying hair is challenging for the masking model, as shown in Figure 1. To address this limitation, we use textual inversion, as explained in § 2.2. Additionally, many bald samples wearing hats exist in the non-blond class, which results in corrupted generated images that need to be pruned. An example of this case is shown in Figure 5.

2.4. Imagenet-A

For further experiments and to evaluate DDB on more general tasks, we assessed our approach on two classes from ImageNet-A [2]—natural adversarial examples: *Cockroach* and *Bee*. This dataset contains samples on which standard machine learning models typically perform poorly. Our method improved the average accuracy of the ERM baseline on this dataset from 43.26% to 62.58%. We used ImageNet-1k [1] as the source of training samples. The ERM model was trained using the SGD optimizer with a learning rate of 0.001, while the retraining phase used the Adam optimizer with a learning rate of 0.0005. The batch size was set to 32 for both stages. We also set the hyperparameters $\gamma_1 = 1$ and $\gamma_2 = 1$ in this experiment. Qualitative images are available in Fig. 6.

2.5. Qualitative images

In this section, we provide additional qualitative examples across datasets. Figures 9 and 10 present generated images from the Waterbirds dataset, Figures 7 and 8 show results from the CelebA dataset, and Figures 11 and 12 show results from the Metashift dataset. Each figure showcases successful generations of minority group samples across different classes.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 3

[2] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples, 2021. 3

[3] Weixin Liang and James Zou. Metashift: A dataset of datasets for evaluating contextual distribution shifts and training conflicts, 2022. 1

[4] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild, 2015. 1

[5] Fahimeh Hosseini Noohdani, Parsa Hosseini, Aryan Yazdan Parast, Hamidreza Yaghoubi Araghi, and Mahdieh Soleymani Baghshah. Decompose-and-compose: A compositional approach to mitigating spurious correlation, 2024. 1, 2

[6] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization, 2020. 1

[7] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. 2010. 1

[8] Shirley Wu, Mert Yuksekgonul, Linjun Zhang, and James Zou. Discover and cure: Concept-aware mitigation of spurious correlation, 2023. 1, 2

[9] Nanyang Ye, Kaican Li, Haoyue Bai, Runpeng Yu, Lanqing Hong, Fengwei Zhou, Zhenguo Li, and Jun Zhu. Ood-bench: Quantifying and understanding two dimensions of out-of-distribution generalization, 2022. 1

[10] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Antonio Torralba, and Aude Oliva. Places: An image database for deep scene understanding, 2016. 1

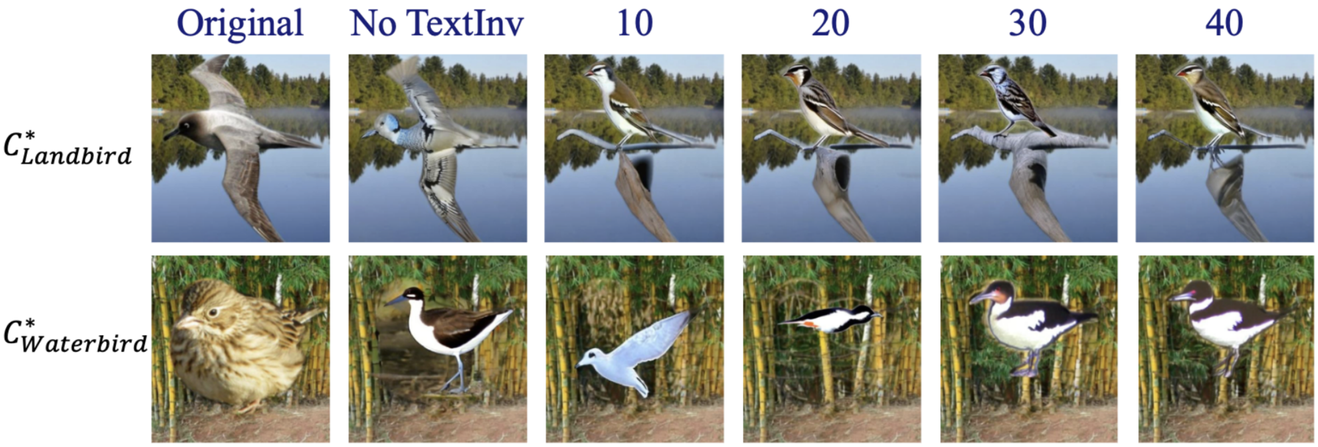

Figure 3. Examples of generated samples for landbirds and waterbirds with varying textual inversion dataset sizes, as well as samples generated without utilizing textual inversion and the initial image.

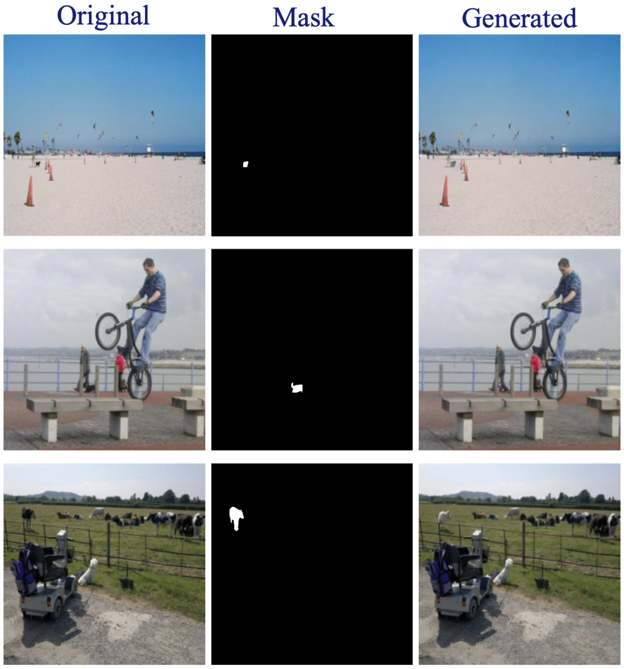

Figure 4. Metashift samples when the stable diffusion cannot perform well because of the initial image.

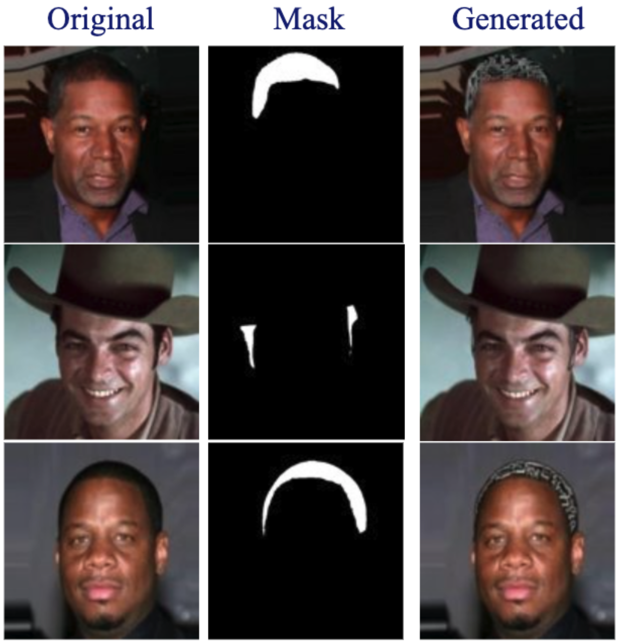

Figure 5. Example of bald samples or samples wearing hat in the CelebA dataset.

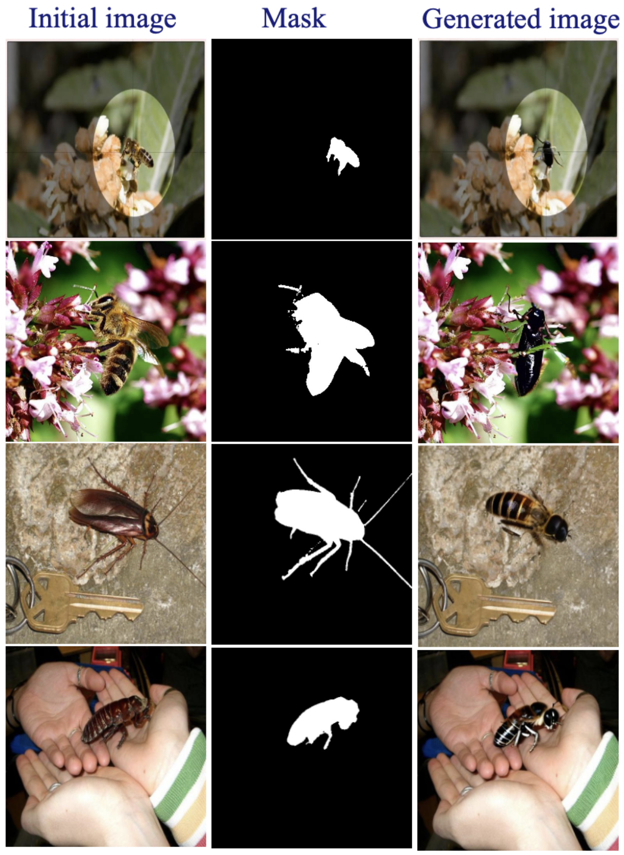

Figure 6. Examples of generated samples along with the corresponding masks and initial images. The generated images belong to the cockroach class, with initial images taken from the bee class, and vice versa.

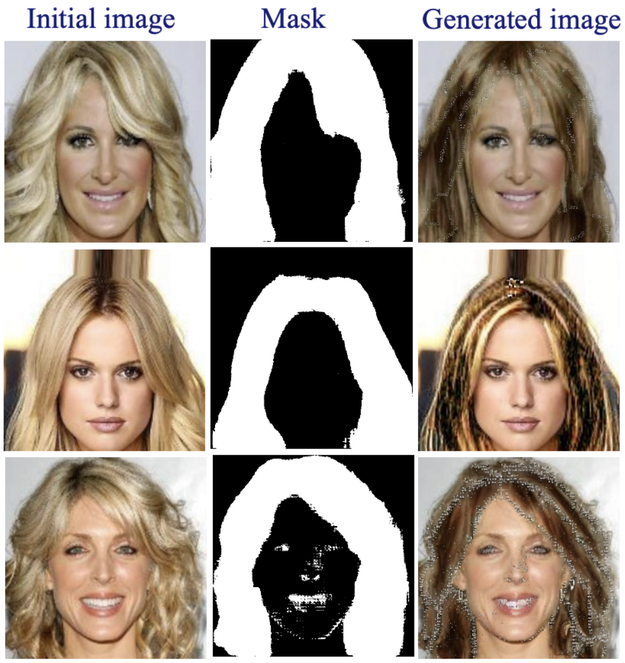

Figure 7. Examples of generated female samples along with the corresponding masks and initial images. The initial images are belong to male class.

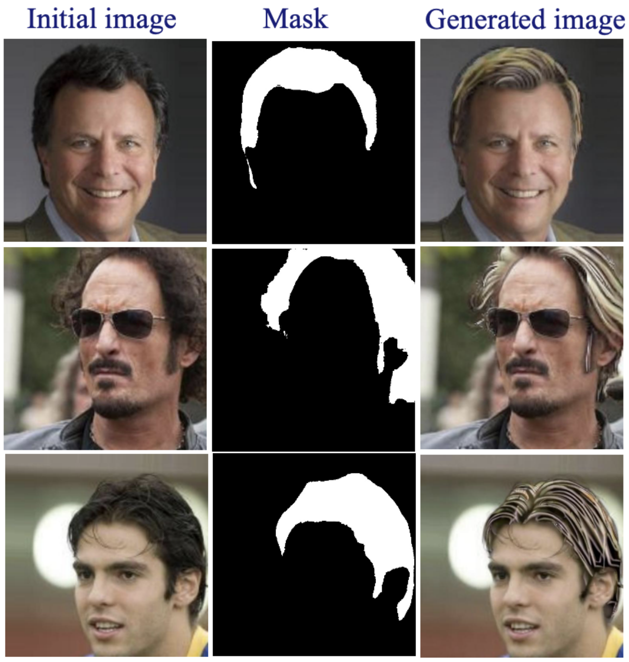

Figure 8. Examples of generated male samples along with the corresponding masks and initial images. The initial images are belong to female class.

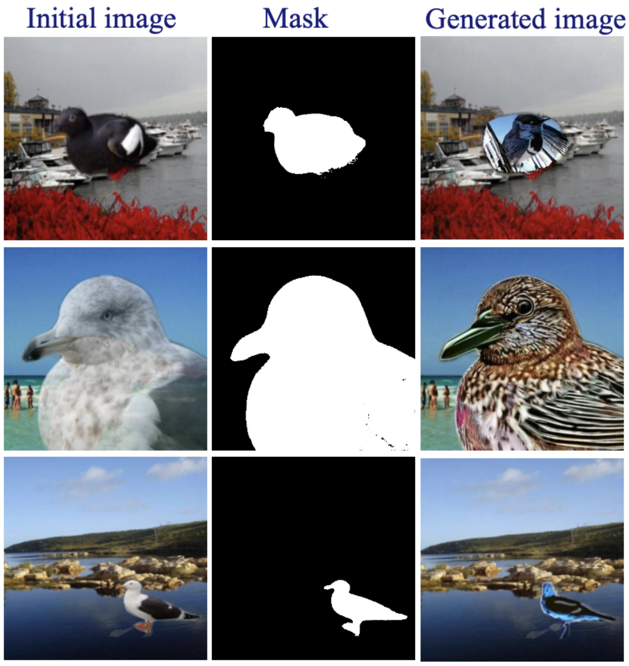

Figure 9. Examples of generated landbird samples along with the corresponding masks and initial images. The initial images are belong to waterbird class.

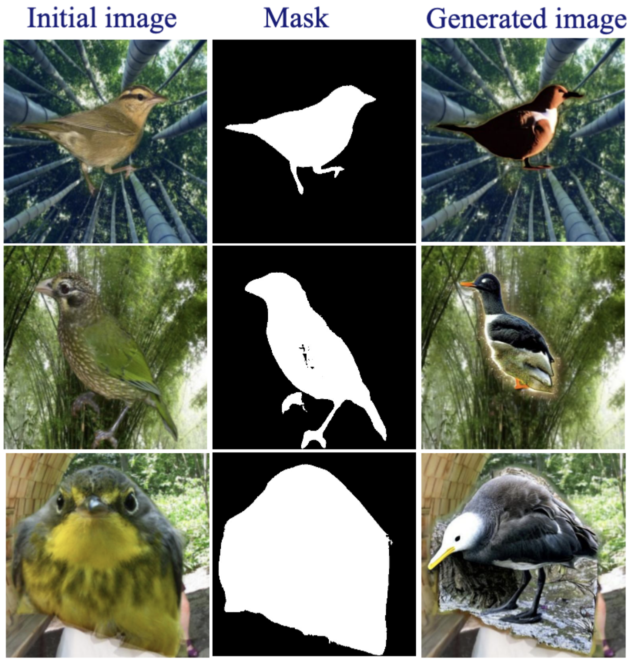

Figure 10. Examples of generated waterbird samples along with the corresponding masks and initial images. The initial images are belong to landbird class.

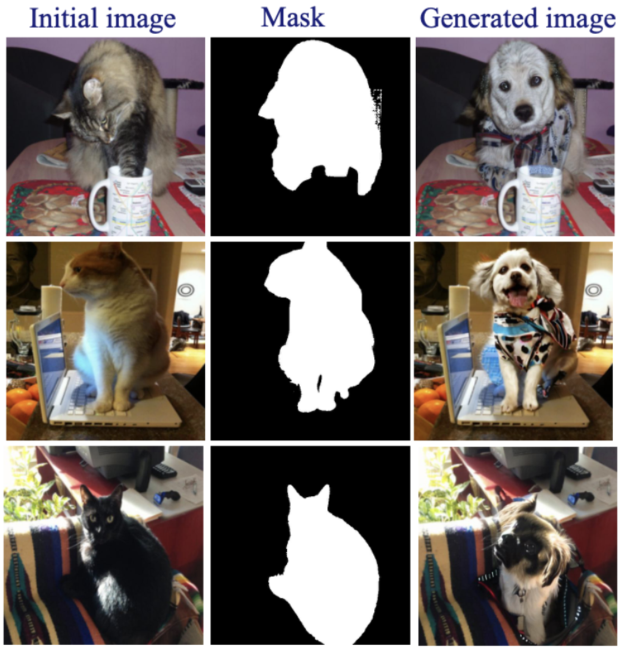

Figure 11. Examples of generated dog samples along with the corresponding masks and initial images. The initial images are belong to cat class.

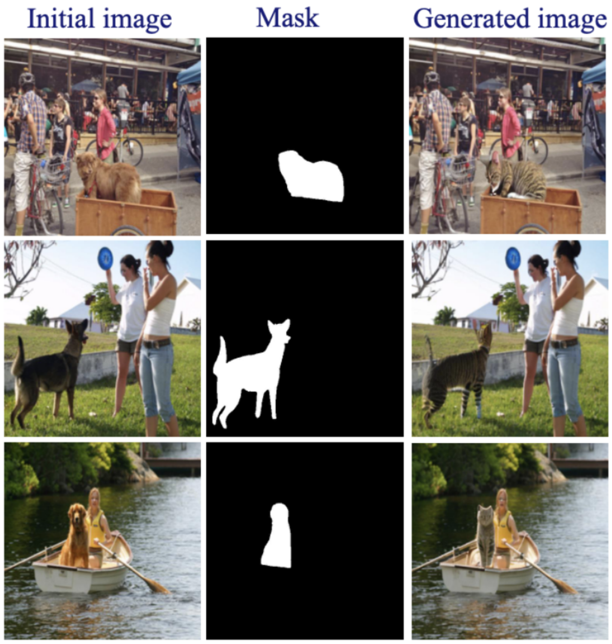

Figure 12. Examples of generated cat samples along with the corresponding masks and initial images. The initial images are belong to dog class.