# Supplementary: Unsupervised Co-generation of Foreground-Background Segmentation from Text-to-Image Synthesis

Yeruru Asrar Ahmed and Anurag Mittal
Department of Computer Science and Engineering,
Indian Institute of Technology Madras
{asrar,amittal}@cse.iitm.ac.in

## 1. Additional Studies

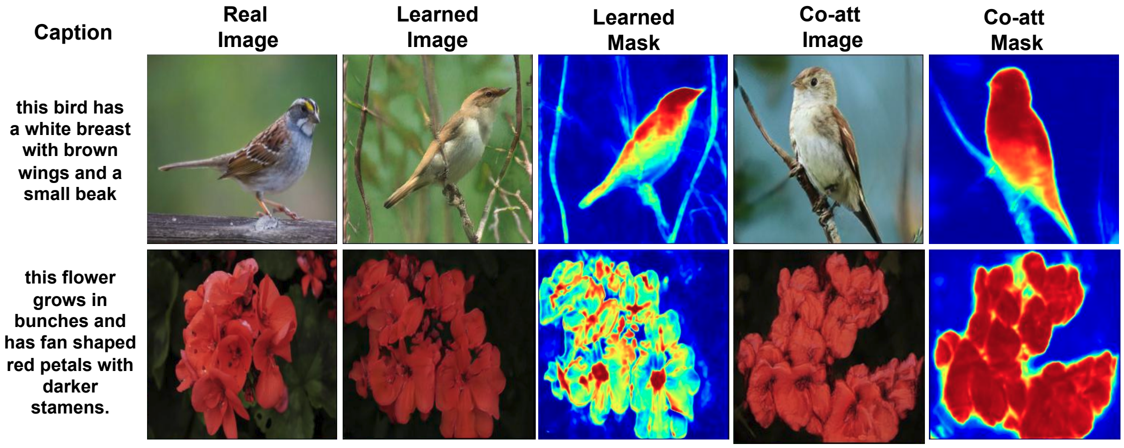
Caption
Real Image
Learned Image
Learned Mask
Co-att Image
Co-att Mask
this bird has a white breast with brown wings and a small beak
this flower grows in bunches and has fan shaped red petals with darker stamens.

Figure 1. Visual comparision of masks generated by Co-attention (Co-att) and Learned attention (Learned) approaches conditioned on the caption for CUB and Oxford-102 datasets.

| Dataset | Method | IS | FID | R% | ACC | IoU | mIoU |
|---|---|---|---|---|---|---|---|
| **CUB** | **Learned** | 5.12 ± .01 | 13.37 | 81.5 | 87.5 | 71.7 | 79.8 |
| | **Co-attn** | **5.24 ± .06** | **12.42** | **86.53** | **94.6** | **73.2** | **83.3** |
| **Oxford** | **Learned** | 4.14 ± .03 | 30.39 | 78.47 | 85.4 | 69.9 | 76.5 |
| | **Co-attn** | **4.28 ± .09** | **28.63** | **79.63** | **90.9** | **77.2** | **81.7** |

Table 1. Quantitative comparison of model using Co-attention based mechanism (Co-att) and that of learned attention (Learned) approach for T2I on CUB and Oxford-102 datasets.

### 1.1. Learned attention mask vs. Co-attention based segmentation mask

COS-GAN uses a co-attention-based SCM predictor to estimate segmentation masks on image attended features. This approach involves extra computation to perform a correlation of image feature and reference feature. To address this issue, we attempt to use a learned attention approach, as in SSA-GAN [7], to predict FG-BG segmentation masks directly on image features. However, as shown in Figure 1, this approach leads to a drop in the quality of the predicted segmentation masks.

Therefore, it is evident that performing a correlation between simultaneously generated multiple images leads to high-quality segmentation masks.

In Table 1, we have quantitatively compared the performance of Text-to-Image synthesis and quality of masks generated by COS-GAN using the co-attention-based SCM predictor (against the model with learned attention masks in Table 1). Co-attention approach achieves slightly better performance, which we attribute to its ability to extract high-confidence segmentation masks used in the Spatial Conditioning blocks. Moreover, co-attention-based approach results in a significant enhancement in the quality of the FG-BG masks produced. This improvement is because the co-attention approach considers the overall global structure of the reference images when predicting the masks, leading to better quality in generated masks.

## 1.2. Sentence with Words for conditioning vectors

Proposed method uses conditioning augmentation with sentence features concatenated with noise and words to generate initial conditioning. Previous methods have utilised only sentence and noise to generate low-resolution features using convolutional layers and have improved performance with cross-modal attention in deep layers [6, 19, 20, 22]. We propose using transformers on sentences with words to capture better global representation.

| Dataset | Method | IS | FID | R% | ACC | IoU | mIoU |
|---|---|---|---|---|---|---|---|
| CUB | S+N | $5.12 \pm .06$ | 14.19 | 81.76 | 92.3 | 71.5 | 81.6 |
| | W+S+N | $\mathbf{5.24 \pm .06}$ | **12.42** | **86.53** | **94.6** | **73.2** | **83.3** |
| Oxford | S+N | $4.26 \pm .05$ | 33.20 | 78.36 | 89.5 | 68.1 | 77.1 |
| | W+S+N | $\mathbf{4.28 \pm .09}$ | **28.63** | **79.63** | **90.9** | **77.2** | **81.7** |

Table 2. Quantitative comparison of proposed model which uses Words+Sentence+Noise(W+S+N) for initial condition with the model that uses Sentence+Noise(S+N).

To investigate the effect of generating images without using words and transformers in initial condition, we compare our model trained with Words, Sentence and Noise with popular approaches [19, 22] that generate low-resolution features using only sentence and noise. Table 2 compares the models quantitatively. Our model trained with Words+Sentence+Noise outperforms model using only Sentence+Noise for initial condition. This is because the transformers used in early layers of the network on words learn long-range dependencies and capture the global structure better than traditional up-sampling convolutions.

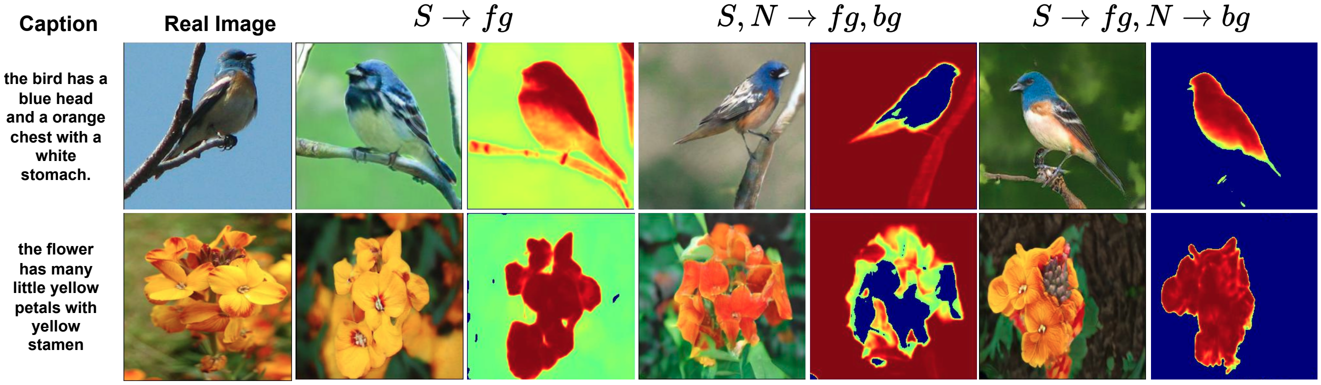
Caption
Real Image
$S \rightarrow fg$
$S, N \rightarrow fg, bg$
$S \rightarrow fg, N \rightarrow bg$
the bird has a blue head and a orange chest with a white stomach.
the flower has many little yellow petals with yellow stamen

Figure 2. Examples of images generated using different conditioning in spatial blocks and their effect on generation of segmentation masks on CUB and Oxford-102 datasets.

## 1.3. Decoupled Conditioning

Proposed Spatial Conditioning Blocks use separate foreground and background conditioning, allowing precise control over mask generation. To validate this claim, we have experimented comparing our model, which utilises sentence and noise conditioning for foreground and background (S → fg, N → bg), with two variants of conditioning: (i) using only Sentence for Foreground (S → fg) as in SSA-GAN and (ii) using Sentence concatenated with Noise for both Foreground and Background (S,N → fg,bg). We have evaluated the quality of generation of image and mask on the CUB and Oxford-102 datasets and presented the results in Table 3.

Our proposed model with decoupled conditioning has achieved the best performance, allowing for focused conditioning for different segments in the image. In contrast, the Sentence for Foreground model (S → fg) has failed to separate foreground and background in the generated segmentation masks. Models using concatenated Sentences and Noise for both Foreground and Background (S,N → fg,bg) have achieved similar performance as that of our model but loses control over generating segmentation masks. Furthermore, we show examples generated by each model and its segmentation masks in Figure 2. Our approach, with dedicated conditioning for foreground and background, prompts SCM predictors to produce segmentation masks with a probability close to 1 for the foreground and 0 for the background. This contrasts the approach using the same conditioning for foreground and background in SCM predictors, which only predicts segmentation to separate the features.

| Dataset | Conditioning | IS | FID | R% | ACC | IoU | mIoU |
|---|---|---|---|---|---|---|---|
| **CUB** | **S → fg** | $5.04 \pm .03$ | 14.83 | 82.63 | 85.1 | 53.3 | 68.1 |
| | **S,N → fg,bg** | $5.22 \pm .08$ | 13.98 | 81.33 | 90.9 | 77.1 | 89.1 |
| | **S → fg, N → bg** | $\mathbf{5.24 \pm .06}$ | **12.42** | **86.53** | **94.6** | **73.2** | **83.3** |
| **Oxford** | **S → fg** | $4.15 \pm .07$ | 30.24 | 77.47 | 87.5 | 64.2 | 74.9 |
| | **S,N → fg,bg** | $4.18 \pm .05$ | 30.19 | 78.31 | 90.9 | 71.8 | 78.7 |
| | **S → fg, N → bg** | $\mathbf{4.28 \pm .09}$ | **28.63** | **79.63** | **90.9** | **77.2** | **81.7** |

Table 3. Quantitative comparison of proposed model which uses Sentence (S) for foreground (fg) and Noise (N) for background (bg) in different settings for Spatial Conditioning Blocks.

### 1.4. Linear-SCM for Finer Predictions

In our proposed COS-GAN model, we incorporate the Linear-SCM, which employs the Spatial Co-Attention Mechanism with the Linformer [16] technique. This technique involves the application of a linear layer to spatial dimension of the reference features, resulting in extraction of finer segmentation masks with a minimal increase in overall computation. To evaluate the impact of Linear-SCM on our model, we train a variant of the model that excludes Linear-SCM. In this variant, segmentation masks required for Spatial Conditioning Blocks are generated using the last SCM prediction with bilinear upsampling of masks.

| Dataset | L-SCM | IS | FID | R% | ACC | IoU | mIoU |
|---|---|---|---|---|---|---|---|
| **CUB** | ✗ | $5.18 \pm .02$ | 13.28 | 81.25 | 92.2 | 65.4 | 76.9 |
| | ✓ | $\mathbf{5.24 \pm .06}$ | **12.42** | **86.53** | **94.6** | **73.2** | **83.3** |
| **Oxford** | ✗ | $4.13 \pm .03$ | 29.52 | 78.71 | 89.1 | 73.7 | 78.1 |
| | ✓ | $\mathbf{4.28 \pm .09}$ | **28.63** | **79.63** | **90.9** | **77.2** | **81.7** |

Table 4. Quantitative comparison of the proposed model which uses Linear-SCM (L-SCM) for segmentation mask predictions with the model that does not use L-SCM.

We present a quantitative comparison of the two models in Table 4, where we observe that the model using Linear-SCM achieves a significant improvement in performance by generating finer segmentation masks at higher resolutions. Figure 3 further supports this finding by showing that the segmentation masks generated by the Linear-SCM model are sharper and have captured minor details, compared to the model that only uses SCM for predicting segmentation masks, which are then used with bilinear upsampling in deeper layers. Applying a co-attention approach at high spatial resolutions enables the model to generate more meaningful and precise segmentation masks, resulting in higher-quality generated images.

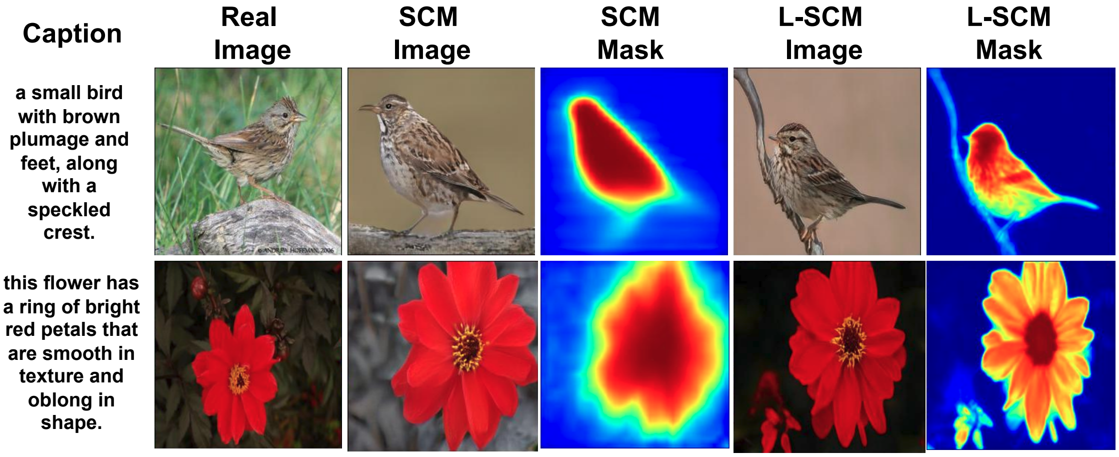
Caption
Real Image
SCM Image
SCM Mask
L-SCM Image
L-SCM Mask
a small bird with brown plumage and feet, along with a speckled crest.
this flower has a ring of bright red petals that are smooth in texture and oblong in shape.

Figure 3. Visual comparision of masks generated by our model using Linear-SCM and model with only SCM approaches conditioned on the caption for CUB and Oxford-102 datasets.

### 1.5. Fixed Reduction in Linear-SCM

We have evaluated the impact of different values of $k$ on performance of our model. Our model uses Linear-SCM to project the spatial dimensions of reference features to a fixed-size dimension ($k = 128$) to reduce the computation required for extracting the correlation matrix between the two generated features when the spatial size is greater than 32. We have also trained another model with $k = 64$ to further reduce the overall computation of our model.

| Dataset | k-dim | IS | FID | R% | ACC | IoU | mIoU |
|---|---|---|---|---|---|---|---|
| **CUB** | 64 | $5.09 \pm .07$ | 14.25 | 81.48 | 90.4 | 69.4 | 78.2 |
| | 128 | $\mathbf{5.24 \pm .06}$ | **12.42** | **86.53** | **94.6** | **73.2** | **83.3** |
| **Oxford** | 64 | $4.05 \pm .09$ | 30.57 | 78.19 | 89.8 | 70.6 | 77.1 |
| | 128 | $\mathbf{4.28 \pm .09}$ | **28.63** | **79.63** | **90.9** | **77.2** | **81.7** |

Table 5. Quantitative comparison of the proposed model, which uses fixed $k = 128$ reduction of spatial dimensions of reference features, with model using $k = 64$.

We have compared the performance of both the models in Table 5. We have found that when the $k$ value is small (i.e., $k = 64$), there is a noticeable drop in the quality of the generated images, suggesting that larger values of $k$ should be used to achieve better performance.

### 1.6. Generalisation of SCM and SCB blocks

To validate the efficacy of our proposed novel components, namely Spatial Co-Attention Mask (SCM) Predictor and Spatial Conditioning Blocks (SCB), we incorporate them into an existing text-to-image approach, namely SSA-GAN [7]. The results are presented in Table 6. In the SCM block, we use varied noise vectors with the same sentence vector to generate multiple image features. It can be observed from Table 6 that inclusion of SCM and SCB blocks into SSA-GAN (SSA-GAN+SCM, SSA-GAN+SCB) consistently improves performance, resulting in improved FID scores for text-to-image generation.

To evaluate the quality of the segmentation masks generated by our proposed model, we have conducted experiments on the CUB dataset using a weakly-supervised UNet [13] model. Specifically, we have used synthetic data generated by our SSA-GAN, COS-GAN, and ”SSA-GAN with SCM and SCB” models to train the UNet model; the standard test split of the CUB dataset is used for evaluation. Our results show that including SCM and SCB blocks in the SSA-GAN model has resulted in significantly improved foreground-background (FG-BG) masks of higher quality. This enhanced quality of masks benefits the performance of the SSA-GAN model and enhances the training of other models in weakly supervised learning.

| Method | IS | FID | R% | ACC | IoU | mIoU |
|---|---|---|---|---|---|---|
| SSA-GAN | $5.17 \pm .08$ | 15.61 | 85.4 | 61.6 | 20.4 | 39.4 |
| SSA-GAN+SCB | $5.16 \pm .06$ | 14.72 | 86.11 | 64.6 | 41.8 | 52.3 |
| SSA-GAN+SCM | $5.12 \pm .04$ | 14.93 | 83.43 | 75.7 | 52.3 | 61.3 |
| SSA-GAN+SCM+SCB | $5.19 \pm .09$ | 13.98 | 85.36 | 85.8 | 66.5 | 72.1 |
| **COS-GAN** | $\mathbf{5.24 \pm .06}$ | **12.42** | **86.53** | **94.6** | **73.2** | **83.3** |

Table 6. SSA-GAN with SCM and SCB blocks on CUB dataset for T2I image generation.

Our experiments show that the synthetic training data generated by COS-GAN outperforms SSA-GAN's, making it a viable option for training models in weakly-supervised settings for downstream tasks.

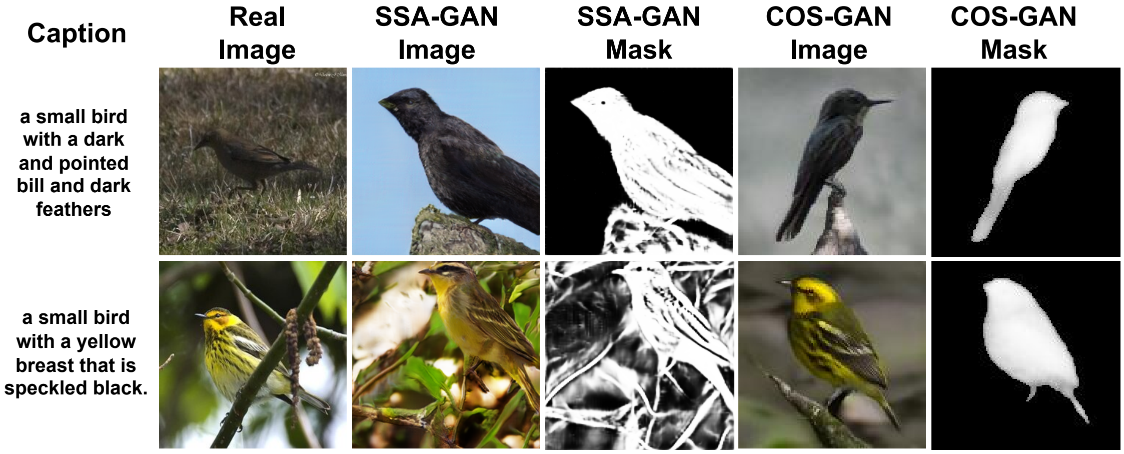
Caption
Real Image
SSA-GAN Image
SSA-GAN Mask
COS-GAN Image
COS-GAN Mask
a small bird with a dark and pointed bill and dark feathers
a small bird with a yellow breast that is speckled black.

Figure 4. Visual comparison of masks generated by our model and SSA-GAN [7] for CUB Dataset [17].

We have visually compared the images and masks generated by both SSA-GAN and COS-GAN in Figure 4. The mask generated by SSA-GAN has exhibited insufficient segmentation of the foreground and background. On the other hand, our approach leverages global reference images to predict masks, resulting in a superior ability to separate the foreground and background. This is due to introducing a global structure in our approach, in contrast to the local prediction strategy employed by SSA-GAN.

## 2. Details of the COS-GAN Architecture

In this section, we elaborate on the internal architecture details of COS-GAN. Our model is implemented using Pytorch [12] framework. The COS-GAN architecture uses a single Generator (Section 3.1) and Discriminator (Section 3.2) for generating images at resolution 256 × 256.

## 3. Implementation Details

We implement the models using PyTorch framework [12][1] and optimise the network using Adam optimiser [4] with the following hyperparameters: $\lambda_1 = 1$, $\lambda_2 = 1$, $\lambda_3 = 1$, $\lambda_4 = 1$ and $\lambda_5 = 1$, batch size = 24, learning rate = $0.0002$, $\beta_1 = 0.5$, and $\beta_2 = 0.999$ of Adam optimiser. Spectral Normalisation [9] is used for all fully connected and convolution layers in generator and discriminator. The model is trained for 600 epochs on CUB, Oxford-102 datasets (takes ~5 days in 3 NVIDIA 1080Ti GPUs) and 120 epochs for the COCO dataset (takes ~9 days in 3 NVIDIA 1080Ti GPUs). During inference, we report results with exponential moving average weights, with a decay rate of 0.999. We obtain global image features of sentence contrastive loss for R-precision from the discriminator network.

[1] The code will be released in GitHub upon acceptance.

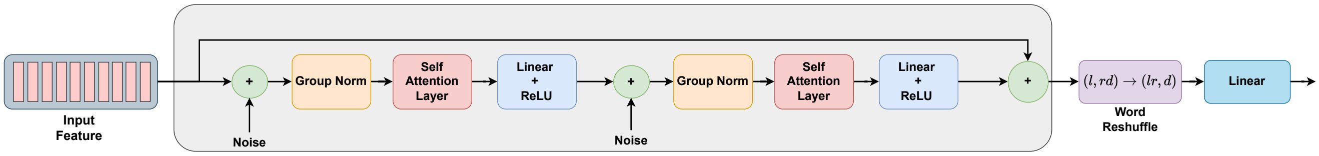

Figure 5. Overview of Self-Attention block with Word Reshuffle used in early layer of COS-GAN to generate low-resolution image features. We use Group Normalization [18]. To increase the stochastic capability of our model, we add noise as in StyleGAN [3, 11].

## 3.1. Generator

The proposed COS-GAN Generator generates two images simultaneously and extracts segmentation masks for the generated images. Overall, Generator architecture details are described in Table 7. For generating low-resolution features in our Generator, we use initial condition of Sentence+Noise+Words passing through a series of self-attention layers to generate low-resolution features. The use of self-attention layers allows our model to capture a better global structure [5]. The two low-resolution generated features are passed through a series of SCM predictors and spatial conditioning blocks for better refinement of the image features. The final generated features are passed through a linear layer for image generation.

Table 7. Generator architecture of COS-GAN. Self Attention Block 3.1.1 are used in early layers to generate low resolutions features capturing global structure. Residual connection is used along with the proposed SCM and Linear-SCM with spatial conditioning blocks 3.1.2.

| $z\epsilon\mathbb{R}^{256} \sim \mathcal{N}(0, I)$ , $S \ \epsilon \ \mathbb{R}^{512}$<br>$W \ \epsilon \ \mathbb{R}^{512}$, $l$ (length of the sequence) |
|:---:|
| Conditional Augmentation $\longrightarrow 256$ |
| Self Attention Block $\longrightarrow l \times 256$ |
| Self Attention Block $\longrightarrow 32 \times 256$ |
| Self Attention Block $\longrightarrow 64 \times 256$ |
| Self Attention Block $\longrightarrow 128 \times 256$ |
| Self Attention Block $\longrightarrow 256 \times 256$ |
| Reshape $\longrightarrow 256 \times 16 \times 16$ |
| Residual Block $\longrightarrow 256 \times 16 \times 16$ |
| Residual Conditioning $\longrightarrow 256 \times 16 \times 16$ |
| Residual Conditioning $\longrightarrow 256 \times 32 \times 32$ |
| Residual Conditioning $\longrightarrow 128 \times 64 \times 64$ |
| Residual Conditioning $\longrightarrow 128 \times 128 \times 128$ |
| Residual Conditioning $\longrightarrow 64 \times 256 \times 256$ |
| Convolution Block $\longrightarrow 64 \times 256 \times 256$ |
| $1 \times 1$ Convolution $\longrightarrow 3 \times 256 \times 256$ |

### 3.1.1 Self-Attention with Word Reshuffle

Given the initial condition (sentence concatenated with noise and words) to our network, we add positional encodings [15] to provide notion of the position of words and sentences to the network. To increase the stochasticity of our model, we add noise at each layer [3] similar to the proposed Styleformer [11]. As shown in Figure 5, these features are passed through a residual block consisting of Group Normalization [18], Self Attention Layer [15], and Linear layer with ReLU [1] activation. For increasing the sequence length, we reshape the features $(l, d \times r) \rightarrow (l \times r, d)$, where $l$ is the number of tokens, $d$ is the channel dimension, and $r$ is the factor for increasing the number of tokens. A linear layer follows each shuffle to increase the channel dimension. When the features are of size $256 \times d$, we reshape the features to $d \times 16 \times 16$ ($d$ is the channel dimension and 16 are spatial sizes of the feature) to generate initial low-resolution features.

### 3.1.2 Residual Spatial Conditioning Blocks

Given two generated low-resolution features, we pass the features through a residual block [2] and a series of residual conditioning blocks consisting of DM-Block [22] to incorporate word-based refinement, followed by $2\times$ Spatial Conditioning Blocks. We use Residual Spatial Conditioning Block without up-sampling (as shown in Figure 6) for low-resolution generated, followed by Residual Spatial Conditioning Block with up-sampling using Spatial Co-Attention mask (as shown in Figure 7) for up-sampling these low-resolution features. For feature size greater than 32, we use Linear-SCM to extract segmentation masks for usage in Spatial Conditioning Blocks (as shown in Figure 8). To reduce the overall computations, we use shared weights for both the features.

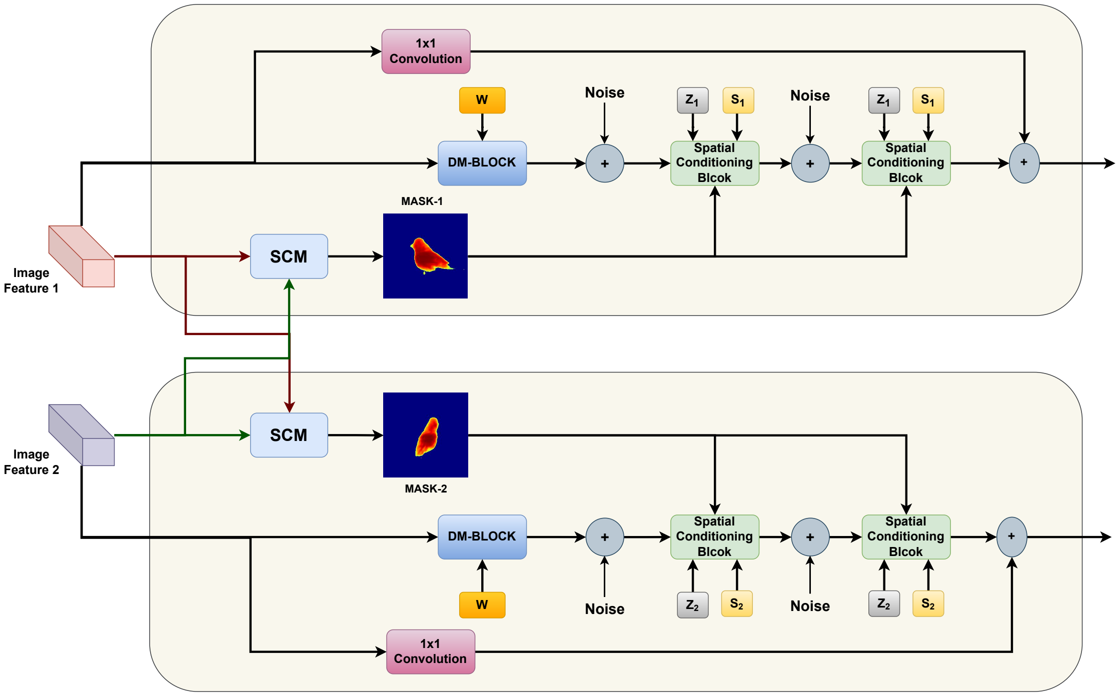

Figure 6. Overview of our Residual Spatial Conditioning Blocks using Spatial Co-attention Mask predictor without upsampling.

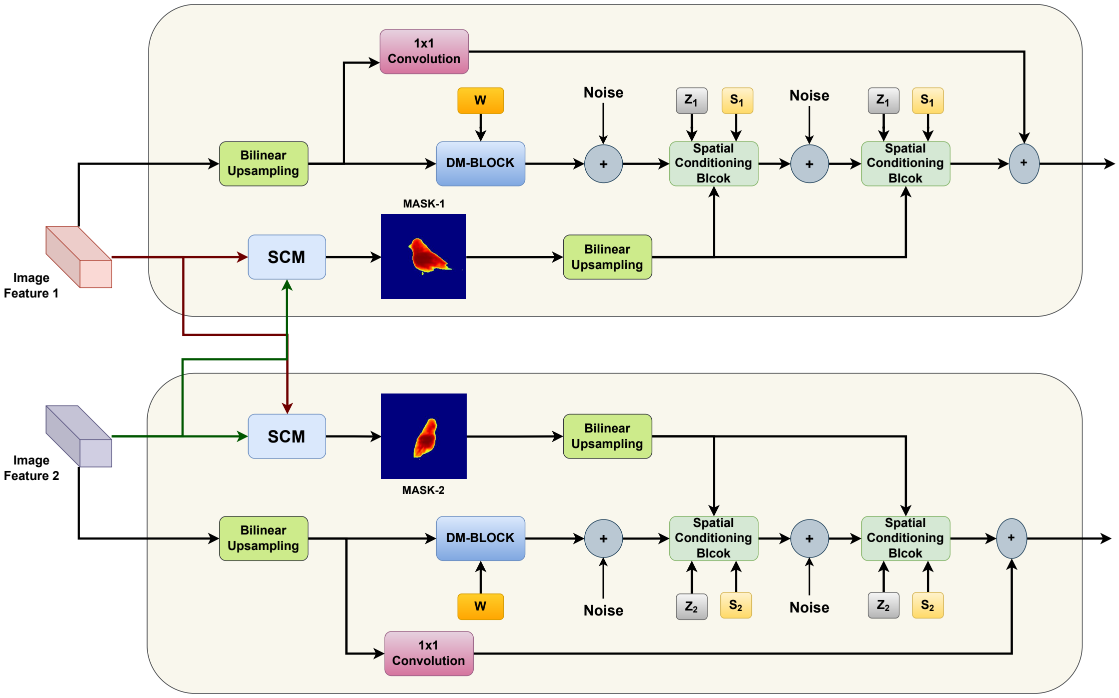

Figure 7. Overview of our Residual Spatial Conditioning Blocks using Spatial Co-attention Mask predictor with upsampling.

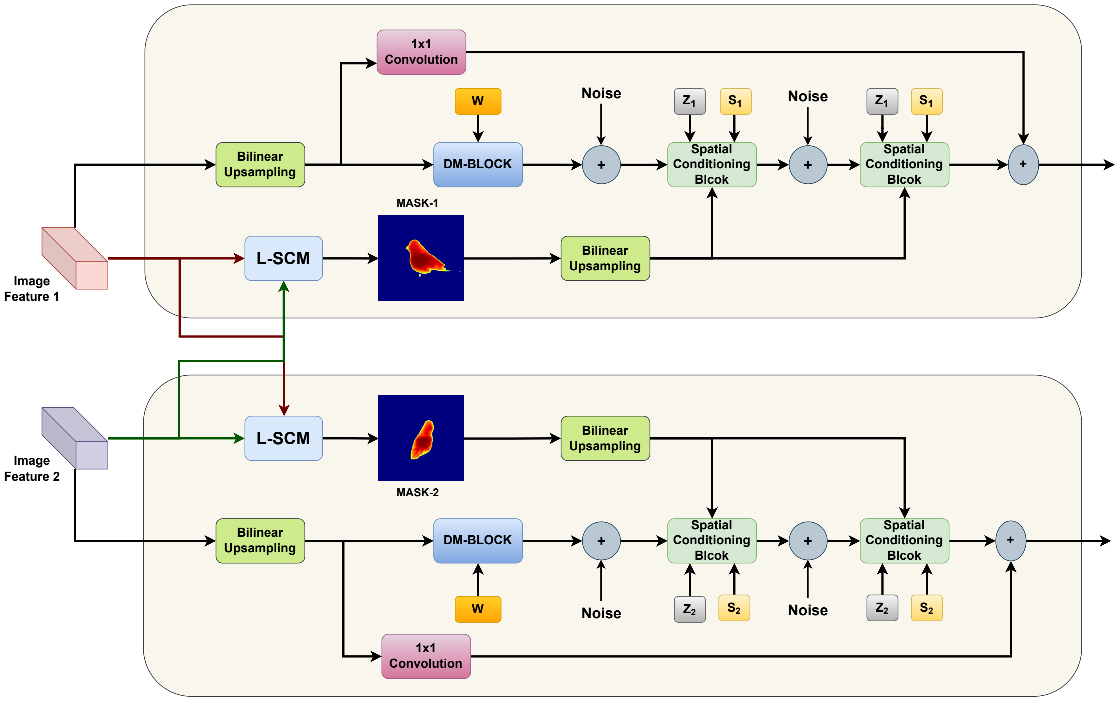

Figure 8. Overview of our Residual Spatial Conditioning Blocks using Linear Spatial Co-attention Mask predictor with upsampling.

Table 8. Discriminator architecture of COS-GAN.

| RGB images 3 × 256 × 256, $S \ \epsilon\ \mathbb{R}^{512}$, $W \ \epsilon\ \mathbb{R}^{512}$ | |
|---|---|
| DownBlock $\longrightarrow$ 64 × 128 × 128 | |
| DownBlock $\longrightarrow$ 128 × 64 × 64 | |
| DownBlock $\longrightarrow$ 256 × 32 × 32 | |
| DownBlock $\longrightarrow$ 256 × 16 × 16 | |
| DownBlock $\longrightarrow$ 512 × 8 × 8 | Word Contrastive Loss |
| DownBlock $\longrightarrow$ 512 × 4 × 4 | |
| Fully Connected(512 × 4 × 4) $\longrightarrow$ 512 | |
| Linear(512) $\longrightarrow$ 1 | Linear(512) $\longrightarrow$ 512 |
| Adversarial Loss | Sentence Contrastive Loss |

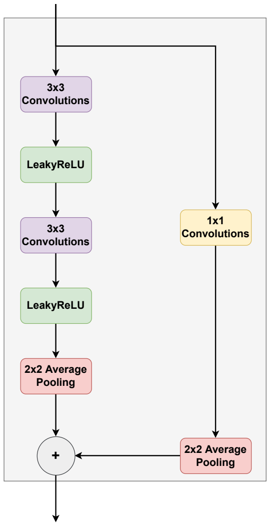

Figure 9. Residual Downsampling blocks used in discriminator.

### 3.2. Discriminator

Unlike multi-stage approaches having multiple discriminators [19,22], we use a single discriminator to predict if an image is real / fake and also extract features for text-alignment. The overall architecture of Discriminator is shown in Table 8. The Discriminator takes an image of 256 x 256 spatial resolution and passes through a series of residual downsampling blocks (DownBlocks - Figure 9). Our Discriminator has two final outputs, one for adversarial loss logit predictions and the other for global feature predictor for sentence contrastive loss. When feature's spatial size is 16 x 16, these features are provided for word contrastive loss.

## 4. More Qualitative Results

We provide additional qualitative results of our model on CUB [17] dataset in Figures 10, 11 and 12 compared with DF-GAN [14], on Oxford-102 [10] dataset in Figures 13, 14 and 15 compared with HDGAN [21], and on MS-COCO [8] dataset in Figures 16, 17 and 17 compared with AttnGAN [19]. We also provide visualisation for generated segmentation masks for all SCM and Linear-SCM in Figures 19, 20 and 21 for CUB, Oxford-102 and COCO datasets respectively.

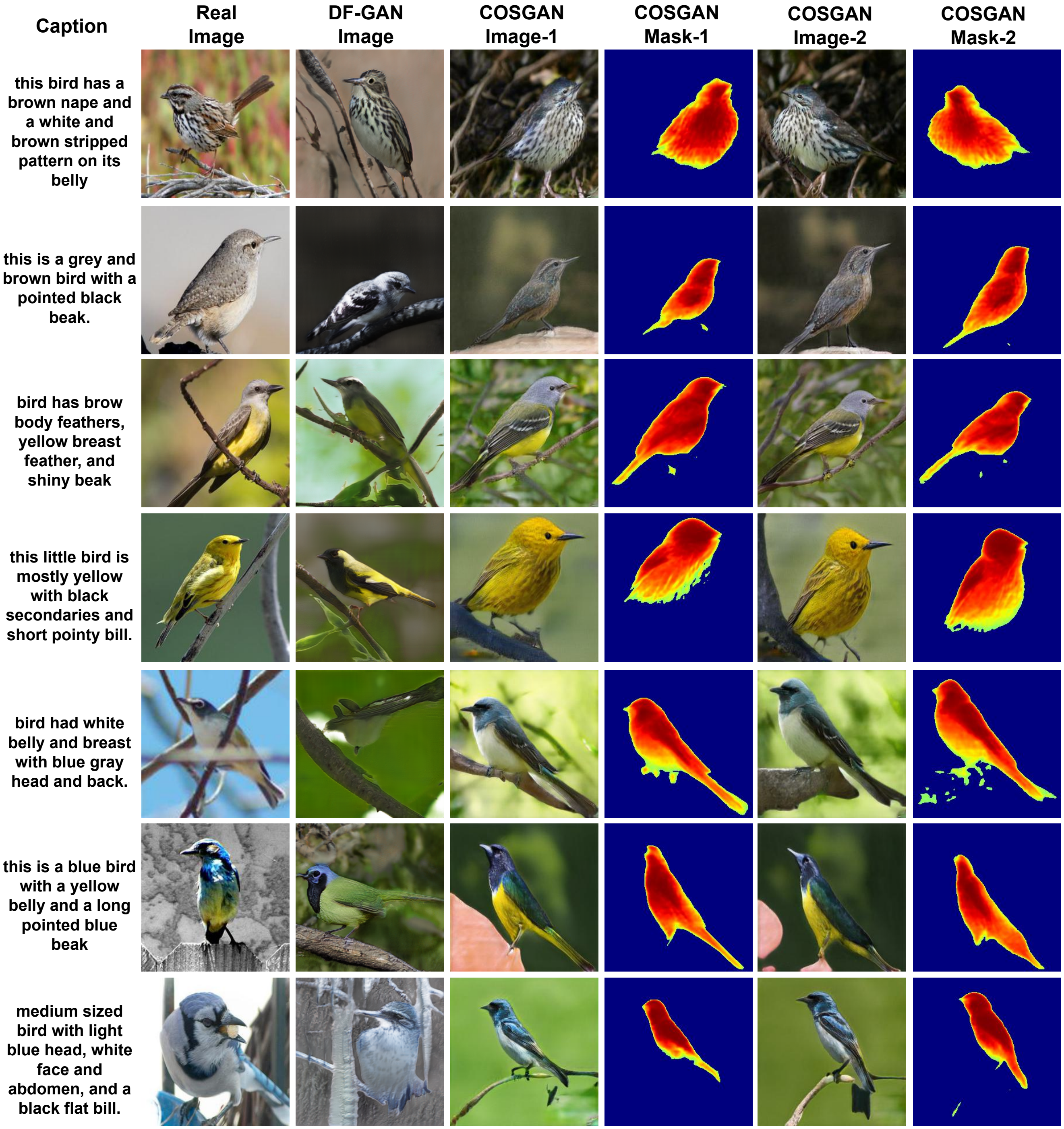
Caption
Real Image
DF-GAN Image
COSGAN Image-1
COSGAN Mask-1
COSGAN Image-2
COSGAN Mask-2
this bird has a brown nape and a white and brown stripped pattern on its belly
this is a grey and brown bird with a pointed black beak.
bird has brow body feathers, yellow breast feather, and shiny beak
this little bird is mostly yellow with black secondaries and short pointy bill.
bird had white belly and breast with blue gray head and back.
this is a blue bird with a yellow belly and a long pointed blue beak
medium sized bird with light blue head, white face and abdomen, and a black flat bill.

Figure 10. Two images generated simultaneously and their segmentation masks by COS-GAN (ours) on CUB dataset [17] and compared with those of DF-GAN [14].

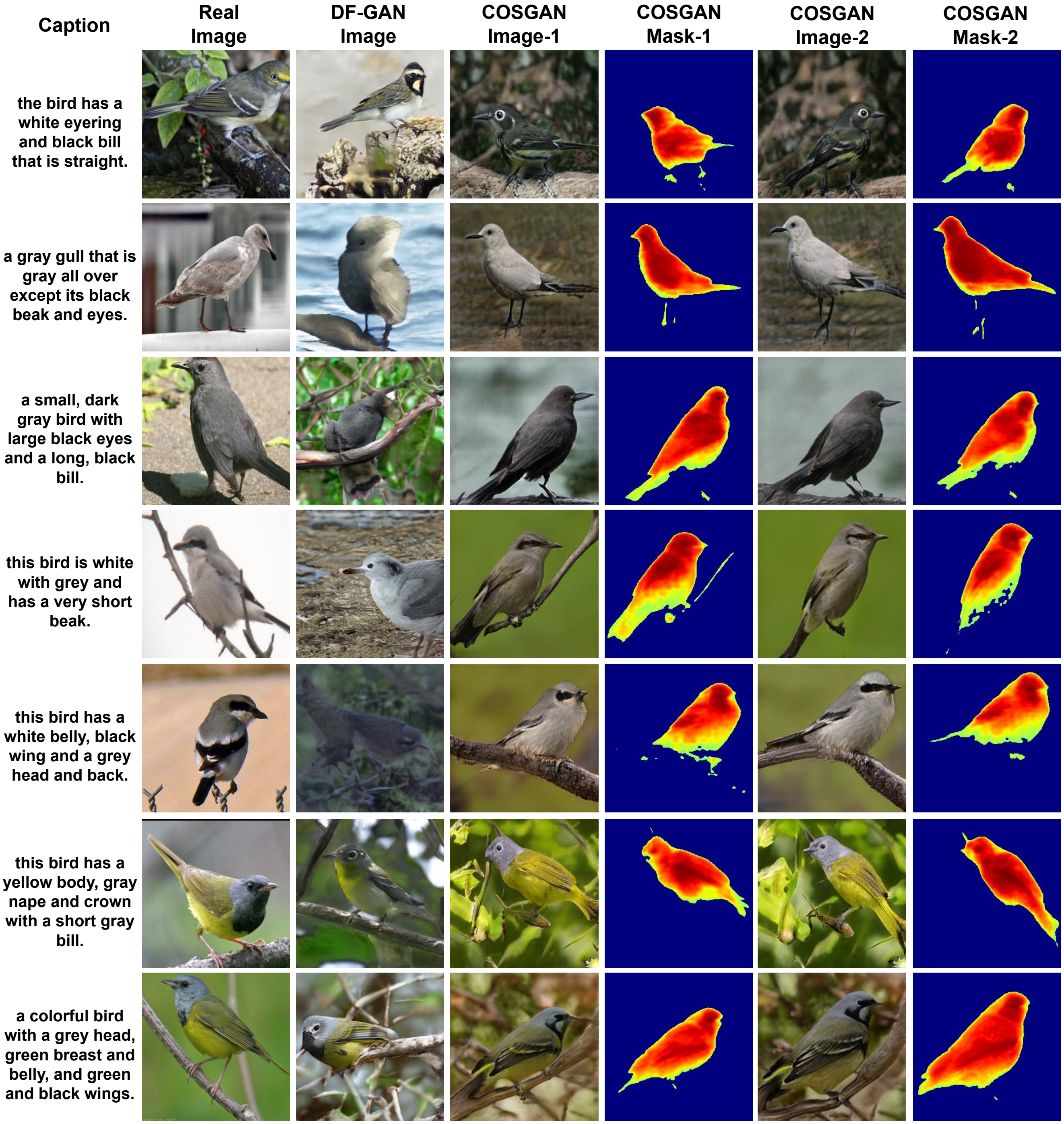
Caption
Real Image
DF-GAN Image
COSGAN Image-1
COSGAN Mask-1
COSGAN Image-2
COSGAN Mask-2
the bird has a white eyering and black bill that is straight.
a gray gull that is gray all over except its black beak and eyes.
a small, dark gray bird with large black eyes and a long, black bill.
this bird is white with grey and has a very short beak.
this bird has a white belly, black wing and a grey head and back.
this bird has a yellow body, gray nape and crown with a short gray bill.
a colorful bird with a grey head, green breast and belly, and green and black wings.

Figure 11. Two images generated simultaneously and their segmentation masks by COS-GAN (ours) on CUB dataset [17] and compared with those of DF-GAN [14].

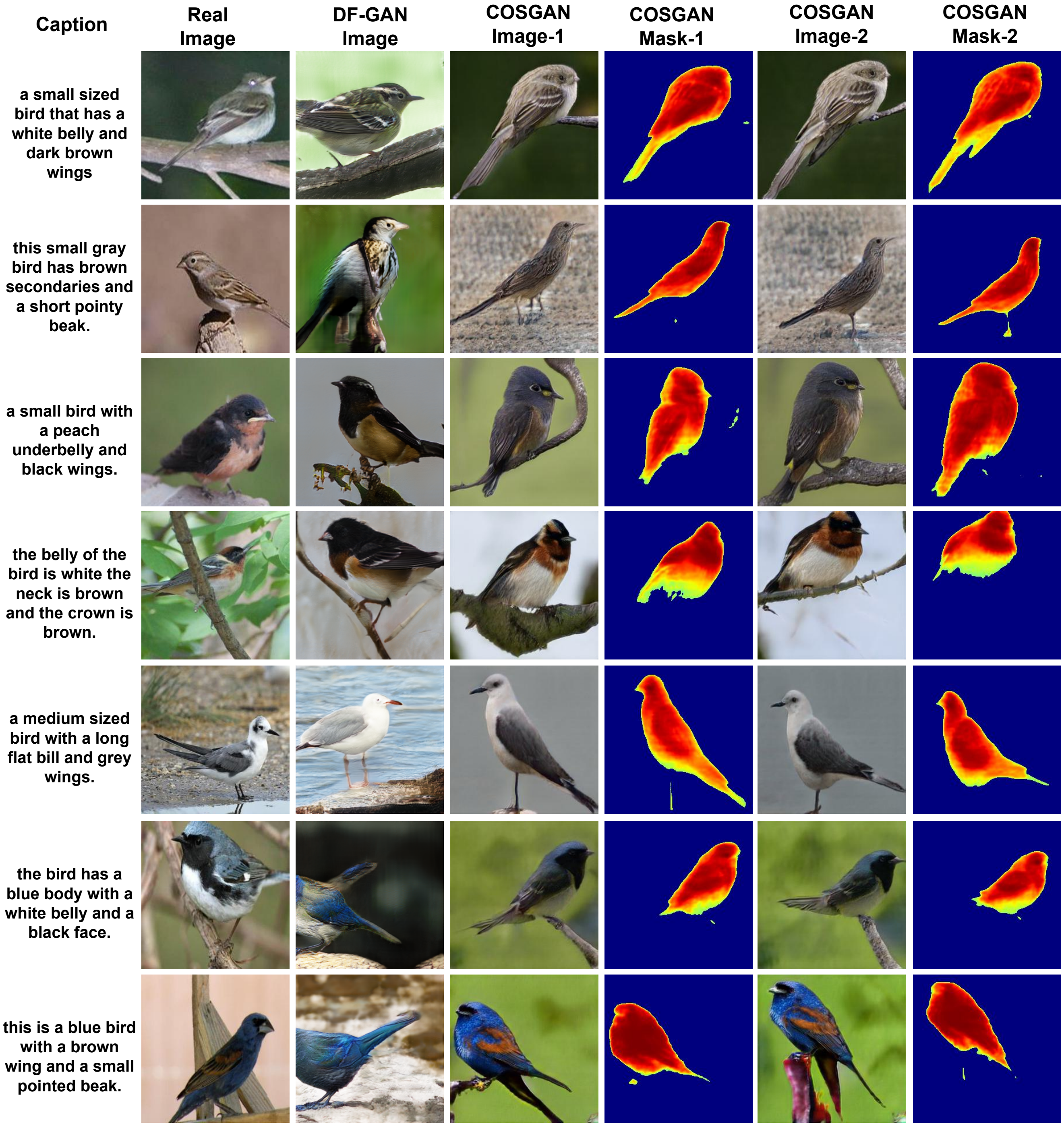
Caption
Real Image
DF-GAN Image
COSGAN Image-1
COSGAN Mask-1
COSGAN Image-2
COSGAN Mask-2
a small sized bird that has a white belly and dark brown wings
this small gray bird has brown secondaries and a short pointy beak.
a small bird with a peach underbelly and black wings.
the belly of the bird is white the neck is brown and the crown is brown.
a medium sized bird with a long flat bill and grey wings.
the bird has a blue body with a white belly and a black face.
this is a blue bird with a brown wing and a small pointed beak.

Figure 12. Two images generated simultaneously and their segmentation masks by COS-GAN (ours) on CUB dataset [17] and compared with those of DF-GAN [14].

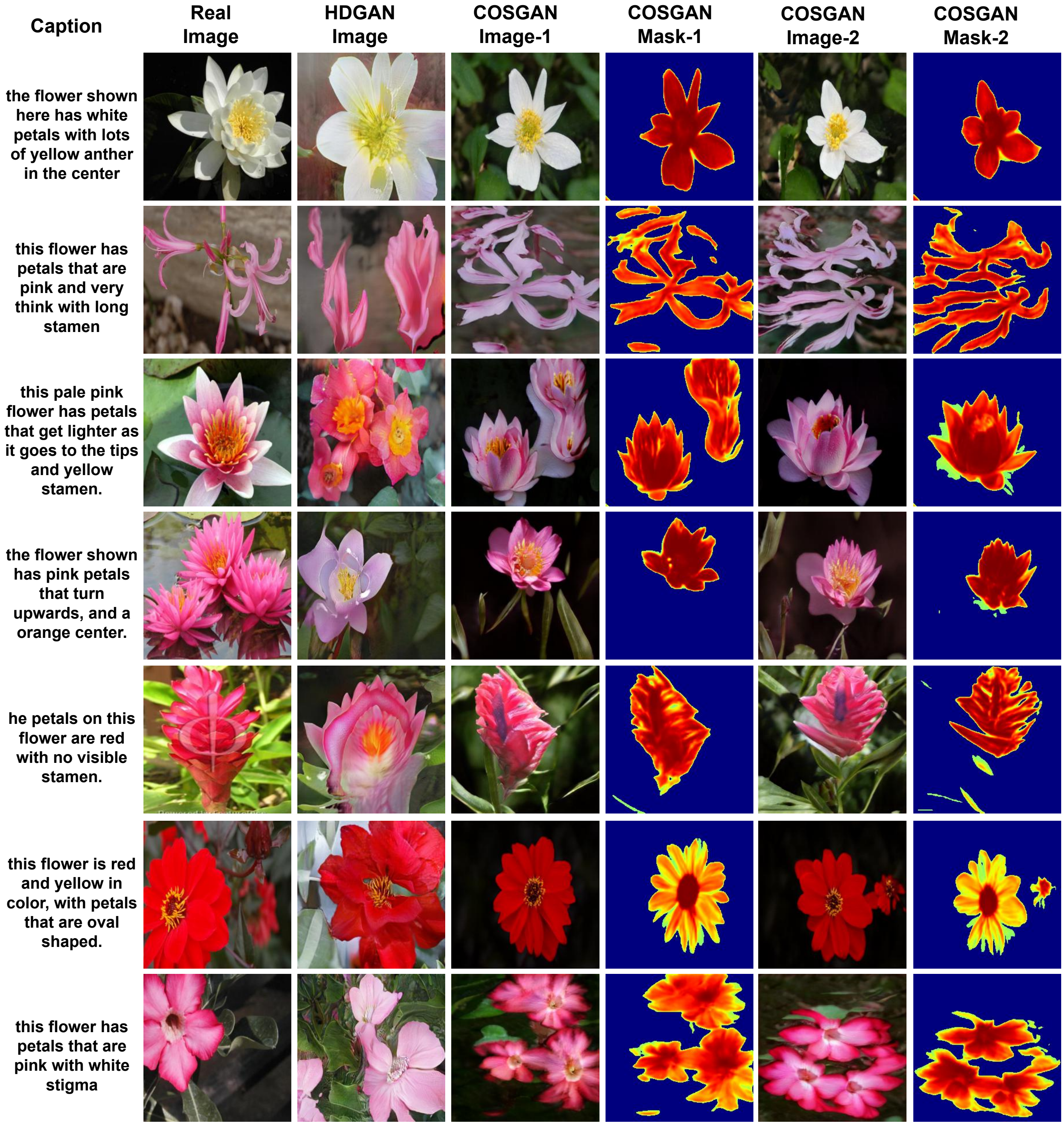
Caption
Real Image
HDGAN Image
COSGAN Image-1
COSGAN Mask-1
COSGAN Image-2
COSGAN Mask-2
the flower shown here has white petals with lots of yellow anther in the center
this flower has petals that are pink and very think with long stamen
this pale pink flower has petals that get lighter as it goes to the tips and yellow stamen.
the flower shown has pink petals that turn upwards, and a orange center.
he petals on this flower are red with no visible stamen.
this flower is red and yellow in color, with petals that are oval shaped.
this flower has petals that are pink with white stigma

Figure 13. Two images generated simultaneously and their segmentation masks by COS-GAN (ours) on Oxford-102 dataset [10] and compared with those of HDGAN [14].

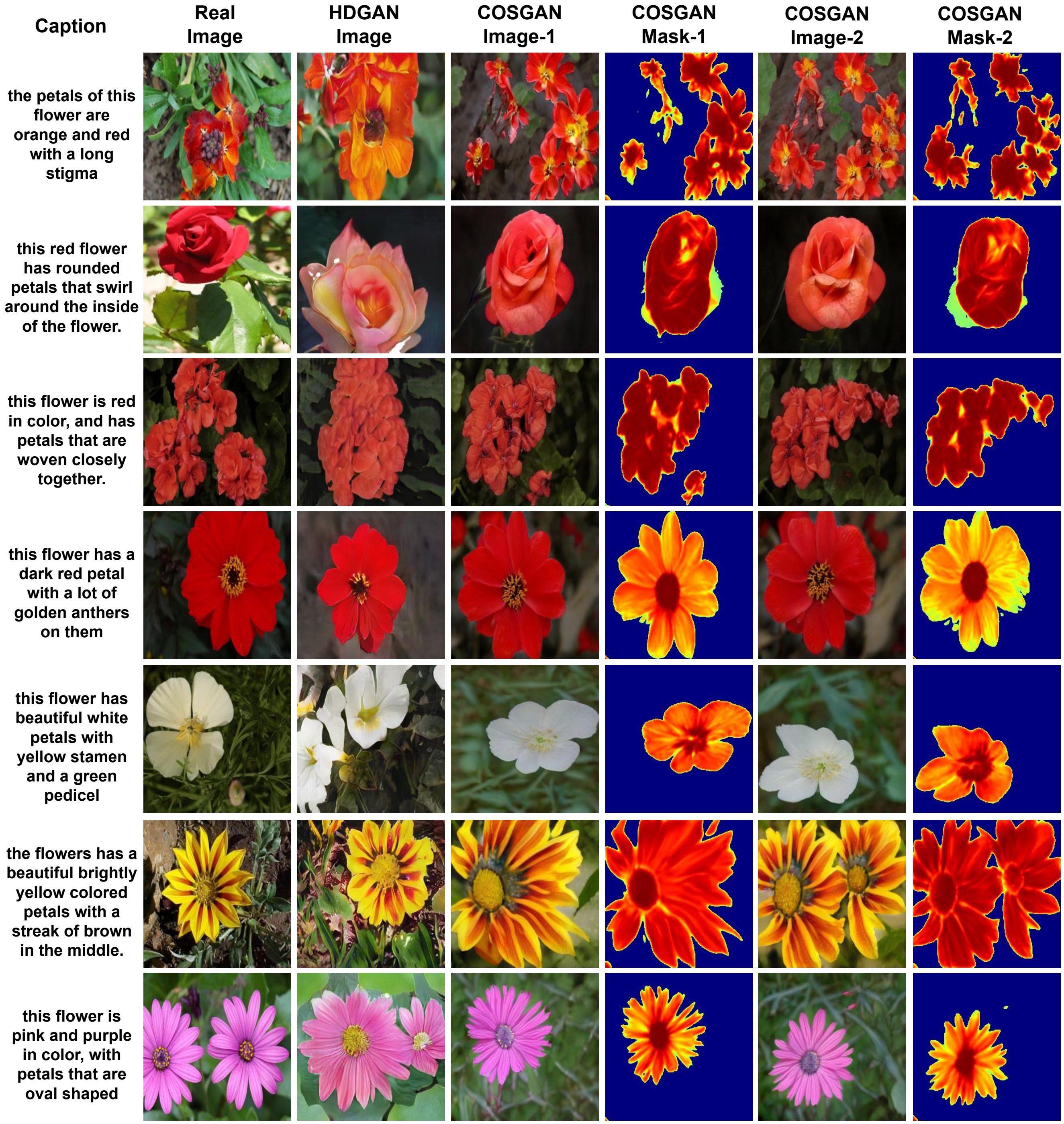
Caption
Real Image
HDGAN Image
COSGAN Image-1
COSGAN Mask-1
COSGAN Image-2
COSGAN Mask-2
the petals of this flower are orange and red with a long stigma
this red flower has rounded petals that swirl around the inside of the flower.
this flower is red in color, and has petals that are woven closely together.
this flower has a dark red petal with a lot of golden anthers on them
this flower has beautiful white petals with yellow stamen and a green pedicel
the flowers has a beautiful brightly yellow colored petals with a streak of brown in the middle.
this flower is pink and purple in color, with petals that are oval shaped

Figure 14. Two images generated simultaneously and their segmentation masks by COS-GAN (ours) on Oxford-102 dataset [10] and compared with those of HDGAN [14].

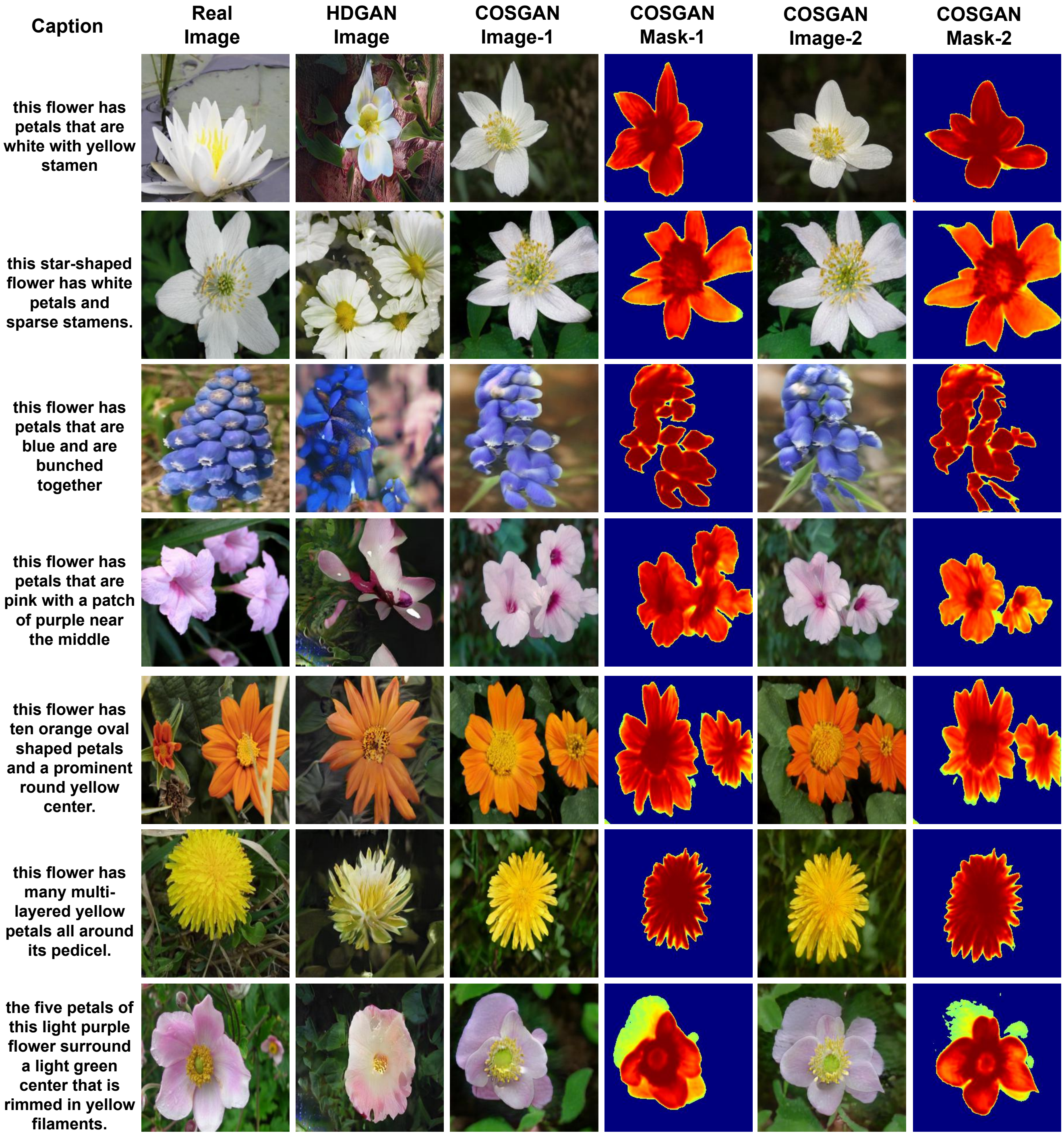
Caption
Real Image
HDGAN Image
COSGAN Image-1
COSGAN Mask-1
COSGAN Image-2
COSGAN Mask-2
this flower has petals that are white with yellow stamen
this star-shaped flower has white petals and sparse stamens.
this flower has petals that are blue and are bunched together
this flower has petals that are pink with a patch of purple near the middle
this flower has ten orange oval shaped petals and a prominent round yellow center.
this flower has many multi-layered yellow petals all around its pedicel.
the five petals of this light purple flower surround a light green center that is rimmed in yellow filaments.

Figure 15. Two images generated simultaneously and their segmentation masks by COS-GAN (ours) on Oxford-102 dataset [10] and compared with those of HDGAN [14].

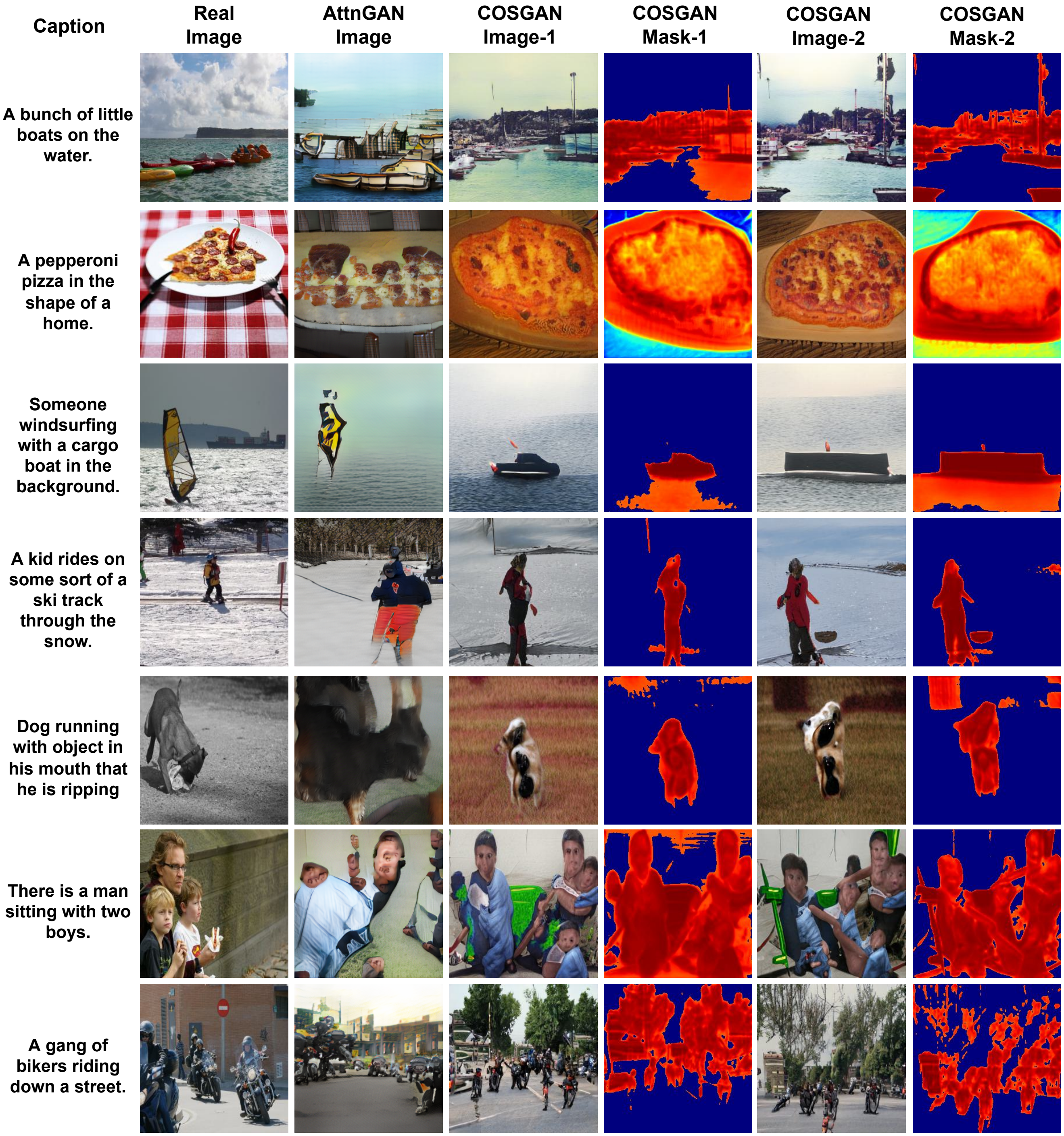
Caption
Real Image
AttnGAN Image
COSGAN Image-1
COSGAN Mask-1
COSGAN Image-2
COSGAN Mask-2
A bunch of little boats on the water.
A pepperoni pizza in the shape of a home.
Someone windsurfing with a cargo boat in the background.
A kid rides on some sort of a ski track through the snow.
Dog running with object in his mouth that he is ripping
There is a man sitting with two boys.
A gang of bikers riding down a street.

Figure 16. Two images generated simultaneously and their segmentation masks by COS-GAN (ours) on COCO dataset [8] and compared with those of AttnGAN [19].

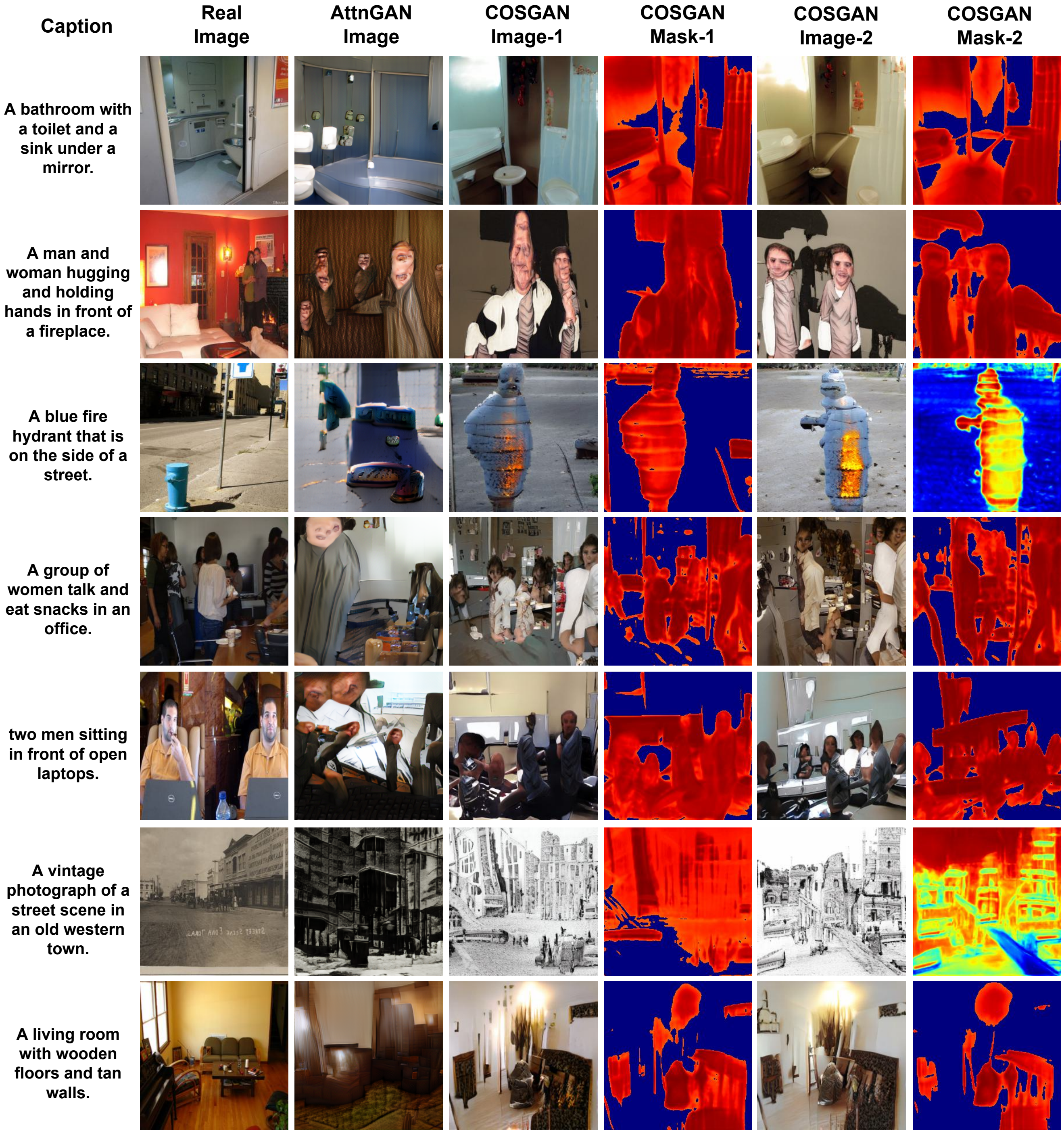
Caption
Real Image
AttnGAN Image
COSGAN Image-1
COSGAN Mask-1
COSGAN Image-2
COSGAN Mask-2
A bathroom with a toilet and a sink under a mirror.
A man and woman hugging and holding hands in front of a fireplace.
A blue fire hydrant that is on the side of a street.
A group of women talk and eat snacks in an office.
two men sitting in front of open laptops.
A vintage photograph of a street scene in an old western town.
A living room with wooden floors and tan walls.

Figure 17. Two images generated simultaneously and their segmentation masks by COS-GAN (ours) on COCO dataset [8] and compared with those of AttnGAN [19].

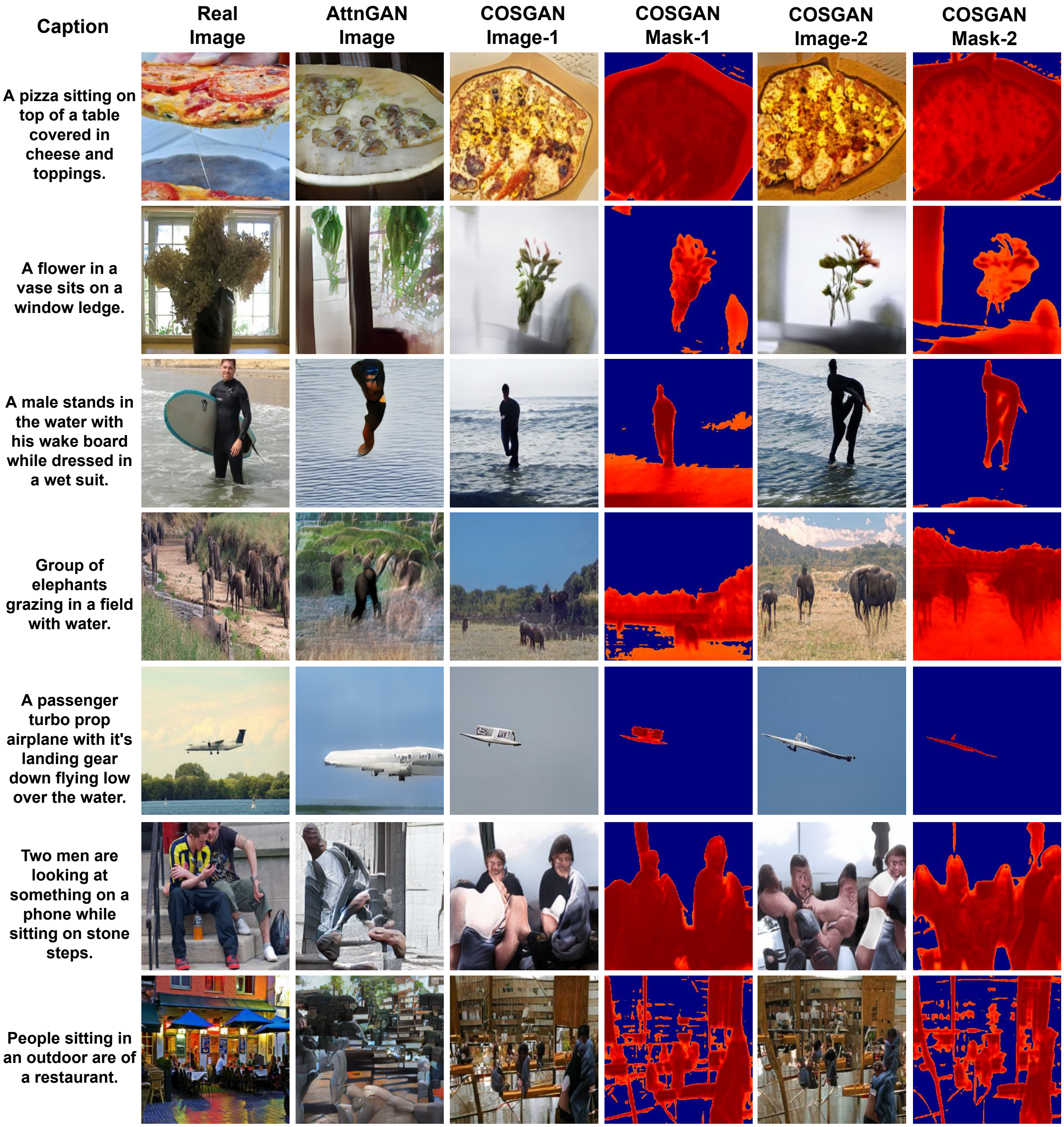
Caption
Real Image
AttnGAN Image
COSGAN Image-1
COSGAN Mask-1
COSGAN Image-2
COSGAN Mask-2
A pizza sitting on top of a table covered in cheese and toppings.
A flower in a vase sits on a window ledge.
A male stands in the water with his wake board while dressed in a wet suit.
Group of elephants grazing in a field with water.
A passenger turbo prop airplane with it's landing gear down flying low over the water.
Two men are looking at something on a phone while sitting on stone steps.
People sitting in an outdoor are of a restaurant.

Figure 18. Two images generated simultaneously and their segmentation masks by COS-GAN (ours) on COCO dataset [8] and compared with those of AttnGAN [19].

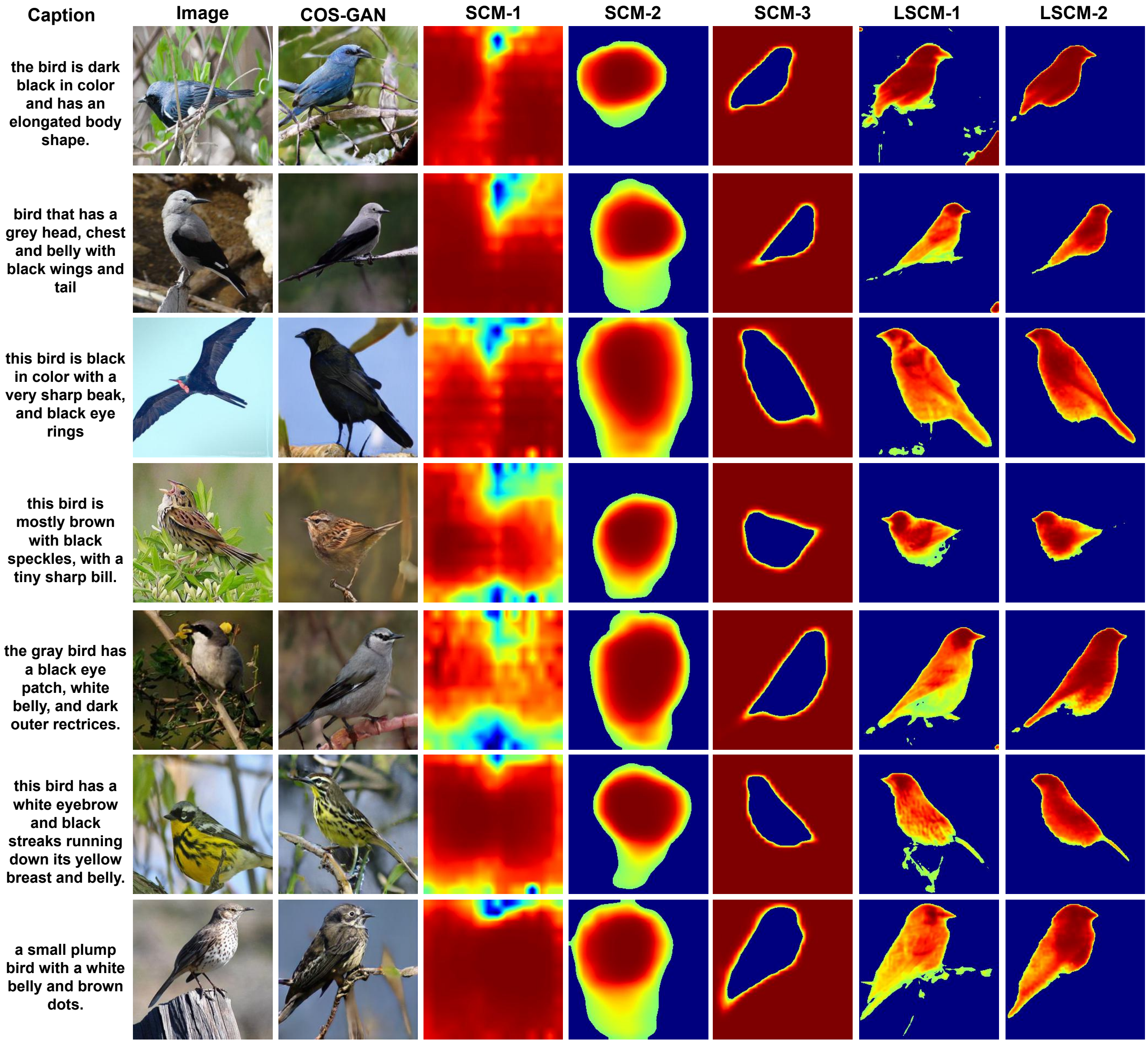
Caption
Image
COS-GAN
SCM-1
SCM-2
SCM-3
LSCM-1
LSCM-2
the bird is dark black in color and has an elongated body shape.
bird that has a grey head, chest and belly with black wings and tail
this bird is black in color with a very sharp beak, and black eye rings
this bird is mostly brown with black speckles, with a tiny sharp bill.
the gray bird has a black eye patch, white belly, and dark outer rectrices.
this bird has a white eyebrow and black streaks running down its yellow breast and belly.
a small plump bird with a white belly and brown dots.

Figure 19. Segmentation masks generated by all SCM and Linear-SCM blocks for the generated images on CUB dataset.

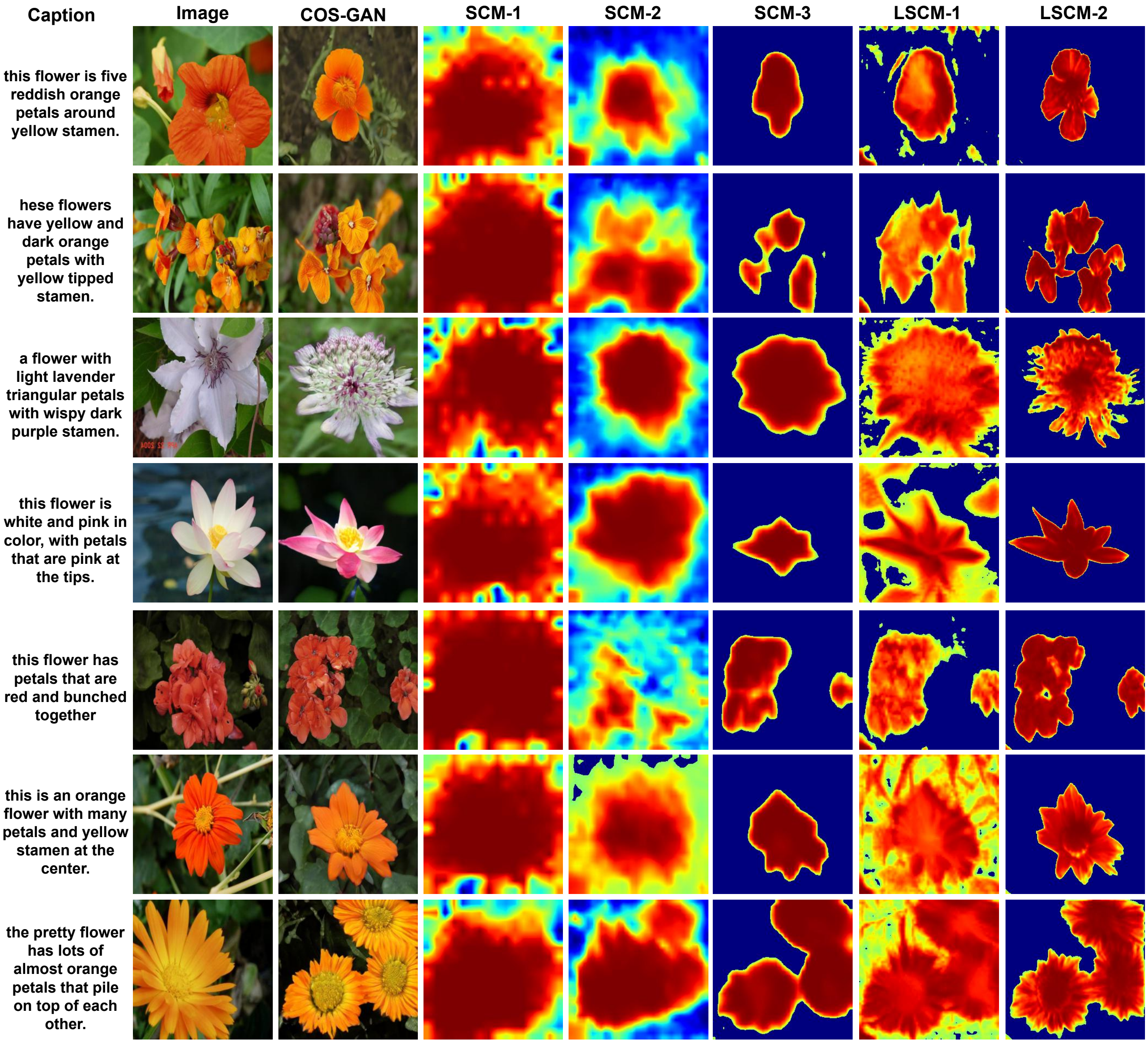
Caption
Image
COS-GAN
SCM-1
SCM-2
SCM-3
LSCM-1
LSCM-2
this flower is five reddish orange petals around yellow stamen.
hese flowers have yellow and dark orange petals with yellow tipped stamen.
a flower with light lavender triangular petals with wispy dark purple stamen.
this flower is white and pink in color, with petals that are pink at the tips.
this flower has petals that are red and bunched together
this is an orange flower with many petals and yellow stamen at the center.
the pretty flower has lots of almost orange petals that pile on top of each other.

Figure 20. Segmentation masks generated by all SCM and Linear-SCM blocks for the generated images on CUB dataset.

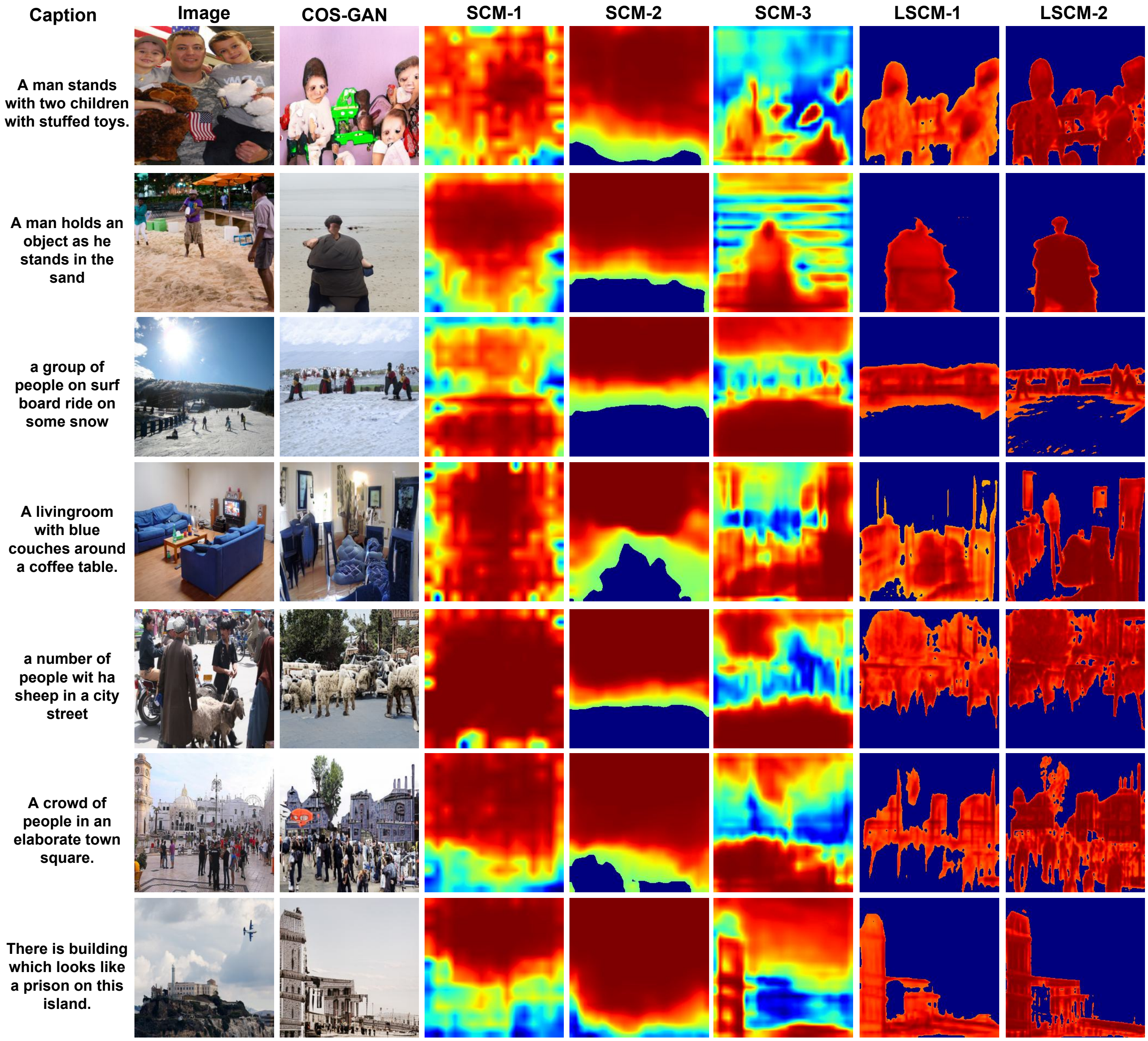
Caption
Image
COS-GAN
SCM-1
SCM-2
SCM-3
LSCM-1
LSCM-2
A man stands with two children with stuffed toys.
A man holds an object as he stands in the sand
a group of people on surf board ride on some snow
A livingroom with blue couches around a coffee table.
a number of people wit ha sheep in a city street
A crowd of people in an elaborate town square.
There is building which looks like a prison on this island.

Figure 21. Segmentation masks generated by all SCM and Linear-SCM blocks for the generated images on COCO dataset.

## References

[1] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In Geoffrey Gordon, David Dunson, and Miroslav Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 315–323, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR. 6

[2] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 7

[3] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 6

[4] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 5

[5] Alexander Kolesnikov, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Jakob Uszkoreit, Lucas Beyer, Matthias Minderer, Mostafa Dehghani, Neil Houlsby, Sylvain Gelly, Thomas Unterthiner, and Xiaohua Zhai. An image is worth 16x16 words: Transformers for image recognition at scale. 2021. 6

[6] Jiadong Liang, Wenjie Pei, and Feng Lu. Cpgan: Content-parsing generative adversarial networks for text-to-image synthesis. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 491–508, Cham, 2020. Springer International Publishing. 2

[7] Wentong Liao, Kai Hu, Michael Ying Yang, and Bodo Rosenhahn. Text to image generation with semantic-spatial aware gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18187–18196, June 2022. 1, 4, 5

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, 2014. 9, 16, 17, 18

[9] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018. 5

[10] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008. 9, 13, 14, 15

[11] Jeeseung Park and Younggeun Kim. Styleformer: Transformer based generative adversarial networks with style vector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8983–8992, June 2022. 6

[12] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 5

[13] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. 4

[14] Ming Tao, Hao Tang, Fei Wu, Xiao-Yuan Jing, Bing-Kun Bao, and Changsheng Xu. Df-gan: A simple and effective baseline for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16515–16525, June 2022. 9, 10, 11, 12, 13, 14, 15

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 6

[16] Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity, 2020. 3

[17] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010. 5, 9, 10, 11, 12

[18] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 6

[19] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2, 9, 16, 17, 18

[20] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 833–842, June 2021. 2

[21] Zizhao Zhang, Yuanpu Xie, and Lin Yang. Photographic text-to-image synthesis with a hierarchically-nested adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 9

[22] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2, 7, 9