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

Abstract

Text-to-Image (T2I) synthesis is a challenging task requiring modelling both textual and image domains and their relationship. The substantial improvement in image quality achieved by recent works has paved the way for numerous applications such as language-aided image editing, computer-aided design, text-based image retrieval, and training data augmentation. In this work, we ask a simple question: Along with realistic images, can we obtain any useful by-product (e.g., foreground / background or multi-class segmentation masks, detection labels) in an unsupervised way that will also benefit other computer vision tasks and applications?.

In an attempt to answer this question, we explore generating realistic images and their corresponding foreground / background segmentation masks from the given text. To achieve this, we experiment the concept of co-segmentation along with GAN. Specifically, a novel GAN architecture called Co-Segmentation Inspired GAN (COS-GAN) is proposed that generates two or more images simultaneously from different noise vectors and utilises a spatial co-attention mechanism between the image features to produce realistic segmentation masks for each of the generated images. The advantages of such an architecture are two-fold: 1) The generated segmentation masks can be used to focus on foreground and background exclusively to improve the quality of generated images, and 2) the segmentation masks can be used as a training target for other tasks, such as object localisation and segmentation. Extensive experiments conducted on CUB, Oxford-102, and COCO datasets show that COS-GAN is able to improve visual quality and generate reliable foreground / background masks for the generated images.

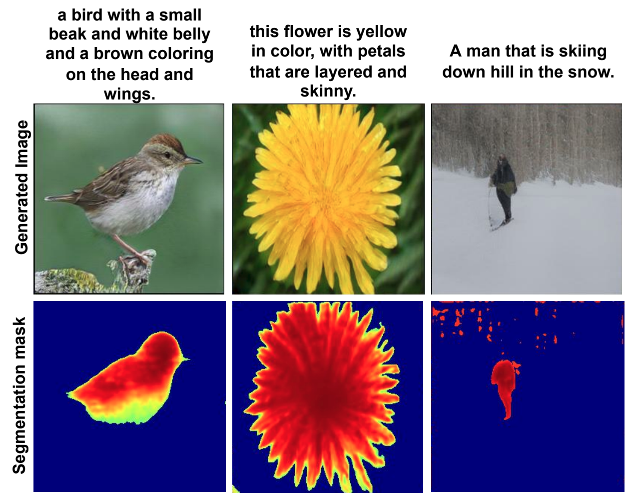

Figure 1. Examples of images and unsupervised segmentation masks generated by our method (COS-GAN) on CUB (left), Oxford-102 (middle), and COCO (right) datasets.

1. Introduction

The computer vision community has recently garnered extreme interest in the text-to-image (T2I) [16, 19, 21, 26, 53, 62, 80] synthesis task because of its wide range of applications such as language-aided image editing, computer-aided design, and text-based image retrieval. Further, T2I models can be used to produce numerous novel images (to be able to train machine learning models). In general, T2I task requires generative models to understand complex intra-modal and inter-modality relationships between both text and image domains to produce meaningful and realistic images.

T2I approaches typically utilise conditional GANs [43, 45, 48] by taking noise and sentences as conditional inputs to generate images. Earlier approaches [13, 56, 57] have successfully generated low-resolution images on single-class datasets [47, 78]. Generating high-resolution images is made possible by multi-stage GAN architectures [88] that generate images at different resolutions, with a high-resolution stage generator conditioned on images generated from low-resolution stage. Recent attempts have used various improved techniques such as intermediate discriminators [90], visually-aligned textual features [82, 91], multi-stage refinement with local attention [8, 32, 35, 36, 40, 84], layer-wise fusion of text features [37, 71, 86, 89], codebook based visual discrete representations [16, 17, 54, 75] and denoising diffusion models [15, 58] to improve quality of image generation.

The current trend of approaches solely focuses on improving the quality of the generated images. In this paper, we take a step back and ask a simple question: Along with realistic images, can we obtain any useful by-product (e.g., foreground / background or detection labels) in an unsupervised way that will also benefit other computer vision tasks? To answer this question, we explore a new direction in the T2I synthesis to generate images along with their foreground-background segmentation masks in an unsupervised way.

Foreground-Background (FG-BG) segmentation is a special case of segmentation where each image pixel is classified as foreground or background. Unsupervised approaches to generate images and their corresponding FG-BG masks rely on training Generative Adversarial Networks (GANs) in a layered approach [1,4,5,42,66,83]. This involves using multiple generators to produce foreground and background separately and then predicting the mask from the foreground. This predicted mask is then used to combine both the foreground and background to create final FG-BG mask output. Further, many approaches also utilise the underlying class distinction in the dataset to generate images and corresponding masks. We propose using a GAN to generate images and their corresponding FG-BG masks from the text as a novel approach. These masks and images can be used to train other vision applications.

We hypothesise that generating multiple images for the same text and performing a notion of co-segmentation between those images will provide with a foreground mask, specific to the common object present in the images as defined by text. To validate this hypothesis, inspired by deep co-segmentation related approaches [9, 34, 68, 69, 74], we propose a novel architecture named "Co-Segmentation Inspired GAN" (COS-GAN) for generating images and segmentation masks from the given text. COS-GAN generates image and FG-BG masks as an intermediate output (refer Figure 1) in a completely unsupervised way on text-image datasets. Specifically, we propose a GAN model that generates two image features for the same text but is conditioned with different noise vectors. Then, we perform a spatial co-attention between these image features to promote the notion of co-segmentation. Further, co-attention logits are used to predict a 2-class segmentation mask (foreground (FG)-background (BG)) to signify FG / BG regions in the images. Further, these masks are used to enhance FG / BG regions exclusively. We conduct comprehensive experiments on CUB [78], Oxford-102 [47], and COCO [38] datasets to validate the performance of COS-GAN in terms of quality of the generated image and segmentation masks. We summarise contribution of our paper as follows:

- We formulate a novel framework to generate images and extract segmentation masks for the generated images conditioned on the text.
- We propose a Spatial Co-attention Mask (SCM) predictor to extract segmentation masks for the generated images and novel spatial conditioning blocks that use segmentation masks from SCM predictor to improve quality of the images generated.
- We formulate generating two image features and extracting segmentation masks. Using Co-Attention Mechanism produces higher quality segmentation mask and improves image generation quality.

2. Related Work

In this section, we discuss briefly some of the relevant works in the literature relating to this paper.

2.1. Text-to-Image Synthesis

For the past few years, Generative Adversarial Networks (GANs) [20] approaches have been used for generating images. With larger GAN models [6, 27], and with regularisation methods [2, 7, 22, 44], GAN approaches can generate images on large datasets like ImageNet [14]. Conditional GANs [43, 45, 48] with sentence conditioning can generate images at low resolutions [13, 56, 57]. StackGAN [88] generates images at intermediate resolutions using a stage-wise approach and uses them as conditioning in high-resolution generators. HDGAN [90] trains a single generator and multiple discriminators for each resolution to provide intermediate signals to the generator. AttnGAN [82] has introduced cross-domain attention for local refinement using image aligned text embeddings. DM-GAN [91] uses memory refinement-based attention to capture text-image interactions. MirrorGAN [51] has proposed generating captions from discriminator to boost text *vs.* image semantic consistency. SD-GAN [84] applies contrastive loss between two generated images for two different captions of the same image to capture better text-image relations. ControlGAN [32] has introduced a fine-grained discriminator to improve discriminator's capability to understand complex relations between text and image. CPGAN [36] extracts salient features of the image for each word to provide image representation to the generator along with the word. XMC-GAN [86] increases mutual information between image and text using inter-modality and intra-modality contrastive losses between images and text. DF-GAN [71] has introduced a single generator to generate images with affine conditioning of text with Matching-Aware zero-centered Gradient Penalty (MA-GP) to improve text-image alignment. SSA-GAN [37] uses semantic masks to improve spatial conditioning of the images. These generated semantic maps can also be used as FG-BG masks.

With the introduction of Neural Discrete representation [75], images are represented as low-level tokens of the visual codebook [17]. Images are treated as a sequence of tokens; and to generate images, models have to generate the sequence in an autoregressive approach using transformer [76]. Images represented as low-level tokens allow models to scale up to large models to generate images at high-resolutions [16, 54]. The current focus of such models is to learn compact visual codebook to reduce the number of autoregressive predictions at the time of inference [31] and further boost image quality generations by capturing multiple representations [19, 55, 81].

Another approach for generating images from text is Denoising Diffusion Probabilistic Models (DDPM) [67]. DDPM generates images by reversing the forward markovian chain by removing noise at each step which allows to generate images at high-resolutions [15, 25, 46]. Guided diffusion using language models [52] allows models to capture non-natural interaction between text and images [53, 62]. Diffusion models with visual codebooks are also used to generate images at low-resolutions [18, 21]. The incorporation of self-attentions [76] and dynamic convolutions [12] in the recently proposed scaling up of GANs [26, 64] enables faster generation of images, while maintaining comparable quality of the Diffusion approach. Moreover, GANs offer enhanced control over the process of image generation.

In this space of Text-to-Image generation models, all the approaches focus on improving the quality of the generated images and further boost text-image compatibility. With abundant availability of text-image pairs, in our proposed work, we generate text-conditioned segmentation masks for generated image features and further use the masks to improve quality of the images generated.

2.2. Foreground-Background Mask Generations

Various GAN-based models have been utilised for generation of FG-BG masks. Typically, these models adopt a layered training approach, where the generator is trained to produce foreground and background components separately. Subsequently, these components are combined using a mask predicted from the foreground. Although several approaches employing Information Maximisation [3, 42, 65, 85] have been proposed for FG-BG mask extraction, the quality of generated maps is generally inferior compared to that of the methods utilising GAN-based approaches.

FineGAN [66] is one such model that generates the background, foreground, and mask in a hierarchical tree-type neural network architecture with bounding boxes as inputs. OneGAN [4] uses a complete unsupervised training approach to generate FG-BG masks, with reconstruction losses applied between pose, style, and shape vectors that are predicted from both the generator and discriminator in a layered approach. Labels4Free [1] employs a pre-trained StyleGAN model for generating masks using a layered approach. Melas-Kyriazi et al. [42] use the latent spaces of pre-trained large-scale GAN models to generate masks. Yang et al. [83] generate FG-BG segmentation mask using layered GANs and alternate training of GAN and segmentation networks for the generated mask.

Several GAN-based models can generate segmentation masks with human intervention [33] or off-the-shelf mask prediction techniques [73] for pre-trained large-scale image synthesis networks [6, 28]. In the case of DDPM-based models [46, 67], pre-trained mask-generated networks [23] are employed to predict the mask for the features extracted from the trained DDPM models [58]. Some T2I (Text-to-Image) models can generate semantic maps. For instance, TReCS [29] employs text and mouse localisation to generate both images and their corresponding segmentation maps. Another T2I model, SSA-GAN [37], utilises a segmentation approach to generate semantic maps alongside pictures based on a given text. In contrast to current methods, our proposed approach introduces a novel GAN-based model that leverages Co-Segmentation to extract foreground-background masks conditioned on text. This approach distinguishes itself from layered GAN approaches, relying on additional interventions such as pre-trained models or human input and using segmentation approaches like CBAM [79] to predict segmentation maps.

3. Methodology

Our goal is to generate realistic images along with their foreground-background masks from the given text. To achieve this, we propose a simple architecture involving co-segmentation [9, 34] between two image features simultaneously generated from the same text (with different noise vectors). Specifically, we propose a novel framework called *"Co-Segmentation Inspired GAN (COS-GAN)"* that accepts text T as input and encodes it into a sentence vector S. This sentence vector S is instantiated into two sentence vectors s_1, s_2 using conditional augmentation [88] and, further, augmented with two different noise vectors to yield two conditioning vectors v_1, v_2. These vectors are transformed into low-resolution spatial maps and then passed through multiple stages where every stage consists of a Spatial Co-attention Mask (SCM) predictor followed by upsample convolutions to finally output generated images. SCM predictor employs co-attention between two image features over whole spatial dimension to indice a notion of co-segmentation and capture global information for prediction of foreground-background (FG-BG) segmentation mask for each image. Further, predicting FG-BG masks on image attended feature maps results in superior quality masks over maps predicted simply on image features. Apart from the

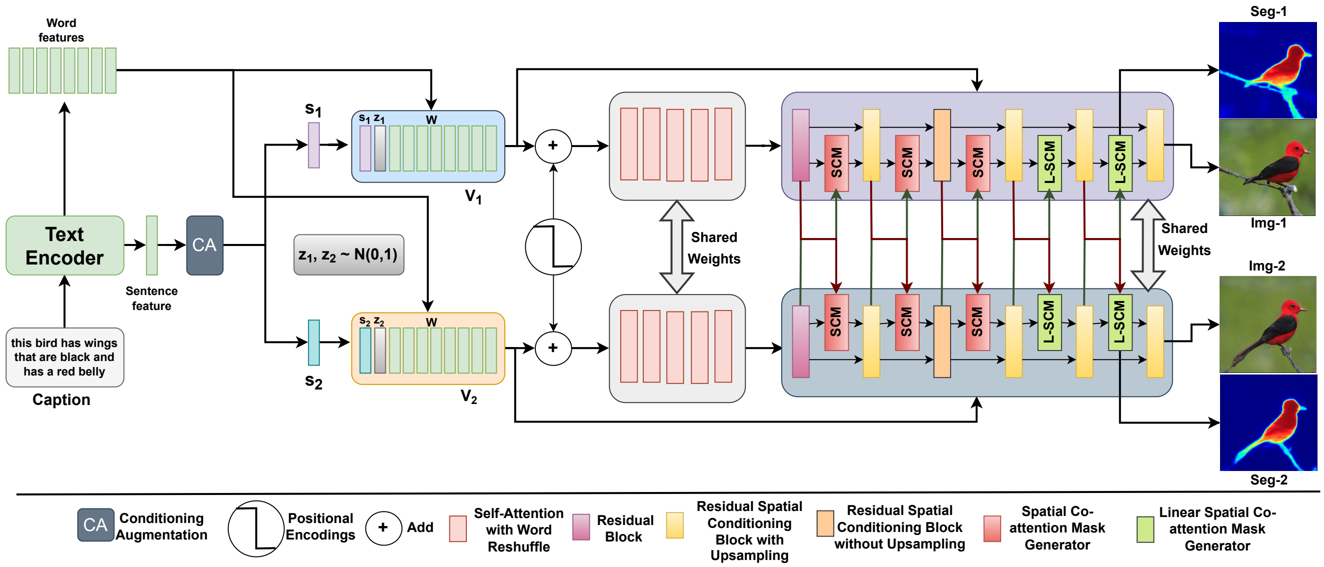

Figure 2. Overall architecture of COS-GAN generator 3.1. Given the caption text, the model creates two "conditioning" vectors by appending the text feature with two different noise vectors. Further, from these conditioning vectors, two different images are generated through multiple stages of upsampling convolutions and Spatial Co-attention Mask (SCM) predictors.

advantage of being a useful by-product, the prediction of FG-BG mask allows the model to individually act on foreground and background to enhance image quality. The network is trained using a combination of simple adversarial loss and text-image alignment loss. In the following sections, Sec. 3.1 explains the architecture of the generator (G) and Sec. 3.2 introduces the discriminator (D).

3.1. Generator Architecture

Generator **G** aims to generate two images simultaneously from the same text by ensuring that the images possess enough variations. Adopting co-segmentation concept between image features results in segmenting the predominant common object in those images. To achieve this, G accepts the given text T as input and passes it through a pre-trained text encoder [82] to yield sentence vector S and word vectors W. Then, to generate two different images for the same caption, two conditioning vectors are prepared as follows: First, S is instantiated into two sentence vectors s_1, s_2 using conditioning augmentation [88]. By essentially sampling from a conditional distribution $\mathcal{N}\left(\mu\left(s\right), \Sigma\left(s\right)\right)$, the conditioning augmentation process enables the generator to introduce stochastic nature and variability into the generation process. Further, s_1, s_2 are appended with two different noise vectors z_1, z_2 produced from Standard Gaussian Distribution $\mathcal{N}(0, 1)$ and word features W to result in two conditioning vectors v_1, v_2.

As shown in Figure 2, the conditioning vectors v_1, v_2 are added with positional encoding [76] and passed through a set of self-attention layers [76] to capture long-range dependencies for improving the global structure of the generated images and capture complex interactions between the sentence, noise and word features. In self-attention layers, we follow PixelShuffle [30] style of reshaping method to increase the number of tokens, *i.e.,* reshaping $(l, d \times r) \rightarrow (l \times r, d)$, where l is the number of tokens, d is the channel dimension, and r is the factor for increasing the number of tokens. Each of these shuffles is followed by a linear layer to increase the channel dimension. Finally, after self-attention layers, we end up with two features of size $256 \times d$.

We reshape the self-attended features to $d \times 16 \times 16$ and get initial low-resolution spatial feature maps. These low-resolution feature maps are passed through a series of upsampling blocks to result in the high-resolution image of dimension 256×256. Here, each upsampling block consists of a Spatial Co-attention Mask (SCM - Sec. 3.1.1) predictor followed by a Spatial conditioning block (Sec. 3.1.2) and upsampling convolutions. SCM predictor employs a co-attention mechanism between its input feature maps to calculate a correlation matrix and predict an FG-BG segmentation mask for each of the input feature. Further, this FG-BG mask is used in the Spatial conditioning block to modulate FG and BG regions of the generated image. To improve our model's stochastic ability, we add noise to each layer similar to StyleGAN [28, 49]. To reduce overall computations, we use shared weights for generating multiple images and use only one generated image for predicting values for generator losses. Adversarial loss $\mathcal{L}^G_{Adv}$ for the generator is:

$$\mathcal{L}^G_{Adv} = \mathbb{E}_{\hat{x} \sim p_G}[-D(\hat{x})] \tag{1}$$

Here $\hat{x}$ is the generated (I_{fake}) image. To generate images reflecting the captions, generator is also trained to minimise sentence contrastive loss $\mathcal{L}^G_{\text{sent}}$ between the global image features $\hat{f}_g$ for generated images predicted by discriminator and sentence features (S) as follows:

$$\mathcal{L}_{\text{sent}}^{G}\left(\hat{f}_{g_i}, S_i\right) = -\log \frac{\exp\left(Sim\left(\hat{f}_{g_i}, S_i\right)\right)}{\sum_{n=1}^{N} \exp\left(Sim\left(\hat{f}_{g_i}, S_n\right)\right)} \tag{2}$$

$$Sim(\hat{f}_g, S) = \cos\left(\hat{f}_g, S\right)/\tau \tag{3}$$

We use cosine similarity $\cos(u,v) = u^T v/\|u\|\|v\|$, between features to calculate similarity scores $Sim(.,.)$ for sentence embeddings S and global visual features $\hat{f}_g$. $\hat{f}_g$, $\hat{f}_p$ are global and patch features extracted from discriminators for the generated image. The patch features $\hat{f}_p$ for generated images extracted must be aligned with the words W in the corresponding sentence. We use previous strategies to learn connections between these words and regions in the image [82, 86], the cosine similarity is computed between all the image regions and words in the sentence and compute the attention values $\alpha_{i,j}$ for word features w_i in the sentence and patch features as $\hat{f}_{p_j}$ as:

$$\alpha_{i,j} = \frac{\exp\left(\rho_1 Sim\left(w_i, \hat{f}_{p_j}\right)\right)}{\sum_{k=1}^{R} \exp\left(\rho_1 Sim\left(w_i, \hat{f}_{p_k}\right)\right)} \tag{4}$$

Here R ($= 256$) is the total number of regions in the patch. $c_i = \sum_{j=i}^{R} \hat{f}_{p_j}\alpha_{i,j}$, is the aligned visual region feature for the i^{th} word in the sentence. The score S_{word} function between all the regions in the patch feature $\hat{f}_p$ and all words W can be defined as:

$$\mathcal{S}_{\text{word}}\left(\hat{f}_p, W\right) = \log\left(\sum_{l=1}^{T} \exp\left(\rho_2 Sim(w_l, c_l)\right)\right)^{\frac{1}{\rho_2}} \tag{5}$$

Here T is the number of words in the sentence. ρ_1 and ρ_2 are hyper parameters; we set it to the same values as in AttnGAN [82]. Word Contrastive loss $\mathcal{L}_{word}^{G}$ for generator is as follows:

$$\mathcal{L}_{\text{word}}^{G}\left(\hat{f}_{p_i}, W_i\right) = -\log \frac{\exp\left(S_{word}\left(\hat{f}_{p_i}, W_i\right)\right)}{\sum_{n=1}^{N} \exp\left(S_{word}\left(\hat{f}_{p_i}, W_n\right)\right)} \tag{6}$$

We employ conditioning augmentation [88] to generate multiple conditioning vectors for the same sentence in the generator; we apply the regularisation term for conditioning augmentation (L_{CA}) on sentence feature vector(s) as:

$$\mathcal{L}_{CA} = D_{KL}\left(\mathcal{N}\left(\mu\left(s\right), \Sigma\left(s\right)\right) \| \mathcal{N}(0, I)\right) \tag{7}$$

Here $\mu(s)$ and $\Sigma(s)$ are mean and diagonal covariance matrices that are computed as functions of the sentence feature vectors. We use KL Divergence between the Standard Gaussian and the conditional Gaussian Distribution for regularisation. λ_1, λ_2 and λ_3 are hyper-parameters. The complete loss for the generator is defined as follows:

$$\mathcal{L}_G = \mathcal{L}_{Adv}^{G} + \lambda_1 \mathcal{L}_{CA} + \lambda_2 \mathcal{L}_{\text{sent}}^{G} + \lambda_3 \mathcal{L}_{\text{word}}^{G} \tag{8}$$

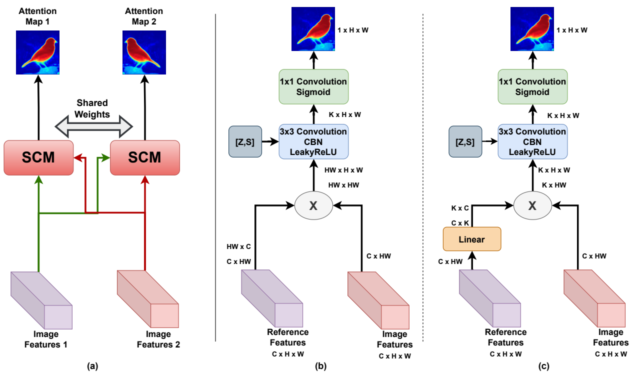

Figure 3. (a) Bidirectional feature sharing for generating attention maps, (b) Spatial Co-attention Mask to predict segmentation masks, given image and reference features, and (c) Linear-SCM to reduce computations for co-attention when the spatial size > 32 (K is set to 128).

3.1.1 Spatial Co-attention Mask Predictor

The prediction of intermediate masks [79] within generative models [1, 4, 37, 66, 83] typically relies on a single convolutional layer. However, this approach heavily depends on the local receptive field in image features for mask prediction, limiting its utilisation of global information within the image. To overcome this limitation, our proposed approach introduces generation of two images and implements spatial co-attention between their respective image features. This strategy aims to facilitate co-segmentation, enabling us to predict masks on image-attended feature maps that primarily highlight the common object present in both the images. We effectively integrate global information from both images by leveraging attention across all spatial locations between the images.

In the SCM block as shown in Figure 3, we combine an image feature with its reference feature across all spatial locations. This forms a correlation matrix that captures global information. Afterward, we process this matrix through a convolution block, which includes a convolutional layer followed by Conditional Batch Normalization [11] and a LeakyReLU activation [41], followed by a linear layer with a Sigmoid activation to predict an FG-BG mask. When dealing with larger spatial resolutions (≥ 64), co-attention across all spatial locations can result in significant memory usage. To address this concern, we employ a linear layer with a LinFormer [77] approach (using K = 128) to achieve co-attention. We refer to this modified version of SCM as the Linear-SCM (L-SCM) predictor.

3.1.2 Spatial Conditioning Block

In contemporary methods, predicted intermediate masks primarily enhance underlying tasks by regulating the impact of image features [37, 79]. However, these masks often focus only on the foreground or the object of interest, limiting their scope in generative models. This limitation stems from the importance of both foreground and background in generating high-quality images. In contrast, our approach utilises predicted masks for foreground and background, leading to enhanced performance in T2I tasks and better mask generation. Supplementary material provides additional evidence to support and validate this assertion.

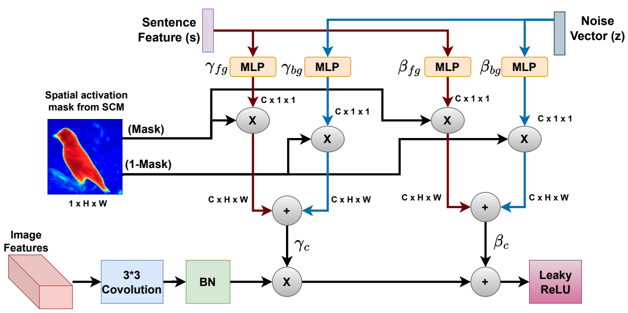

Figure 4. We use segmentation masks from SCM predictor in the Spatial Conditioning Block. We use sentence vector for foreground and noise vector for background conditioning.

We propose applying spatial conditioning to features using the FG-BG segmentation mask from the SCM predictor. In Figure 4, we utilise Conditional Batch Normalization [11] with modulation parameters (γ_c and β_c) for spatial conditioning on input features. For foreground conditioning related to text concepts, we use text features to estimate γ_{fg} and β_{fg}. As for the background, we use a noise vector z ($z \sim N(0,1)$) to estimate γ_{bg} and β_{bg}. The modulation parameters of Conditional Batch Normalization are derived from the mask (M) in the following manner:

$$\mathrm{BN}(x \mid s, z) = (\gamma_c) \cdot \frac{x - \mu(x)}{\sigma(x)} + (\beta_c) \tag{9}$$

$$\gamma_{fg} = FC_{\gamma_{fg}}(s) \tag{10}$$

$$\beta_{fg} = FC_{\beta_{fg}}(s) \tag{11}$$

$$\gamma_{bg} = FC_{\gamma_{bg}}(z) \tag{12}$$

$$\beta_{bg} = FC_{\beta_{bg}}(z) \tag{13}$$

$$\gamma_c = M \times \gamma_{fg} + (1 - M) \times \gamma_{bg} \tag{14}$$

$$\beta_c = M \times \beta_{fg} + (1 - M) \times \beta_{bg} \tag{15}$$

FC is a fully connected layer + Leaky ReLU here, and for the foreground, we use the $mask$ and $(1 - mask)$ for the background. Using segmentation masks from the SCM predictor for spatial conditioning prompts the SCM predictor to generate better segmentation masks for using suitable conditioning for the foreground, and the background, as the network is trained to enhance the quality of the generated image and be consistent with the text. The SCM predictor's ability to produce meaningful and high-quality segmentation masks, which are then used for dedicated modulations in the spatial conditioning block, contributes to enhancing image quality in the generated outputs.

3.2. Discriminator

The Discriminator **D** is used for two purposes: (1) to predict whether the image is real or fake, and (2) to be a feature encoder for extracting features of the image for multi-modal contrastive loss. The image is passed through a series of residual downsampling blocks to extract three features, as shown in Figure 5. Logit values and global sentence feature are extracted from the final fully connected layer. Patch features for word contrastive loss are extracted when the feature size is 16×16. The logit values are used for Adversarial Hinge loss [44]. Adversarial loss $\mathcal{L}_{Adv}^{D}$ for discriminator is as follows:

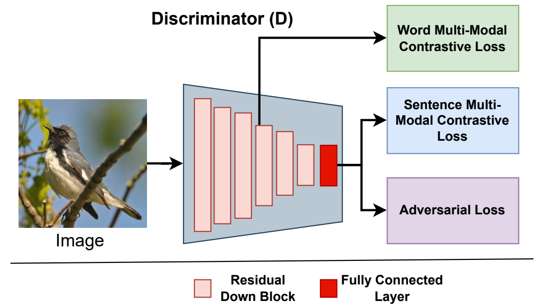

Figure 5. Overview of COS-GAN discriminator architecture. Discriminator consists of three outputs: i) Logits for adversarial loss, ii) Global image features for sentence contrastive loss, and iii) Patch features for word contrastive loss.

$$\begin{aligned} \mathcal{L}_{Adv}^{D} &= \mathbb{E}_{x \sim p_{\text{data}}}[\max(0, 1 - D(x))] \\ &+ \mathbb{E}_{\hat{x} \sim p_G}[\max(0, 1 + D(\hat{x}))] \end{aligned} \tag{16}$$

Here, x and $\hat{x}$ are real I_{real} and fake I_{fake} images. Global feature extracted from the final layer with linear projections is also used for sentence contrastive loss. Sentence Contrastive loss $\mathcal{L}_{sent}^{D}$ and Word Contrastive loss $\mathcal{L}_{word}^{D}$ for discriminator are as follows:

$$\mathcal{L}_{\text{sent}}^{D}(f_{g_i}, S_i) = -\log \frac{\exp(Sim(f_{g_i}, S_i))}{\sum_{n=1}^{N} \exp(Sim(f_{g_i}, S_n))} \tag{17}$$

$$\mathcal{L}_{\text{word}}^{D}(f_{p_i}, W_i) = -\log \frac{\exp(S_{word}(f_{p_i}, W_i))}{\sum_{n=1}^{N} \exp(S_{word}(f_{p_i}, W_n))} \tag{18}$$

f_g and f_p are global and patch features extracted from Discriminator for real images. For training of $\mathcal{L}_{\text{sent}}^{D}$ and $\mathcal{L}_{\text{word}}^{D}$, we use only real image pairs and not the generated image pairs as the images generated in early stages are not recognisable [86]. λ_4 and λ_5 are hyper-parameters. The final objective function for the Discriminator is defined as:

$$\mathcal{L}_D = \mathcal{L}_{GAN}^{D} + \lambda_4 \mathcal{L}_{\text{sent}}^{D} + \lambda_5 \mathcal{L}_{\text{word}}^{D} \tag{19}$$

4. Experiments

In this section, we introduce datasets and evaluation metrics used in our experiments. We then evaluate the proposed model on the datasets and compare qualitatively and quantitatively with the current approaches in the literature. The supplementary material further explains the specific details of the network, its training specifications, hyperparameters, and additional studies.

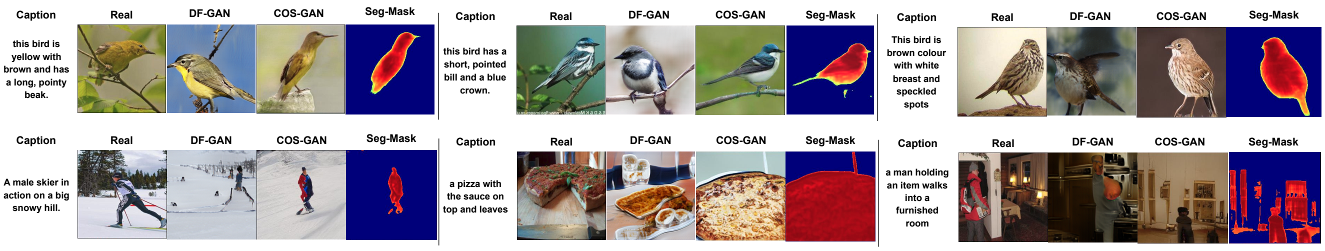

Figure 6. Visual comparison of images generated by DF-GAN [71] and our proposed COS-GAN with the FG-BG masks on CUB Dataset [78] (Top) and COCO Dataset [38] (Bottom).

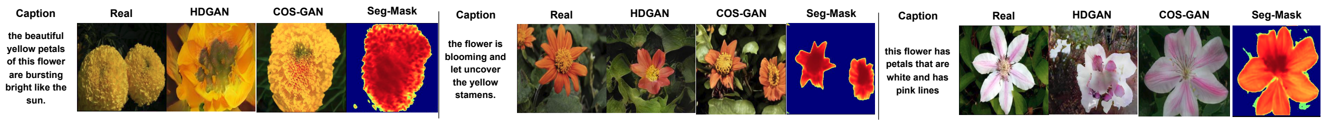

Figure 7. Images generated by HDGAN [90] and compared with images generated by COS-GAN and FG-BG masks on Oxford-102 [47].

Method	**CUB**			**COCO**		
	IS	**FID**	**R%**	**FID**	**R%**	**NoP**
StackGAN [88]	$3.70 \pm .04$	-	-	-	-	-
AttnGAN [82]	$4.36 \pm .02$	23.98	67.82	35.49	83.82	230M
MirrorGAN [51]	$4.56 \pm .06$	18.34	-	34.71	84.71	-
DM-GAN [91]	$4.75 \pm .07$	16.09	72.32	32.64	88.56	46M
CP-GAN [36]	-	-	-	35.41	-	318M
XMC-GAN [86]	-	-	-	9.30	-	166M
DAE-GAN [61]	$4.42 \pm .04$	15.19	85.45	26.72	92.61	98M
TIME [40]	$4.86 \pm .04$	14.81	71.57	31.14	-	120M
DF-GAN [72]	$4.86 \pm .04$	14.81	71.57	19.32	-	19M
SSA-GAN [37]	$5.17 \pm .08$	15.61	85.4	19.37	90.6	26M
COS-GAN	$5.24 \pm .06$	12.42	86.53	19.54	91.42	9M

Table 1. Quantitative comparison between COS-GAN and other models on CUB [78] and COCO [38] datasets. ”-” indicates values are unreported.

Datasets: We evaluate our model on three datasets, namely, 1) Caltech-UCSD birds (CUB) [78], 2) Oxford-102 flowers [47], and 3) MS COCO [39] datasets. The CUB and Oxford-102 datasets have single-class with ten captions provided for each image. For CUB and Oxford-102 datasets, we adopt a training and validation partition similar to StackGAN [88]. The MS-COCO dataset is a multi-class dataset with around 80k training and 40k validation images; and for every image, five captions are provided in the dataset.

Evaluation metrics: We use mainly three metrics to measure quality of the images generated: 1) *Inception Score (IS)* [63], 2) *Fréchet Inception Distance (FID)* [24], and 3) *R-precision (R%)* [82]. FID and IS are used to measure quality of the generated images. R% is used for measuring text-to-image consistency. FID calculates the Fréchet distance between two multivariate Gaussians fitted over the global features extracted from the Inception-v3 [70] on real and synthetic images. Lower FID means generated images are closer to real images. IS calculates the Kullback-Leibler (KL) divergence between a conditional distribution and marginal distribution for class probabilities from Inception-v3 [70] model. The higher IS suggests high quality images with more diverse classes. R-precision measures whether generated images can be used to retrieve the text (to determine the text-to-image alignment).

We also employ three other metrics to evaluate the quality of the generated FG-BG semantic maps: Mean Intersection over Union (mIoU), Intersection over Union (IoU), and pixel classification accuracy. The mIoU calculates the average intersection over union for both foreground and background. The IoU metric determines the intersection over union value for foreground alone. Lastly, the accuracy metric measures the percentage of correctly classified pixels.

4.1. Qualitative Visualisation

In Figure 6, we compare our results visually for images generated on CUB and COCO datasets with DF-GAN [71]. We also show the extracted segmentation masks from the last level SCM predictor (Linear-SCM) for the generated images. We observe that the images generated by our COS-GAN model reflect the text better than those of DF-GAN due to dedicated spatial conditioning for foreground and

Method	IS	FID
StackGAN [88]	$3.20 \pm .01$	51.89
StackGAN++ [87]	$3.26 \pm .01$	48.68
HDGAN [90]	$3.45 \pm .07$	43.17
LeicaGAN [50]	$3.92 \pm .02$	-
DualAttn-GAN [8]	$4.06 \pm .05$	40.31
COS-GAN	$\mathbf{4.28 \pm .09}$	**28.63**

Table 2. Quantitative comparison between COS-GAN and other models on Oxford-102 Dataset [47].

background. In Figure 7, images for the Oxford-102 dataset are generated and compared with those of HD-GAN [90]. Our generated images capture better semantics and appear more realistic. Images generated by our model are plausible and aligned with the text. As shown in Figures 6 and 7, generating two images and applying a co-attention mechanism allow us to generate high-quality segmentation masks. For visualisation of segmentation masks, we set a threshold of 0.5 and consider the values above 0.5 as foreground.

Method	ACC	IoU	mIoU
Supervised U-Net	98.0	88.8	93.2
GrabCUT [60]	72.6	36.0	52.3
FineGAN [66]	-	44.5	-
OneGAN [4]	-	55.5	-
PerturbGAN [5]	-	-	38.0
DRC [85]	-	56.4	-
Chen et al. [10]	84.5	42.6	-
IEM + SegNet [65]	89.3	55.1	71.4
Melas-Kyriazi et al. [42]	92.1	66.1	-
Yang et al. [83]	94.3	69.7	81.7
SSA-GAN [37]	61.6	20.4	39.4
COS-GAN	**94.6**	**73.2**	**83.3**

Table 3. Quantitative comparison of FG-BG semantic maps between our approach and that of other models on CUB dataset [78].

4.2. Quantitative evaluation

In Table 1, we compare the proposed COS-GAN with current GAN-based state-of-the-art models for text-to-image synthesis on CUB [78] and COCO [38] datasets. Our model improves the FID from 14.81 to 12.42 and IS from $5.17 \pm .08$ to $5.24 \pm .06$ on CUB dataset. Our model does not use any extra network to improve Text-Image alignment but only uses the discriminator to capture this alignment; so we notice a small drop in R-Precision values. On COCO dataset, in Table 1, we achieve similar performance as that of SSA-GAN method [37]. COS-GAN's ability to extract meaningful segmentation masks for the generated images can be seen as an added advantage over other models. For the COCO dataset, we report only FID and R-precision as IS scores do not reflect the quality of the generated images for larger datasets [37, 71, 89]. Compared to other approaches for T2I, our COS-GAN utilises significantly less Number of Parameters (NoP) and still achieves competitive performance with extraction of FG-BG semantic maps representation for every image. In Table 2, we compare results for Oxford-102 dataset. We only show quantitative results for IS and FID scores for evaluation, as R-precision scores are not available in the literature. We improve IS score from 4.06 to 4.28 and remarkably decrease FID from 40.31 to 28.63 along with high-quality segmentation masks.

Method	ACC	IoU	mIoU
Supervised U-Net	95.2	79.5	86.8
GrabCUT [60]	82.0	69.2	-
Chen et al. [10]	87.9	76.4	-
IEM [65]	88.3	76.8	79.1
IEM + SegNet [65]	89.6	**78.9**	80.8
COS-GAN	**90.9**	77.2	**81.7**

Table 4. Quantitative comparison of FG-BG semantic maps between our approach and other models on Oxford-102 dataset [47]

If the quality of the generated FG-BG images and masks is exceptional, they can serve the purpose of training segmentation networks using weak supervision. To evaluate the efficacy of the generated FG-BG masks, we have trained UNet [59] in weakly supervised approach using images and masks generated by COS-GAN for predicting segmentation maps for real image. We have evaluated the predicted masks on the standard test splits of the CUB and Oxford-102 datasets. The comparison of other approaches for generating FG-BG masks for CUB dataset is presented in Table 3 and for Oxford-102 dataset in Table 4. The maps produced by COS-GAN exhibit superior quality and represent a viable option for training various models in weakly supervised learning scenarios. Compared to SSA-GAN [37], which employs a segmentation approach for mask prediction and addresses only the foreground, our proposed method surpasses it by generating FG-BG masks of higher quality, as demonstrated in Table 3.

5. Conclusion

In this paper, we have proposed a novel GAN framework (COS-GAN) for text-to-image synthesis, which generates two images simultaneously and extracts their FG-BG segmentation masks using Co-attention mechanism. The presented method has illustrated that predicting segmentation maps on image attended features produces high-quality segmentation masks and improves the quality of images generated. We also propose a novel Spatial Conditioning Block that focuses on dedicated conditioning to the foreground and background, further boosting the model's performance and prompting the network to generate meaningful segmentation masks. We comprehensively have studied our model on CUB, Oxford-102, and COCO datasets and compared it with other state-of-the-art approaches.

References

[1] Rameen Abdal, Peihao Zhu, Niloy J. Mitra, and Peter Wonka. Labels4free: Unsupervised segmentation using stylegan. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13970–13979, October 2021. 2, 3, 5

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan, 2017. 2

[3] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, Dec 2017. 3

[4] Yaniv Benny and Lior Wolf. Onegan: Simultaneous unsupervised learning of conditional image generation, foreground segmentation, and fine-grained clustering. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI*, page 514–530, Berlin, Heidelberg, 2020. Springer-Verlag. 2, 3, 5, 8

[5] Adam Bielski and Paolo Favaro. *Emergence of Object Segmentation in Perturbed Generative Models*. Curran Associates Inc., Red Hook, NY, USA, 2019. 2, 8

[6] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis, 2019. 2, 3

[7] Andrew Brock, Theodore Lim, J. M. Ritchie, and Nick Weston. Neural photo editing with introspective adversarial networks, 2017. 2

[8] Yali Cai, Xiaoru Wang, Zhihong Yu, Fu Li, Peirong Xu, Yueli Li, and Lixian Li. Dualattn-gan: Text to image synthesis with dual attentional generative adversarial network. *IEEE Access*, 7:183706–183716, 2019. 2, 8

[9] Hong Chen, Yifei Huang, and Hideki Nakayama. Semantic aware attention based deep object co-segmentation. *arXiv preprint arXiv:1810.06859*, 2018. 2, 3

[10] Mickaël Chen, Thierry Artières, and Ludovic Denoyer. Unsupervised object segmentation by redrawing. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 8

[11] Ting Chen, Mario Lučić, Neil Houlsby, and Sylvain Gelly. On self-modulation for generative adversarial networks. In *International Conference on Learning Representations*, 2019. 5, 6

[12] Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu. Dynamic convolution: Attention over convolution kernels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[13] Ayushman Dash, John Cristian Borges Gamboa, Sheraz Ahmed, Marcus Liwicki, and Muhammad Zeshan Afzal. Tac-gan - text conditioned auxiliary classifier generative adversarial network, 2017. 1, 2

[14] 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. 2

[15] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat GANs on image synthesis. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. 2, 3

[16] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. Cogview: Mastering text-to-image generation via transformers. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 19822–19835. Curran Associates, Inc., 2021. 1, 2, 3

[17] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. *CoRR*, abs/2012.09841, 2020. 2, 3

[18] Zhengcong Fei, Mingyuan Fan, Junshi Huang, Xiaoming Wei, and Xiaolin Wei. Progressive denoising model for fine-grained text-to-image generation, 2022. 3

[19] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors, 2022. 1, 3

[20] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27, 2014. 2

[21] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10696–10706, June 2022. 1, 3

[22] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In I. Guyon, U. V. 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. 2

[23] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 3

[24] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. V. 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. 7

[25] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020. 3

[26] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scal-

ing up gans for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10124–10134, June 2023. 1, 3

[27] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation, 2018. 2

[28] 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. 3, 4

[29] Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Text-to-image generation grounded by fine-grained user attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 237–246, January 2021. 3

[30] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 4

[31] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11523–11532, June 2022. 3

[32] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip Torr. Controllable text-to-image generation. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 2

[33] Daiqing Li, Huan Ling, Seung Wook Kim, Karsten Kreis, Sanja Fidler, and Antonio Torralba. Bigdatasetgan: Synthesizing imagenet with pixel-wise annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21330–21340, June 2022. 3

[34] Weihao Li, Omid Hosseini Jafari, and Carsten Rother. Deep object co-segmentation. *arXiv preprint arXiv:1804.06423*, 2018. 2, 3

[35] Wenbo Li, Pengchuan Zhang, Lei Zhang, Qiuyuan Huang, Xiaodong He, Siwei Lyu, and Jianfeng Gao. Object-driven text-to-image synthesis via adversarial training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[36] 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

[37] 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. 2, 3, 5, 7, 8

[38] 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. 2, 7, 8

[39] 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*, pages 740–755, Cham, 2014. Springer International Publishing. 7

[40] Bingchen Liu, Kunpeng Song, Yizhe Zhu, Gerard de Melo, and Ahmed Elgammal. Time: Text and image mutual-translation adversarial networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(3):2082–2090, May 2021. 2, 7

[41] Andrew L. Maas, Awni Y. Hannun, and Andrew Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *in ICML Workshop on Deep Learning for Audio, Speech and Language Processing*, 2013. 5

[42] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Finding an unsupervised image segmenter in each of your deep generative models, 2021. 2, 3, 8

[43] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets, 2014. 1, 2

[44] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018. 2, 6

[45] Takeru Miyato and Masanori Koyama. cgans with projection discriminator, 2018. 1, 2

[46] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models, 2021. 3

[47] 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. 1, 2, 7, 8

[48] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier GANs. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2642–2651. PMLR, 06–11 Aug 2017. 1, 2

[49] 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. 4

[50] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. Learn, imagine and create: Text-to-image generation from prior knowledge. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 8

[51] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. Mirrorgan: Learning text-to-image generation by redescrip-

tion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2, 7
[52] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 3
[53] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. 1, 3
[54] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *CoRR*, abs/2102.12092, 2021. 2, 3
[55] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 3
[56] Scott Reed, Zeynep Akata, Santosh Mohan, Samuel Tenka, Bernt Schiele, and Honglak Lee. Learning what and where to draw, 2016. 1, 2
[57] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1060–1069. PMLR, 2016. 1, 2
[58] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022. 2, 3
[59] 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. 8
[60] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. "grabcut": Interactive foreground extraction using iterated graph cuts. In *ACM SIGGRAPH 2004 Papers*, SIGGRAPH '04, page 309–314, New York, NY, USA, 2004. Association for Computing Machinery. 8
[61] Shulan Ruan, Yong Zhang, Kun Zhang, Yanbo Fan, Fan Tang, Qi Liu, and Enhong Chen. Dae-gan: Dynamic aspect-aware gan for text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13960–13969, October 2021. 7
[62] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022. 1, 3
[63] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 2234–2242, Red Hook, NY, USA, 2016. Curran Associates Inc. 7
[64] Axel Sauer, Tero Karras, Samuli Laine, Andreas Geiger, and Timo Aila. Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis. volume abs/2301.09515, 2023. 3
[65] Pedro Savarese, Sunnie S. Y. Kim, Michael Maire, Greg Shakhnarovich, and David McAllester. Information-theoretic segmentation by inpainting error maximization, 2020. 3, 8
[66] Krishna Kumar Singh, Utkarsh Ojha, and Yong Jae Lee. Finegan: Unsupervised hierarchical disentanglement for fine-grained object generation and discovery. In *CVPR*, 2019. 2, 3, 5, 8
[67] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265, Lille, France, 07–09 Jul 2015. PMLR. 3
[68] Arulkumar Subramaniam, Athira Nambiar, and Anurag Mittal. Co-segmentation inspired attention networks for video-based person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 2
[69] Arulkumar Subramaniam, Jayesh Vaidya, Muhammed Abdul Majeed Ameen, Athira Nambiar, and Anurag Mittal. Co-segmentation inspired attention module for video-based computer vision tasks. *Computer Vision and Image Understanding*, 223:103532, 2022. 2
[70] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 7
[71] 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. 2, 7, 8
[72] Ming Tao, Songsong Wu, Xiaofeng Zhang, and Cailing Wang. Dcfgan: Dynamic convolutional fusion generative adversarial network for text-to-image synthesis. pages 1250–1254, 11 2020. 7
[73] Nontawat Tritrong, Pitchaporn Rewatbowornwong, and Supasorn Suwajanakorn. Repurposing gans for one-shot semantic part segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4475–4485, June 2021. 3

[74] Jayesh Vaidya, Arulkumar Subramaniam, and Anurag Mittal. Co-segmentation aided two-stream architecture for video captioning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2774–2784, January 2022. 2

[75] Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. Neural discrete representation learning. In I. Guyon, U. V. 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. 2, 3

[76] 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. 3, 4

[77] Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity, 2020. 5

[78] 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. 1, 2, 7, 8

[79] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 3, 5

[80] Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. Nüwa: Visual synthesis pre-training for neural visual world creation. *CoRR*, abs/2111.12417, 2021. 1

[81] Fuxiang Wu, Liu Liu, Fusheng Hao, Fengxiang He, and Jun Cheng. Text-to-image synthesis based on object-guided joint-decoding transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18113–18122, June 2022. 3

[82] 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. 1, 2, 4, 5, 7

[83] Yu Yang, Hakan Bilen, Qiran Zou, Wing Yin Cheung, and Xiangyang Ji. Unsupervised foreground-background segmentation with equivariant layered gans. *CoRR*, abs/2104.00483, 2021. 2, 3, 5, 8

[84] Guojun Yin, Bin Liu, Lu Sheng, Nenghai Yu, Xiaogang Wang, and Jing Shao. Semantics disentangling for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[85] Peiyu Yu, Sirui Xie, Xiaojian Ma, Yixin Zhu, Ying Nian Wu, and Song-Chun Zhu. Unsupervised foreground extraction via deep region competition. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 14264–14279. Curran Associates, Inc., 2021. 3, 8

[86] 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, 5, 6, 7

[87] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP, 10 2017. 8

[88] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1, 2, 3, 4, 5, 7, 8

[89] Zhenxing Zhang and Lambert Schomaker. DTGAN: dual attention generative adversarial networks for text-to-image generation. *CoRR*, abs/2011.02709, 2020. 2, 8

[90] 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. 1, 2, 7, 8

[91] 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. 1, 2, 7