Self-Distillation on Conditional Spatial Activation Maps for ForeGround-BackGround Segmentation

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

Abstract

Fine-grained image segmentation offers a simplified yet meaningful representation, but obtaining such representations for training large-scale models demands considerable human effort and cost. Existing strategies aim to predict these maps with limited or no training image pairs. When only a few train-label pairs are available, Semi-Supervised Segmentation (SSS) with the student-teacher paradigm is employed. Without labels, neural networks are designed to extract intermediate activation masks for unsupervised learning, mostly confined to 2-class Foreground-Background (FG-BG) segmentation.

FG-BG unsupervised segmentation typically relies on intricately designed large-scale Generative Adversarial Networks (GANs) to generate intermediate activation maps. Additionally, conditional GANs also are utilised with spatial conditioning maps to generate FG-BG maps for conditional image generations, facilitating the creation of synthetic datasets. Moreover, transferring annotations to real-world data often requires using another segmentation network trained in a weakly supervised manner.

Considering these multi-step approaches, we introduce a simple yet effective single-step approach that directly produces superior conditional FG-BG maps for images using a reconstruction network. Our proposed encoder-decoder network reconstructs the original image from slightly noisy inputs and generates precise conditional attention maps. These conditional attention maps are created by emulating the behaviour of deeper generator layers in spatial conditioning GANs and further refined using the student-teacher paradigm. Our approach stands out for its simplicity and efficiency compared to intricate multi-step methods or GAN-based designs.

1. Introduction

Deep learning has significantly advanced computer vision applications, especially in supervised tasks like classification [27, 35, 59, 61], segmentation [10, 11, 40, 46, 74], and object detection [22, 23, 53]. Real-world datasets [15, 16, 19, 39] have been crucial for early success by enabling large-scale training. Fine-grained datasets [15, 19, 38, 77, 78] have furthered deep learning in various domains like medical imaging [54], image-to-image translation [33], and controllable image generation [21, 67, 73]. But, as neural networks grow exponentially in parameters and complexity [7, 18, 51, 52, 56], training demands vast datasets. Curating large-scale datasets requires a massive amount of labour-intensive effort, cost, and time, leading to challenges in training, thus limiting the complexity and size of large-scale models. Furthermore, creating fine-grained segmentation datasets at such a large scale poses feasibility issues due to their effort-intensive nature compared to other tasks like object detection or classification.

Multiple approaches have been developed to facilitate neural networks based learning without labels or with few training pairs for fine-grained segmentation. These methods include unsupervised learning [41, 55, 69], where semantic maps are produced without labels, and semi-supervised segmentation [32, 43], which leverages a few training pairs. These approaches make use of available large-scale unlabelled data for training.

In SSS learning, the widely adopted method is consistency regularisation [47, 48]. This method requires generating pseudo-labels on label preserving augmented images for labelled and unlabelled data and then using them in standard supervised learning [40]. The generation of pseudo-labels can be accomplished using large-scale pre-trained segmentation networks [13] or neural networks whose weights are updated through Exponential Moving Average (EMA) [20].

In the absence of labelled samples, SSS training faces a challenge known as confirmation bias [5], where models tend to overfit to incorrectly predicted pseudo-labels. In contrast, unsupervised learning for segmentation maps takes a different approach by leveraging the inherent structures of neural networks for extracting intermediate spatial activation maps [66] that can be interpreted as segmentation maps [1, 8, 69]. However, the limitation of such activation maps is that it result in 2-class segmentation maps, classify-

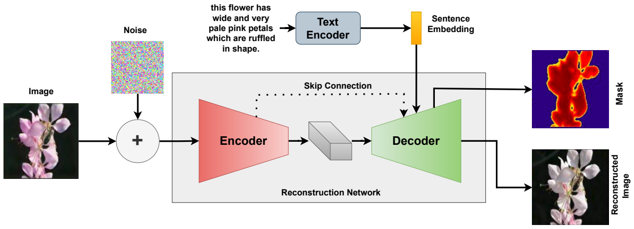

Figure 1. Proposed Encoder-Decoder reconstruction network producing conditional FG-BG maps while reconstructing image.

ing each pixel as either foreground or background.

Previous methods for FG-BG segmentation have relied on GraphCut [55] techniques and information maximisation approaches [12, 57]. While these methods offer faster inference times, they produce lower-quality results than generative techniques. Current strategies for FG-BG segmentation involve training GANs [24] to generate intermediate results for foreground, background, and masks [1, 6, 60, 69]. These masks combine foreground and background to create images and serve as an auxiliary intermediate output. The resulting image-mask pairs are utilised for training segmentation networks [11, 54] to predict masks for real images.

Furthermore, conditional GANs are employed in Text-to-Image synthesis to generate images while producing conditional activation maps, which are interpreted as FG-BG maps [3, 37]. However, such strategies necessitate extensive GANs and separate segmentation network training, increasing computational complexity and fragmented multi-stage learning. Additionally, the quality of masks predicted on real images is influenced by that of images and masks produced by the generative model.

We propose an end-to-end encoder-decoder reconstruction network for generating FG-BG maps as shown in Figure 1 to eliminate the need for multi-stage learning and extensive generative model training. This network reproduces original images from slightly noisy versions while extracting conditional spatial activation maps. Our approach mimics the behaviour of deep generator layers of spatial conditional GANs [68, 70] in the decoder for producing FG-BG maps. Spatial conditioning GANs utilise distinctive conditioning representations for each spatial location derived from conditional embeddings in deeper generator layers. Such conditioning is achieved through attention [68, 70, 72, 81] or predicted spatial activation maps [3, 37] for fine-tuning the spatial characteristics of the image at high resolutions (for creating realistic images). By replicating this spatial conditioning behaviour in the decoder of the proposed reconstruction network using predicted spatial activation maps, FG-BG maps are generated.

For introducing conditional aspects in reconstruction networks decoder, sentence embeddings are used from a pre-trained text encoder for captions associated with the image. Furthermore, these spatial activation maps undergo refinement using the student-teacher training paradigm [63]. Teachers and students possess similar structures, with the weights of teachers being updated through EMA. Our approach generates pseudo activation maps for refinement in a subtle difference to SSS learning. In summary, the contributions of the paper is outlined as follows:

- We propose an image Encoder-Decoder reconstruction network with shared features between encoder and decoder layers. This network is designed to produce FG-BG maps using spatial conditioning layers in the decoder, further refined through the student-teacher training approach. This configuration facilitates direct extraction of FG-BG semantic maps for images, resulting in faster and more straightforward training and enhanced quality of mask.
- We evaluate our approach on Caltech-UCSD birds [65] (CUB) and Oxford-102 flowers [45] datasets to assess its performance. The proposed method outperforms other techniques in FG-BG extraction, demonstrating superior results.

2. Related Work

This section provides an overview of literature related to the current paper.

Segmentation: Segmentation networks, explicitly trained to identify objects in images, often require extensive image-segmentation pairs for training [10, 11, 40, 46, 74]. Recent advancements have introduced dual-branch segmentation networks, focusing on spatial structures and contextual information to enhance predictions and inference efficiency [14, 30, 31, 49, 64, 71]. Our network uses this dual-branch design to predict intermediate spatial attention maps with contextual information and sharper spatial resolution for FG-BG map prediction through thresholding.

Semi-Supervised Segmentation: Unlabelled data utilisation is essential for semi-supervised learning [25]. Consistency regularisation is a popular method [47, 48], significantly impacted by pseudo-label quality [4, 36]. Pre-trained networks often generate these labels [13], and momentum encoders enhance this [75]. Simple EMA models with strong random intensity augmentation improve performance [76]. We employ EMA models with random intensity augmentation for refining spatial maps.

Unsupervised learning for FG-BG maps: Carefully designed GANs are utilised for generating synthetic FG-BG datasets to train segmentation networks in weak supervision. For instance, FineGAN [60] employs a hierarchical tree structural training with bounding boxes to create FG-BG masks, while OneGAN [6] generates FG-BG maps by incorporating a reconstruction loss involving pose, style, and shape vectors from both generators and discriminators. Labels4Free [1] employs a pre-trained StyleGAN to generate foreground while generating intermediary mask and

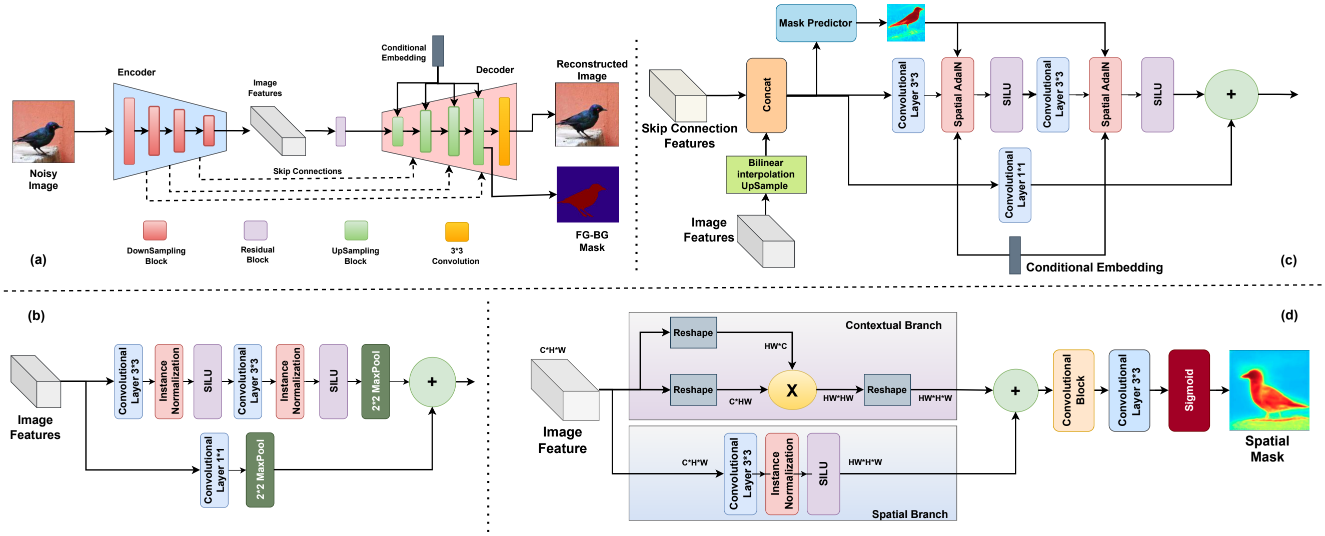

Figure 2. (a) Encoder-Decoder network with skip connection used in both student and teacher networks. Each UpSampling block is provided with conditional embeddings to extract attention maps with high activation for objects of interest. Activation maps from the final UpSampling block are used with thresholding for FG-BG maps. (b) The DownSampling block consists of two convolution blocks connected using a residual connection. (c) UpSampling blocks are designed to replicate spatial conditioning blocks in Text-to-Image synthesis GANs with intermediate masks predicted for specific spatial conditioning [37]. (d) Mask predictor uses dual branching to capture long correlations using contextual branches and preserve spatial structure from spatial branch for sharper mask predictions.

background outputs using separate networks. Yang et al. [69] and OneGAN [6] function without supervision but use fine-grained categories in the datasets. Melas-Kyriazi et al. [41] leverage the latent space of large-scale GAN models trained on extensive datasets and employ multi-stage learning for mask generation. SSA-GAN [37], a Text-to-Image synthesis spatial conditioning model, generates images conditioned on text and produces FG-BG masks from intermediate layers. These existing methods often necessitate fine-grained category information or multi-stage learning, relying heavily on intensive model training or specific dataset structures. In contrast, our method stands out operating unsupervised, requiring solely a conditioning vector representing the object of interest within the images. This unique simplicity and freedom from fine-grained categorisation enhance adaptability and usability of the method across different datasets.

Self-Distillation for Semantic Maps: MoCo [28] has introduced momentum encoder learning representations in unsupervised learning. This approach relies on having many negative samples to ensure robust feature learning. In contrast, BYOL [26] focuses on predicting image representation from a distinct viewpoint as predicted by the momentum encoder, avoiding use of negative examples. The DINO [8] approach utilises a Vision Transformer (ViT) [18] to predict distribution of the momentum encoder while incorporating a local-to-global view, which allows to produce spatial activation maps from self-attention layers of ViT. When graph [2] or spectral [42] methods are applied to these self-attention maps trained via DINO, they enable creation of semantic maps. Unlike self-distillation on final labels, we use self-distillation on intermediate activation maps, significantly enhancing the performance of predicting spatial maps.

Open Vocabulary zero-shot image segmentation: Attend-and-Excite [9] employs a transformer-based cross-attention between images and words to generate masks based on textual input in pre-trained models. MaskCLIP [79] relies on mask pseudo-labels from the vision-language model to train segmentation networks. ZegCLIP [80] uses deep prompt training to fine-tune prompts for learning better text-image matching on fixed CLIP for directly extending CLIP's zero-shot prediction capability from image to pixel-level. ZUTIS [58] generates a selected set of saliency maps as pseudo-labels for each concept (class) from the vision-language model to train a segmentation network. Unlike these approaches, which require text associated with each image, our method can work effectively with text, class labels, and image embeddings.

3. Methodology

This section provides an overview of our architecture (Section 3.1), student-teacher paradigm (Section 3.2) and training approach (Section 3.3). Our method employs a straightforward reconstruction network to restore original images from slightly distorted versions while concurrently generating conditional spatial activation maps to create FG-BG maps. For introducing minimal distortions to image, we use linear noise scheduler [29] for adding noise to images.

Our reconstruction network adopts an encoder-decoder

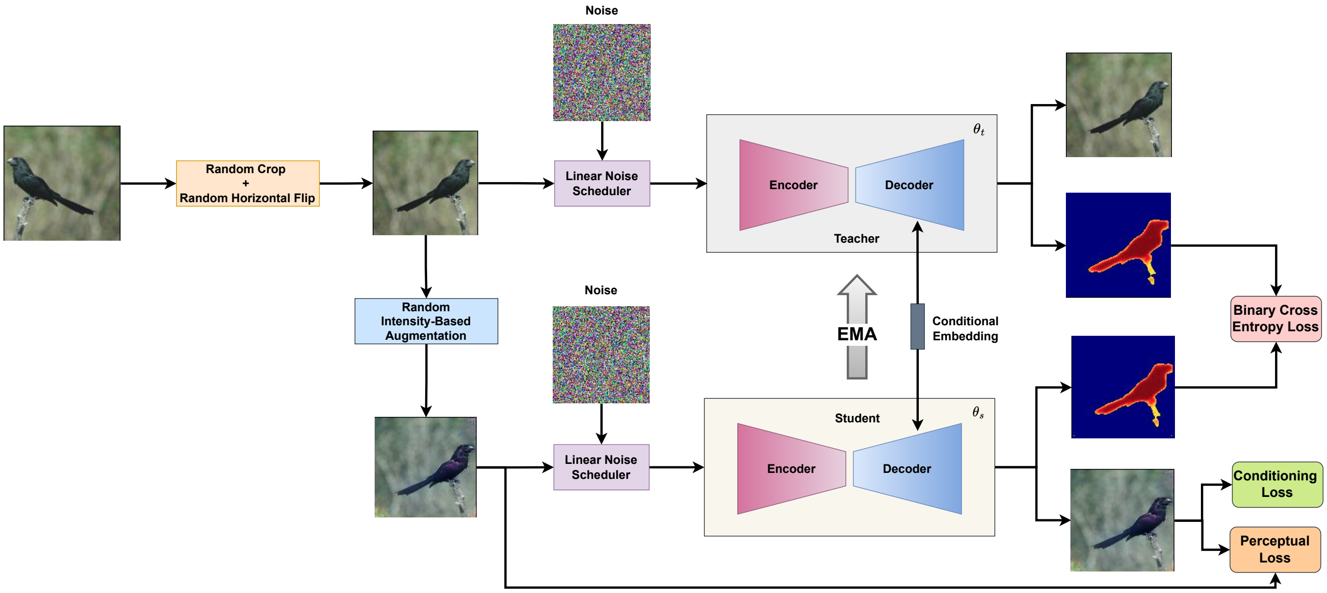

Figure 3. This approach adopts a student-teacher training strategy to refine intermediate conditional activation maps generated for images. Images with random augmentation are fed to student to introduce disagreements in the activation maps, contributing to improved mask quality [76]. The same conditional embeddings are provided to decoders in both networks. Conditioning loss is employed for information maximisation between image and conditional embeddings, guiding parts of the image to have higher spatial activation values. A linear noise scheduler adds minimal noise to images to encourage the network to learn more robust features.

structure, where in the decoder layers resemble the deeper layers found in spatial conditioning GANs [3, 37, 70, 72]. These layers generate intermediate spatial activation maps [3, 37] to refine features based on conditioning. Various representations, such as sentence embeddings for captions, one-hot vectors for class labels, or encoded global image representations from pre-trained models, can serve as conditional embeddings. However, these embeddings must accurately reflect the depicted object to elevate activation values in the spatial maps.

3.1. Network Architecture

The encoder-decoder architecture, illustrated in Figure 2, features skip connections between its components [54]. This design choice aims to maintain spatial structure, especially at higher resolutions, as the network is trained to produce original images by eliminating introduced distortions.

Noise is added to images and projected into a high-dimensional space using a linear convolutional layer. Then, downsampling blocks are utilised until the features reach a size of 8x8. These low-resolution features are then processed through upsampling blocks and a linear convolutional layer, converting them into image space.

Drawing inspiration from the concept of dual-branch segmentation networks [14, 30, 64, 71], which emphasises on spatial structure preservation and capture of long-range contextual information, the presented network employs a dual-branch prediction for attention maps. This is illustrated in mask predictor used in each upsampling block, as shown in Figure 3.

The predicted attention map from mask predictor is employed within two Spatial Adaptive Instance Normalisation (AdaIN) blocks. Each spatial AdaIN block utilises the predicted mask and conditional embeddings to modulate features using decoupled spatial conditioning [3]. Decoupled conditioning uses text to condition the foreground (or object of interest) and noise for the background using predicted spatial mask. As we do not use noise as input (given as input to GANs), for background conditioning, we replace noise with learnable parameters. The integration of conditioning embedding ***e*** and the mask ***Mask*** within spatial AdaIN is executed as follows:

$$\text{AdaIN}(x_t \mid e) = (\gamma_c) \cdot \frac{x_t - \mu(x_t)}{\sigma(x_t)} + (\beta_c) \tag{1}$$

$$\gamma_f = MLP_{\gamma_f}(concat[e, \gamma_1]) \tag{2}$$

$$\beta_f = MLP_{\beta_f}(concat[e, \gamma_2]) \tag{3}$$

$$\gamma_b = MLP_{\gamma_b}(\gamma_3) \tag{4}$$

$$\beta_b = MLP_{\beta_b}(\gamma_4) \tag{5}$$

$$\gamma_c = Mask \times \gamma_f + (1 - Mask) \times \gamma_b \tag{6}$$

$$\beta_c = Mask \times \beta_f + (1 - Mask) \times \beta_b \tag{7}$$

γ_1, γ_2, γ_3 and γ_4 are learnable parameters, and MLP stands for Multilayer Perceptron.

The spatial branch in mask predictor focuses on preserving structural details within the images, aiding in predicting sharper attention maps. Simultaneously, the contextual branch performs cross-spatial correlation over the image features that are normalised across channels, capturing long-range relationships. This configuration enables the integration of local semantics from the spatial branch and global correlations from the contextual branch. Intermediate maps are generated by combining local semantics from the spatial branch and global correlations from the contextual branch, which are further processed through a convolutional block. A sigmoid activation is applied to ensure meaningful activation values are within a normalised range of 0 to 1.

3.2. Student-Teacher Paradigm

As depicted in Figure 3, these maps undergo further enhancement through activation outputs from a momentum encoder. Our approach consists of two networks with similar structures, initialised with same weights, following the established student-teacher paradigm [63]. When presented with an image, the method begins with a random crop and random horizontal flips. This image is provided to teacher network to generate activation maps, similar to pseudo labels in SSS learning for subsequent refinement [13]. For generating activation maps for the student network, firstly, the image undergoes random intensity-based augmentation to introduce variations in the resulting activation maps, a technique shown to improve mask quality [76].

3.3. Training and Loss

Our method requires images and their corresponding noisy versions to train the reconstruction network. Further, this approach can even function without any pixel-level distortions. However, adding minimal distortion to images significantly improves the quality of the resulting FG-BG maps.

To create noisy images, a linear noise scheduler is utilised [29] to add noise gradually at each time step. It allows the generation of noisy images $\boldsymbol{x}_t$ from the original image $\boldsymbol{x}_0$ at any specified time step t. The generation of noisy images is done as follows:

$$\boldsymbol{x}_t = \sqrt{\alpha_t}\boldsymbol{x}_0 + \sqrt{1-\alpha_t}\boldsymbol{z}, \quad \boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{8}$$

$$\alpha_t = \prod_{s=1}^{t} (1-\beta_s) \tag{9}$$

For the network training, we use β_s as a fixed variance linear scheduler [29, 44] with a maximum number of timestamps $T = 1000$. However, t for distortion is uniformly chosen within 1 to 10. This selection ensures that generated images have minimal pixel distortion added to them.

Perceptual loss [34] is used on the student network and employed on extracted features for reconstructed and original images from pre-trained VGGNet [59]. The loss function is crucial for enforcing image reconstruction while preserving spatial structure during training. The reconstruction loss ($\mathcal{L}_{rec}$) can be expressed as follows:

$$\mathcal{L}_{rec} = \|\varphi(\boldsymbol{x_0}) - \varphi(\boldsymbol{\Phi_s}(\boldsymbol{x_t}))\|_2^2 \tag{10}$$

In the context of our method, $\boldsymbol{x}_0$ signifies images without any added noise, while $\boldsymbol{x}_t$ represents images with noise incorporated from the linear scheduler at a specific time-step t. The student network is denoted as $\boldsymbol{\Phi_s}$, and φ refers to the pre-trained VGGNet utilised for feature extractions.

The conditioning loss measures the alignment between the reconstructed image from the student network and the conditional embeddings. This helps information maximisation between the image and the conditioning aspect, influencing spatial activation maps to prioritise certain areas of the image for higher activation values. This directly impacts the Spatial AdaIN in the decoders, as exact conditional embeddings are utilised within this process.

The conditional embeddings employed in determining the foreground are the descriptive caption associated with the image. The conditioning loss, using global conditional embeddings, can be expressed as follows:

$$\mathcal{L}_{cond} = L_{sim}(f_g, C) \tag{11}$$

$$f_g = ENC_{image}(\boldsymbol{\Phi_s}(\boldsymbol{x_t})) \tag{12}$$

$$C = ENC_{cond}(text) \tag{13}$$

We utilise a pre-trained CLIP VIT-B/32 vision transformer encoder [18, 50] as our image encoder. We rely on a pre-trained CLIP sentence encoder for global conditioning embeddings associated with text. We compute similarity scores using Cosine Similarity (cos) between conditional embeddings C and global visual features f_g as $\cos(u, v) = u^T v/|u||v|$. The temperature hyper-parameter τ is involved in this computation. Subsequently, we apply the contrastive loss to maximise information between the image and global conditioning representations, expressed as follows:

$$\mathcal{L}_{\text{sim}}(f_{g_i}, C_i) = -\log \frac{\exp(Sim(f_{g_i}, C_i))}{\sum_{j=1}^{N} \exp(Sim(f_{g_i}, C_j))} \tag{14}$$

$$Sim(f_{g_i}, C_i) = \cos(f_{g_i}, C_i)/\tau \tag{15}$$

We refine spatial maps using binary cross-entropy loss (L_{bce}) using the student-teacher paradigm. This loss compares the intermediate mask predicted by the final decoder layers in the student network (M_s) with activation maps from the teacher network (M_t), acting as pseudo-labels. The formulation is as follows:

$$\mathcal{L}_{mask} = \mathcal{L}_{bce}(M_s, M_t) \tag{16}$$

The overall training loss to train the student network, with hyper-parameters λ_1, λ_2, and λ_3 controlling the impact of each loss individually, is given by:

$$\mathcal{L}_{student} = \lambda_1\mathcal{L}_{rec} + \lambda_2\mathcal{L}_{cond} + \lambda_3\mathcal{L}_{mask} \tag{17}$$

The methodology follows a student-teacher paradigm, depicted in Figure 3. After each training iteration, the teacher model weights gradually adapt using a momentum parameter α set to 0.995, updating based on the students weights through EMA:

$$\theta_t \leftarrow \alpha\theta_t + (1-\alpha)\theta_s \tag{18}$$

4. Experiments

This section presents the datasets and evaluation metrics employed in our experiments. Subsequently, we assess performance of our proposed model on these datasets and compare it with that of existing approaches in the literature. For detailed training particulars and hyperparameters, supplementary material may be referred.

Datasets: We evaluate our model quantitatively on two datasets: Caltech-UCSD birds [65] (CUB) and Oxford-102 flowers [45]. These datasets consist of images, captions, class labels for fine-grained categories of birds and flowers, and FG-BG segmentation maps. The CUB dataset contains 11,788 images, while the Oxford-102 dataset has 8,189 images.

Evaluation metrics: Mean Intersection over Union (mIoU), Intersection over Union (IoU), and per-pixel accuracy (ACC) are used to evaluate the quality of FG-BG maps. The mIoU measures the average intersection over union of accurately classified foreground and background pixels. The IoU metric calculates the intersection over union value specifically for the foreground. Meanwhile, the ACC metric quantifies the percentage of correctly classified pixels.

Method	ACC	IoU	mIoU
Supervised U-Net	98.0	88.8	93.2
GrabCUT [55]	72.6	36.0	52.3
FineGAN [60]	-	44.5	-
OneGAN [6]	-	55.5	-
ReDO [12]	84.5	42.6	-
IEM + SegNet [57]	89.3	55.1	71.4
Melas-Kyriazi et al. [41]	92.1	66.1	-
Yang et al. [69]	94.3	69.7	81.7
DSM [42]	-	-	76.9
DeepCut [2]	-	-	78.2
Ours	**94.2**	**76.2**	**84.5**

Table 1. Quantitative comparison between our approach and those of other weak and unsupervised methods on the CUB dataset that do not use ground-truth segmentation maps. ”-” indicates values are unreported.

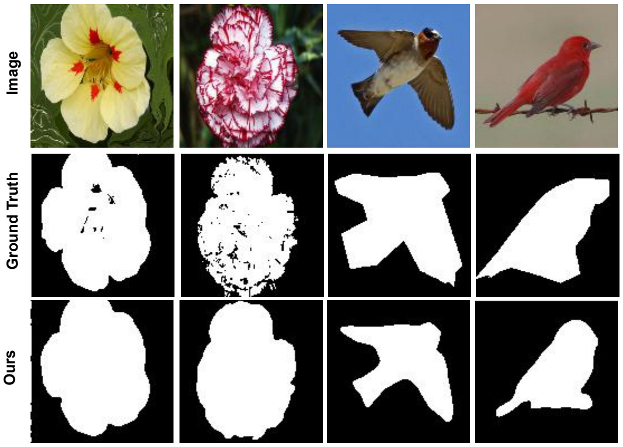

Figure 4. Illustration of FG-BG semantic maps generated by our approach on Oxford-102 and CUB datasets.

Method	ACC	IoU	mIoU
Supervised DeepLabV3 [11]	97.56	79.64	88.47
SSA-GAN [37]	61.6	20.4	39.4
COS-GAN [3]	**94.6**	73.2	83.3
Ours	94.2	**76.2**	**84.5**

Table 2. Quantitative comparison of FG-BG semantic maps predicted by our approach and those of the Text-to-Image synthesis GANs producing intermediate masks.

4.1. Results

The comparison in Table 1 showcases the superiority of our approach in comparision with other methods in unsupervised and weakly supervised FG-BG map generation on the CUB dataset. Consistent with prior approaches, our model generates maps at 128 × 128 resolution. Notably, it surpasses GAN-based methods [69] and techniques relying on DINO features for mask extraction [2, 42]. Impressively, our model achieves this performance with single-stage training, underscoring the simplicity and efficiency of our approach in comparison with those of other methods.

To further highlight mask quality and efficiency of our approach, we have compared our approach with Text-to-Image synthesis conditional GANs like SSA-GAN [37] and COS-GAN [3] on the CUB dataset. The comparison, presented in Table 2 for the standard CUB test split, highlights the superior performance of our approach without training of GANs. By emulating spatial conditioning blocks in the decoder similar to the deeper layers of COS-GAN’s generator, our method demonstrates the benefits of avoiding multi-stage learning and reliance on conditional generative models for synthesising datasets. Figure 4 may be referred for visual representations of the generated FG-BG maps for the CUB and Oxford-102 datasets.

Various conditional embeddings reflecting the object of interest can be used for achieving desired spatial activations in the proposed model. We have trained the model with different conditional embeddings, presented the results in Table 3 and compared with ZUTIS [58], open vocabulary model for segmentation. The embeddings include: ’One-Hot’: Class label as a one-hot vector representation,

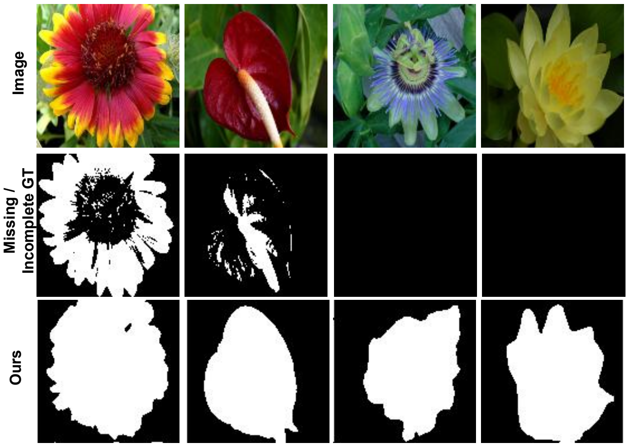

Figure 5. Illustration of missing and incomplete ground truths (GT) masks on Oxford-102 dataset [45] and FG-BG maps using our approach. Ground truth masks are available in public domain.

Method	Emb Type	ACC	IoU	mIoU
ZUTIS [58]	Generic	-	-	72.5
	Name	-	-	72.6
Ours	One-Hot	92.0	66.0	78.3
	Name	92.3	70.6	81.0
	Image	92.6	71.9	81.4
	Text	**94.2**	**76.2**	**84.5**

Table 3. Quantitative comparison of semantic maps from proposed model with those of ZUTIS using different conditional embeddings.

'Name': Name of the bird as a caption from CLIP text-encoder, 'Image': Global embedding of the image with a random crop from CLIP VIT-B/32 image-encoder, 'Text': Global sentence embedding from CLIP text-encoder for the caption associated with the image, and 'Generic': Single word 'BIRD' used as a high-level category.

The 'One-Hot' representation, 'Name' captions, and 'Generic' high-level categories may capture less contextual information than the 'Text' and 'Image' embeddings, potentially impacting the model's ability to recognise fine-grained details. The model's superior performance with sentence embeddings for captions aligns with expectations, as captions offer richer contextual information about the objects in the images. This richer context likely aids the model in better understanding and segmenting the objects of interest.

We have showcased the outcomes of our model on the Flower dataset in Table 4. Although the increase in IoU score is marginal, a deeper analysis of the results has unveiled an important insight that over 300 images in the dataset have complete background as the ground truth. Furthermore, numerous images contain incomplete masks that fail to encompass the entirety of the picture, displayed in Figure 5, emphasising the need for robust models to generate superior-quality masks with an efficient training strategy.

Method	ACC	IoU	mIoU
Supervised U-Net	95.2	79.5	86.8
GrabCUT [55]	82.0	69.2	-
ReDO [12]	87.9	76.4	-
IEM [57]	88.3	76.8	79.0
IEM + SegNet [57]	89.6	78.9	80.8
COS-GAN [3]	**90.9**	77.2	81.7
Ours	90.1	**79.6**	**81.9**

Table 4. Quantitative comparison of FG-BG semantic maps between our approach and other models on Oxford-102 dataset [45].

4.2. Ablation Studies

4.2.1 Contextual and Spatial Branch

We employ a dual-branch setting for intermediate mask predictions, utilising contextual branches to learn long-range correlation and spatial branches to produce sharper activation maps. To assess the impact of each branch on activation maps, we individually have applied them and summarise the results in Table 5. When using only spatial components, strong activation values for high-frequency information like edges and textures are detailed but failed to capture the entire object. Conversely, using only a contextual branch limits the quality of segmentation maps in preserving sharper shapes. These findings emphasise the complementary nature of the branches and their collective impact on generating superior activation maps for segmentation purposes. Visual results may be found in the supplementary material.

Spatial	Contextual	ACC	IoU	mIoU
✓		91.4	69.2	79.4
	✓	92.9	72.1	81.7
✓	✓	**94.2**	**76.2**	**84.5**

Table 5. Quantitative comparison of FG-BG semantic maps using Contextual and Spatial branches for mask prediction on CUB Dataset.

4.2.2 Impact of Losses

Our approach uses various loss functions and augmentation strategies to optimise spatial maps and reconstructed images. Table 6 provides a comprehensive overview of the distinct losses employed in the training process. It illustrates the influence of each loss and the augmentation strategy utilised. The Reconstruction Loss (RL) emphasises spatial structure preservation, resulting in sharper spatial attention maps. However, using RL alone, without Conditioning Loss (CL), does not guarantee the generation of the desired spatial maps. The CL facilitates information maximisation between the image and conditioning aspects, imposing high spatial activation maps for regions in images reflecting the embeddings. Additionally, the Random Augmentation (RA) approach, involving label-preserving augmentation on images, allows for further refinement of maps using Mask Loss (ML). The supplementary material offers visual insights into these impacts for a more comprehensive understanding.

RL	CL	ML	RA	ACC	IoU	mIoU
✓				81.6	50.9	64.1
	✓			91.9	64.6	77.6
✓	✓			93.0	70.7	81.2
✓	✓	✓		93.6	74.1	83.1
✓	✓	✓	✓	**94.2**	**76.2**	**84.5**

Table 6. Impact of losses and Random Augmentation for generating FG-BG semnatic maps on CUB Dataset.

4.2.3 Image Distortion

Introducing slight noise to generate inputs can effectively improve FG-BG maps in our proposed framework. Employing a linear noise scheduler as a data augmentation technique notably improves the quality of produced FG-BG maps, as outlined in Table 7. Limiting the distortion applied to the original image is crucial for improved extraction of FG-BG masks, as excessive distortion adversely impacts both reconstruction quality and attention map sharpness, as observed in Table 8. Visual representations of these findings are available in the supplementary material. Remarkably, even without any distortion, this method demonstrates competitive outcomes.

Method	ACC	IoU	mIoU
No Distortions	92.9	73.4	82.3
Salt and Pepper	93.1	69.9	80.8
Colour Jitter	93.2	70.6	81.3
Gaussian Noise	80.6	43.8	60.5
Linear noise	**94.2**	**76.2**	**84.5**

Table 7. Quantitative comparison of FG-BG semantic maps between various approaches for introducing distortions in images.

4.2.4 Noise Injection

Using a linear noise scheduler to introduce minimal image noise from the initial ten steps has proven effective in maintaining the quality of the generated maps, as highlighted in Table 8. By comparing this approach with noise addition at various steps, we have observed that increased noise additions result in lower map quality, reinforcing our choice of using noise from the first ten steps.

Time-stamp	ACC	IoU	mIoU
1	93.8	73.8	82.6
5	93.2	70.5	81.2
10	93.0	73.7	82.5
20	91.7	70.4	80.0
50	91.0	68.5	78.8
100	90.1	55.8	72.2
Rand(1,10)	**94.2**	**76.2**	**84.5**

Table 8. Quantitative comparison of FG-BG semantic maps generated with different time-stamps using a linear scheduler [29] for adding noise in images on the CUB dataset.

4.2.5 Image and Text Encoders

Using CLIP for text and image encoders in our method, we demonstrate the flexibility across different encoder choices. While we utilise a pre-trained Inception-V3 [62] as the image encoder, we conduct experiments with diverse text encoder strategies, as detailed in Table 9. Our results show competitive performance comparable to that of CLIP encoders by employing various text encoders. This enhancement in CLIP-based results is attributed to extensive transformer training, wherein text embeddings display improved visual alignment with images.

Text	Image	ACC	IoU	mIoU
DAMSM [68]	Inception-V3	92.9	69.2	80.4
BERT [17]	Inception-V3	93.4	71.3	81.7
CLIP	Inception-V3	93.6	74.6	83.4
CLIP	CLIP	**94.2**	**76.2**	**84.5**

Table 9. Quantitative comparison of FG-BG semantic maps using different Text and Image Encoders.

5. Conclusion

Our proposed method uses a reconstruction neural network, reproducing original images from minimally distorted images while extracting intermediate conditioning activation maps and further refining these maps using the student-teacher paradigm. These intermediate maps are generated by replicating the behaviour of spatial conditioning blocks in GANs. A notable advantage of our approach is its simplicity and improvement in predicted mask quality. Unlike other methods that rely on complex generative models for synthetic datasets, we work directly with images to extract semantic maps. This makes our method efficient and effective without need for extra synthetic data.

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, 2021. 1, 2

[2] Amit Aflalo, Shai Bagon, Tamar Kashti, and Yonina Eldar. Deepcut: Unsupervised segmentation using graph neural networks clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 32–41, 2023. 3, 6

[3] Yeruru Asrar Ahmed and Anurag Mittal. Unsupervised co-generation of foreground-background segmentation from text-to-image synthesis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5058–5069, 2024. 2, 4, 6, 7

[4] Eric Arazo, Diego Ortego, Paul Albert, Noel E. O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning, 2020. 2

[5] Eric Arazo, Diego Ortego, Paul Albert, Noel E. O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning, 2020. 1

[6] 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, 6

[7] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020. 1

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 1, 3

[9] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models, 2023. 3

[10] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2018. 1, 2

[11] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *The European Conference on Computer Vision (ECCV)*, 2018. 1, 2, 6

[12] Mickaël Chen, Thierry Artières, and Ludovic Denoyer. Unsupervised object segmentation by redrawing. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. 2, 6, 7

[13] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2, 5

[14] Yinpeng Chen, Xiyang Dai, Dongdong Chen, Mengchen Liu, Xiaoyi Dong, Lu Yuan, and Zicheng Liu. Mobileformer: Bridging mobilenet and transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5270–5279, 2022. 2, 4

[15] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 1

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. 8

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 1, 3, 5

[19] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015. 1

[20] Geoff French, Timo Aila, Samuli Laine, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, high-dimensional perturbations, 2020. 1

[21] 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

[22] Ross Girshick. Fast r-cnn. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015. 1

[23] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 1

[24] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014. 2

[25] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Advances in Neural Information Processing Systems*. MIT Press, 2004. 2

[26] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning, 2020. 3

[27] 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. 1

[28] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[29] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pages 6840–6851. Curran Associates, Inc., 2020. 3, 5, 8

[30] Yuanduo Hong, Huihui Pan, Weichao Sun, and Yisong Jia. Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes. *CoRR*, abs/2101.06085, 2021. 2, 4

[31] Zilong Huang, Yunchao Wei, Xinggang Wang, Wenyu Liu, Thomas S. Huang, and Humphrey Shi. Alignseg: Feature-aligned segmentation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):550–557, 2022. 2

[32] W.-C. Hung, Y.-H. Tsai, Y.-T. Liou, Y.-Y. Lin, and M.-H. Yang. Adversarial learning for semi-supervised semantic segmentation. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018. 1

[33] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[34] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision – ECCV 2016*, pages 694–711, Cham, 2016. Springer International Publishing. 5

[35] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*. 2012. 1

[36] Dong-Hyun Lee. Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)*, 2013. 2

[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, 2022. 2, 3, 4, 6

[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 *Computer Vision – ECCV 2014*, 2014. 1

[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 *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. 1

[40] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015. 1, 2

[41] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Finding an unsupervised image segmenter in each of your deep generative models, 2021. 1, 3, 6

[42] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization. In *CVPR*, 2022. 3, 6

[43] S. Mittal, M. Tatarchenko, and T. Brox. Semi-supervised semantic segmentation with high- and low-level consistency. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(4):1369–1379, 2021. 1

[44] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models, 2021. 5

[45] 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*, 2008. 2, 6, 7

[46] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1520–1528, 2015. 1, 2

[47] Avital Oliver, Augustus Odena, Colin Raffel, Ekin D. Cubuk, and Ian J. Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, page 3239–3250, Red Hook, NY, USA, 2018. Curran Associates Inc. 1, 2

[48] Yassine Ouali, Céline Hudelot, and Myriam Tami. An overview of deep semi-supervised learning. *ArXiv*, abs/2006.05278, 2020. 1, 2

[49] Rudra P K Poudel, Stephan Liwicki, and Roberto Cipolla. Fast-scnn: Fast semantic segmentation network, 2019. 2

[50] 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. 5

[51] 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. 1

[52] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. 1

[53] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2015. 1

[54] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. 1, 2, 4

[55] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. "grabcut": Interactive foreground extraction using iterated graph cuts. In *ACM SIGGRAPH 2004 Papers*, page 309–314, New York, NY, USA, 2004. Association for Computing Machinery. 1, 2, 6, 7

[56] 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

[57] Pedro Savarese, Sunnie S. Y. Kim, Michael Maire, Greg Shakhnarovich, and David McAllester. Information-theoretic segmentation by inpainting error maximization, 2020. 2, 6, 7

[58] Gyungin Shin, Samuel Albanie, and Weidi Xie. Zero-shot unsupervised transfer instance segmentation. In *CVPRW*, 2023. 3, 6, 7

[59] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 1, 5

[60] Krishna Kumar Singh, Utkarsh Ojha, and Yong Jae Lee. Finegan: Unsupervised hierarchical disentanglement for fine-grained object generation and discovery. In *CVPR*, 2019. 2, 6

[61] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015. 1

[62] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016. 8

[63] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 1195–1204, Red Hook, NY, USA, 2017. Curran Associates Inc. 2, 5

[64] Qiang Wan, Zilong Huang, Jiachen Lu, Gang Yu, and Li Zhang. Seaformer: Squeeze-enhanced axial transformer for mobile semantic segmentation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 2, 4

[65] 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. 2, 6

[66] 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)*, 2018. 1

[67] 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

[68] 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)*, 2018. 2, 8

[69] 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. 1, 2, 3, 6

[70] 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)*, 2019. 2, 4

[71] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2, 4

[72] 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, 2021. 2, 4

[73] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 1

[74] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2

[75] Zhen Zhao, Luping Zhou, Yue Duan, Lei Wang, Lei Qi, and Yinghuan Shi. Dc-ssl: Addressing mismatched class distribution in semi-supervised learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9747–9755, 2022. 2

[76] Zhen Zhao, Lihe Yang, Sifan Long, Jimin Pi, Luping Zhou, and Jingdong Wang. Augmentation matters: A simple-yet-effective approach to semi-supervised semantic segmentation. In *CVPR*, 2023. 2, 4, 5

[77] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5122–5130, 2017. 1

[78] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset, 2018. 1

[79] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision (ECCV)*, 2022. 3

[80] Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, and Yifan Liu. Zegclip: Towards adapting clip for zero-shot semantic segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[81] 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)*, 2019. 2