Supplementary: 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

1. Implementation Details

The presented method is implemented using PyTorch framework [5][1] and to optimise the network using Adam optimiser [1] with following hyper parameters: batch size = 32, $\lambda_1 = 0.01$, $\lambda_2 = 1$, $\lambda_3 = 0.001$, learning rate = 0.00001 and default beta values. To obtain the final FG-BG mask, we set a threshold for activation values that are greater than 0.5 and consider them as the foreground. The model is trained for 100 epochs on CUB and Oxford-102 datasets (takes a day in 2 NVIDIA 1080Ti GPUs).

2. Architecture Details

In this section, a more detailed description of the internal architecture of our proposed network is provided. The model has been implemented using the Pytorch [5] framework. The network utilised is based on a single Encoder-Decoder design, similar to UNet, with the encoder and decoder sharing features to retain the overall spatial structure of the images. The network is a simple reconstruction model trained at a resolution of 128×128. Overall network architecture details are described in Table 1.

3×3 Convolution $\longrightarrow 64 \times 128 \times 128$
Downsampling Block $\longrightarrow 128 \times 64 \times 64$
Downsampling Block $\longrightarrow 256 \times 32 \times 32$
Downsampling Block $\longrightarrow 512 \times 16 \times 16$ Downsampling Block $\longrightarrow 1024 \times 8 \times 8$
ResBlock $\longrightarrow 1024 \times 8 \times 8$
Upsampling Block $\longrightarrow 512 \times 16 \times 16$
Upsampling Block $\longrightarrow 256 \times 32 \times 32$
Upsampling Block $\longrightarrow 128 \times 64 \times 64$
Upsampling Block $\longrightarrow 64 \times 128 \times 128$
3×3 Convolution $\longrightarrow 3 \times 128 \times 128$

Table 1. UNet type architecture with shared weights between multiple features to extract attention maps. Feature sharing between DownBlock and UpBlock for preserving spatial features of the images.

The proposed DownBlocks are made up of two convolutional blocks, each consisting of a Convolutional Layer with Instance Normalisation, followed by a SiLU activation function. We repeat this process, passing features through the down-

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

sampling blocks, until the spatial resolution of the features reduces to 8×8. We then use a set of upsampling blocks and a linear convolutional layer to project these low-resolution features back into the image space. Within each UpBlock, feature sharing between the encoder and decoder is implemented to preserve the structural information of the image at higher resolutions.

Each UpBlock comprises of a Bilinear Upsampling followed by a dual-branch mask prediction and two conditional convolutional blocks. Each Conditional Convolutional block consists of a Convolutional Layer, proposed Spatial AdaIN that uses predicted semantic maps, along with conditional embeddings, and a SiLU activation function. The activation maps generated from the mask prediction layer in the final UpBlock are used to produce FG-BG masks. For extracting the FG-BG mask, we threshold the activation maps with values greater than 0.5 to be foreground.

3. More Qualitative Results

To further validate the easy scalability of our approach, we have trained the proposed model, as shown in Table 1, on MS-COCO [2] and on CUHK-PEDES [7] with no modifications. MS-COCO dataset is a multi-object dataset; and each image consists of 5 captions. CUHK-PEDES dataset consists of images of humans trained for the Person Reidentification task. Each image consists of a single human and has two captions associated with it. We illustrate additional qualitative results of our model on CUB [6] dataset in Figure 1, on Oxford-102 [4] dataset in Figure 2, on MSCOC dataset in Figure 8, and on CUHK-PEDES dataset in Figure 9.

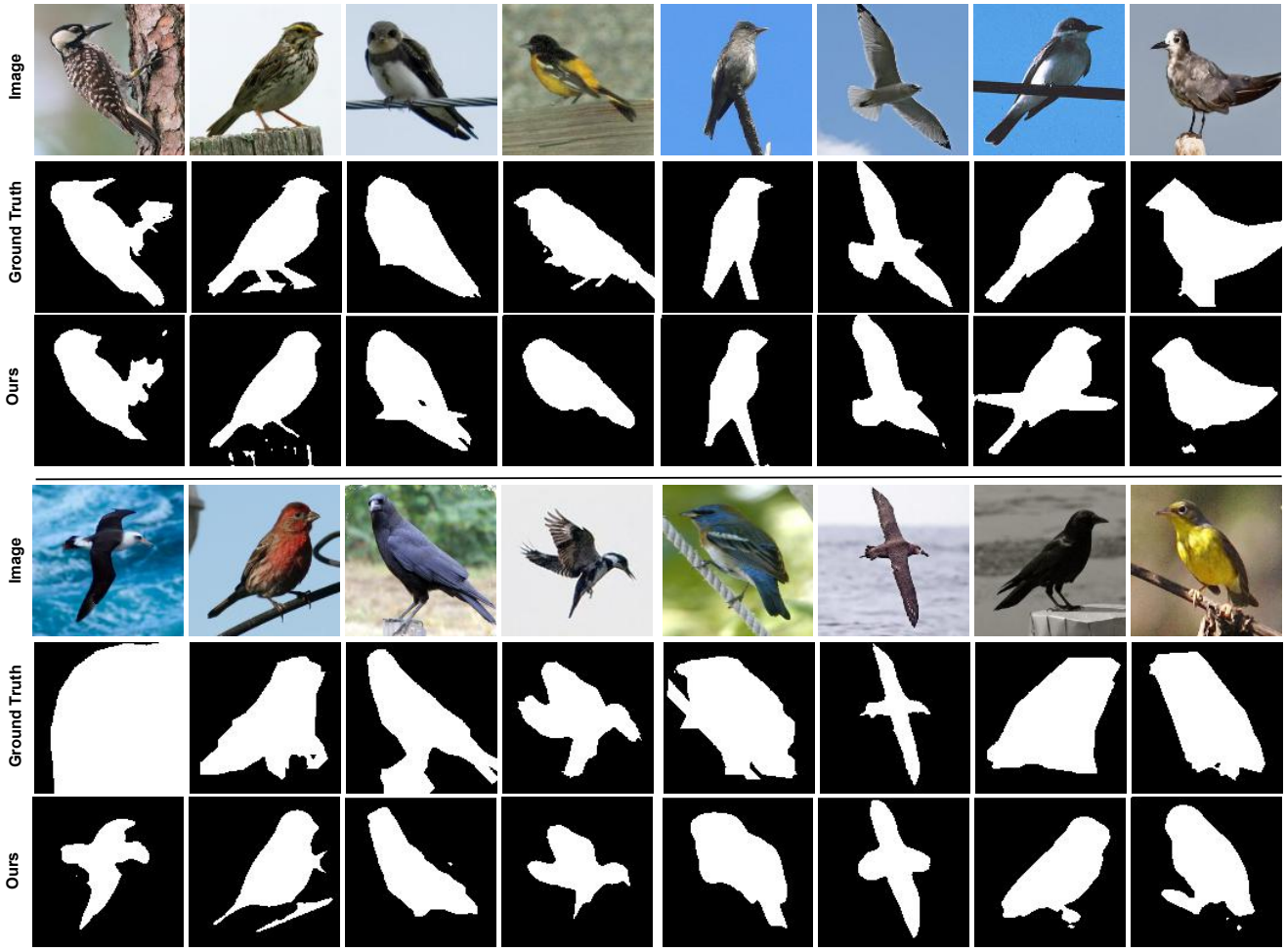


Figure 1. Foreground-Background Mask generated using our approach on CUB dataset [6].

Figure 2. Foreground-Background Mask generated using our approach on Oxford-102 dataset [4].

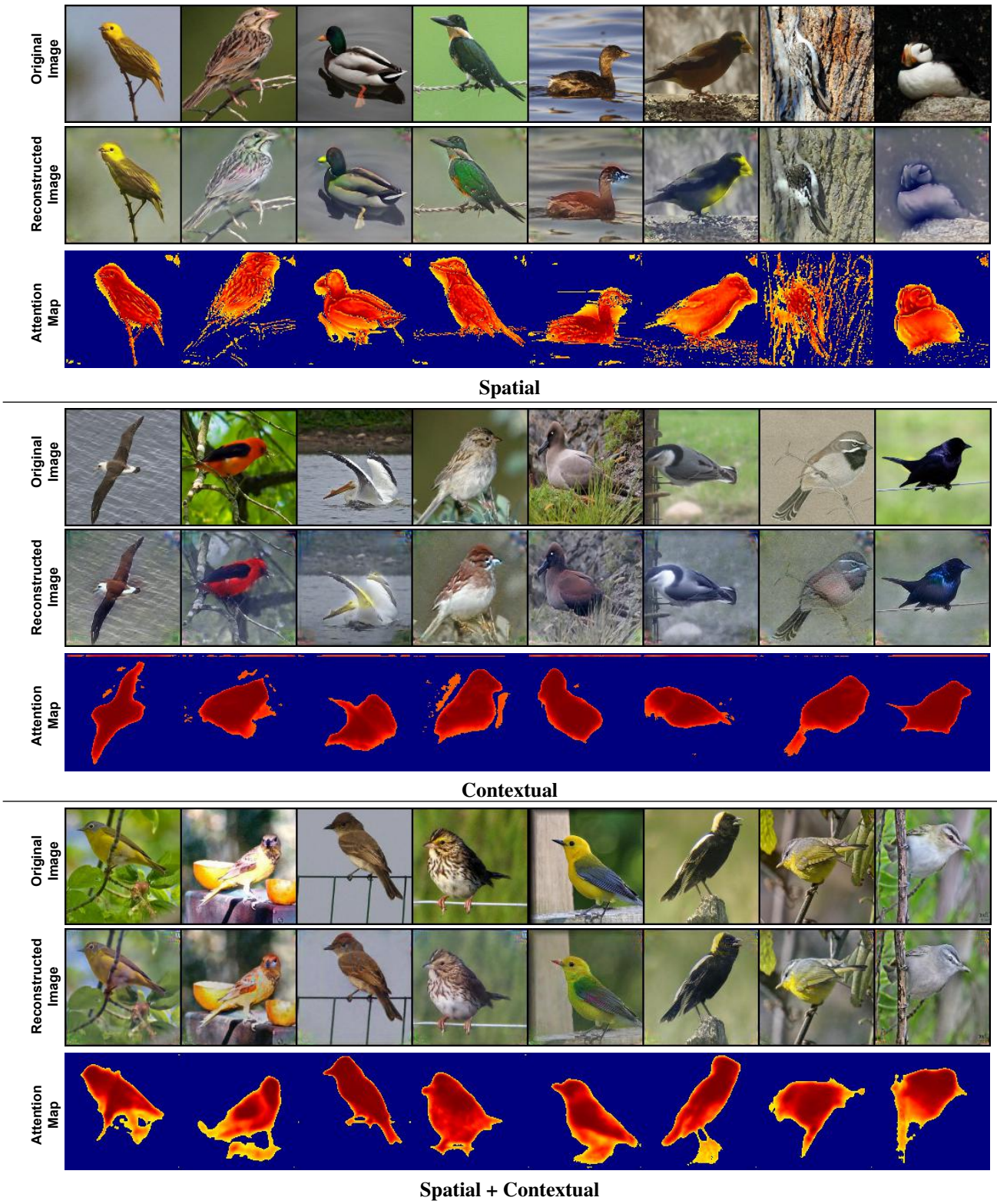


Figure 3. Impact of Spatial and Contextual branches on attention maps in dual branch mask predictor used. Spatial branch captures texture and shapes better than those captured by contextual branch, but fails to capture the overall object. Combining both these branches results in superior quality masks.

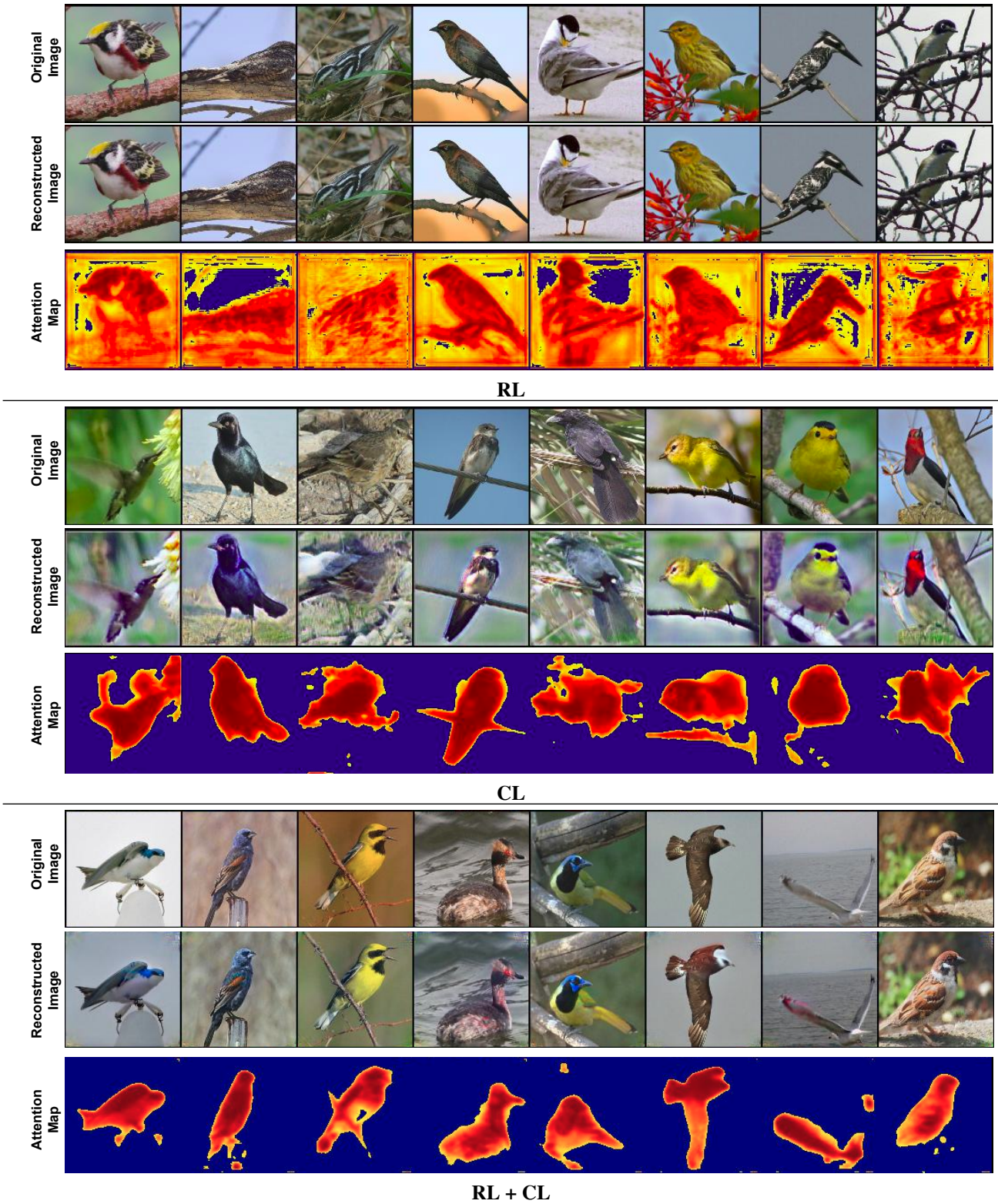


Figure 4. Foreground-Background (FG-BG) masks are generated using various employed losses. When using only Reconstruction Loss (RL), the resulting images are sharp, but the attention maps lack focus on the intended object of interest. On the other hand, applying Conditioning Loss (CL) directs the attention maps towards the object of interest. This loss amplifies the mutual information between the text and image. Combining both losses achieves a right balance, yielding sharp attention maps that are correctly focused on the intended object.

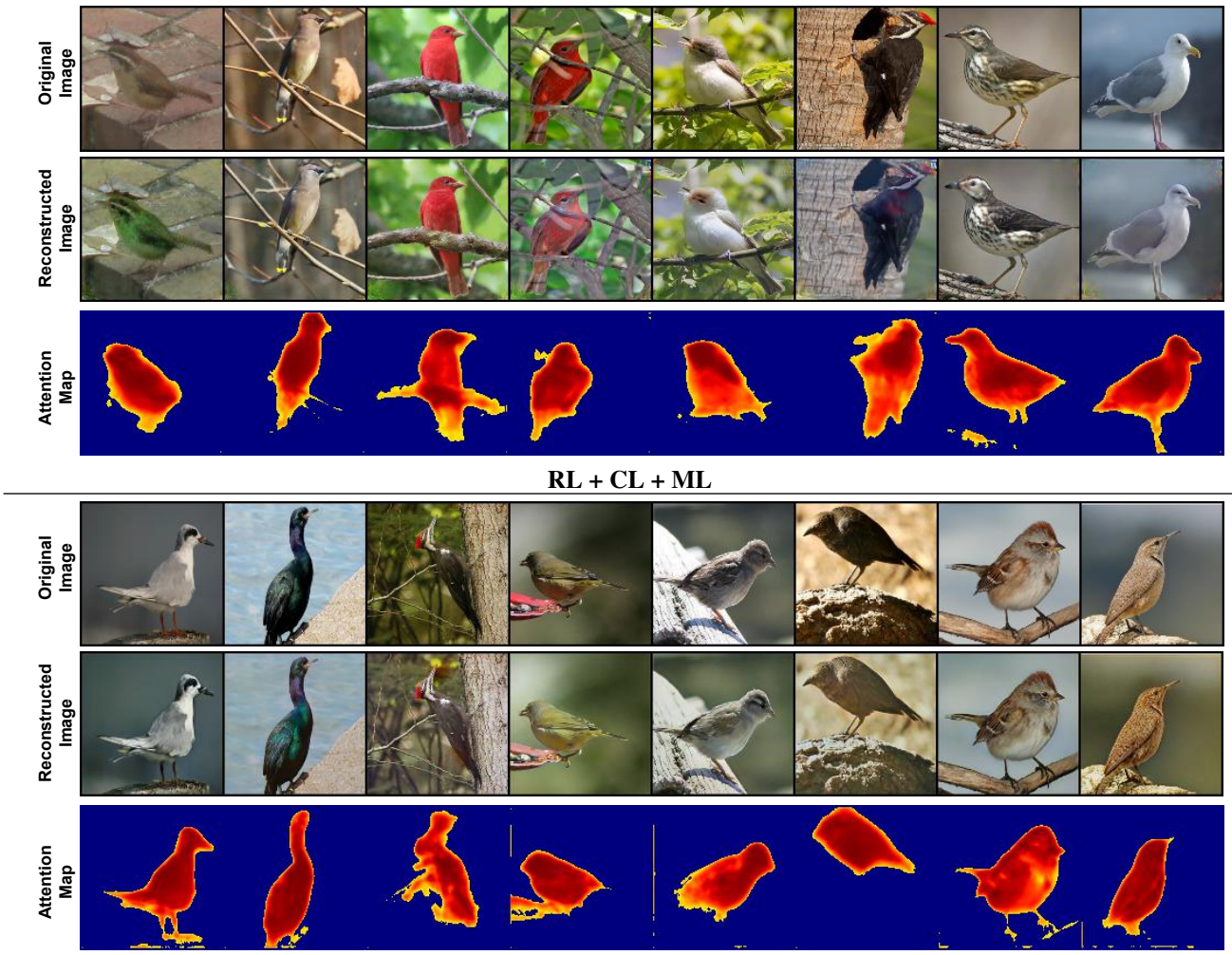
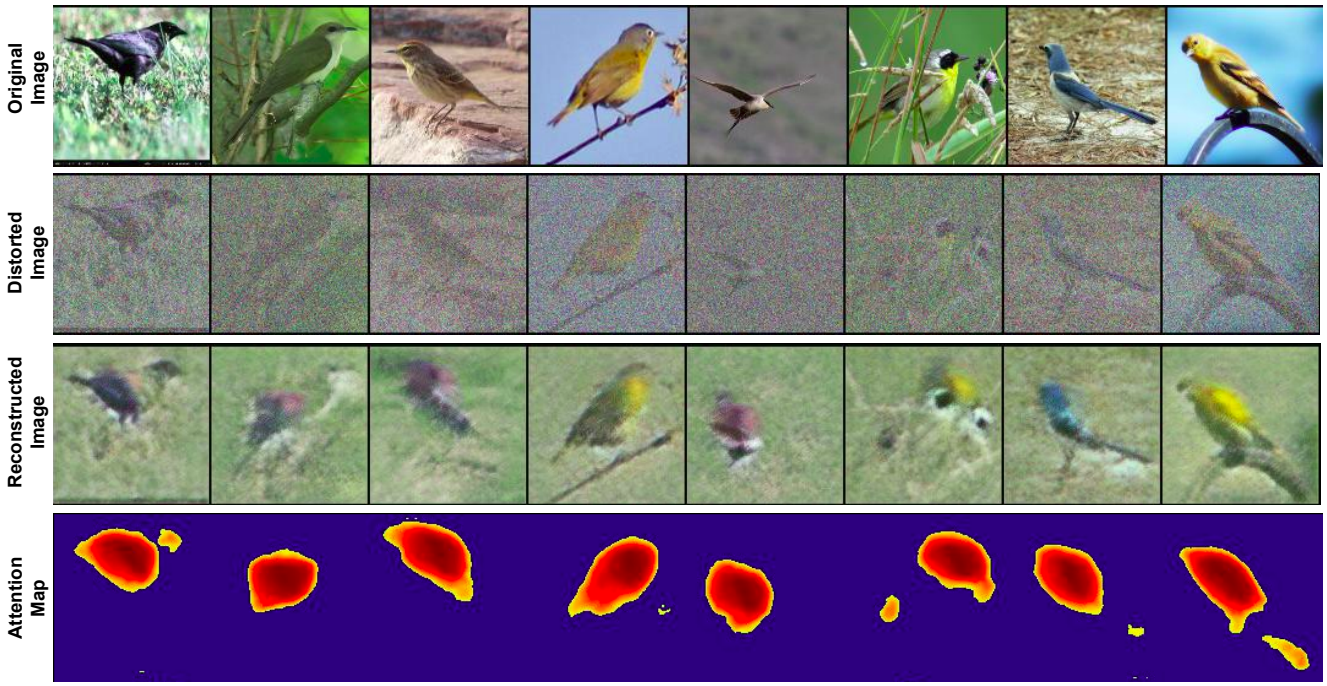
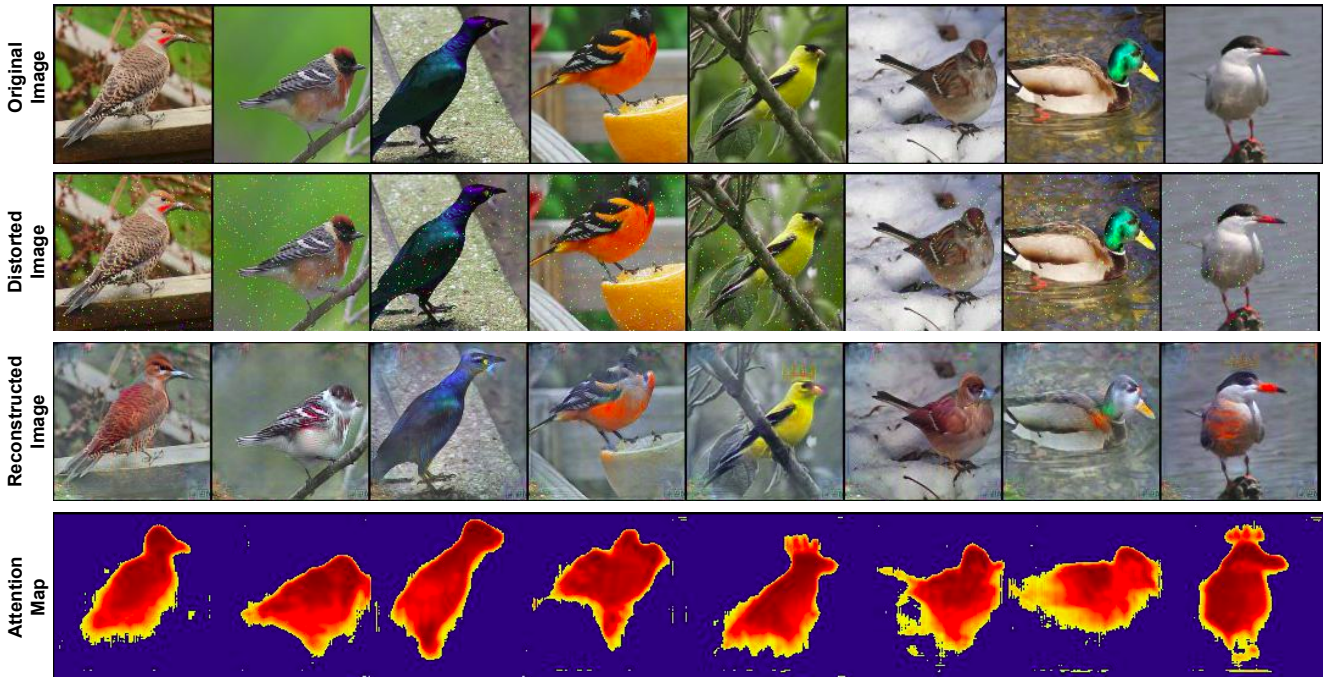


Figure 5. Foreground-Background (FG-BG) masks are generated using self-distillation on the masks. Mask Loss (ML) applied using masks from teacher network and impact on the visual quality of the mask generated when Random Augmentation (RA) is applied on the images provided to student network.

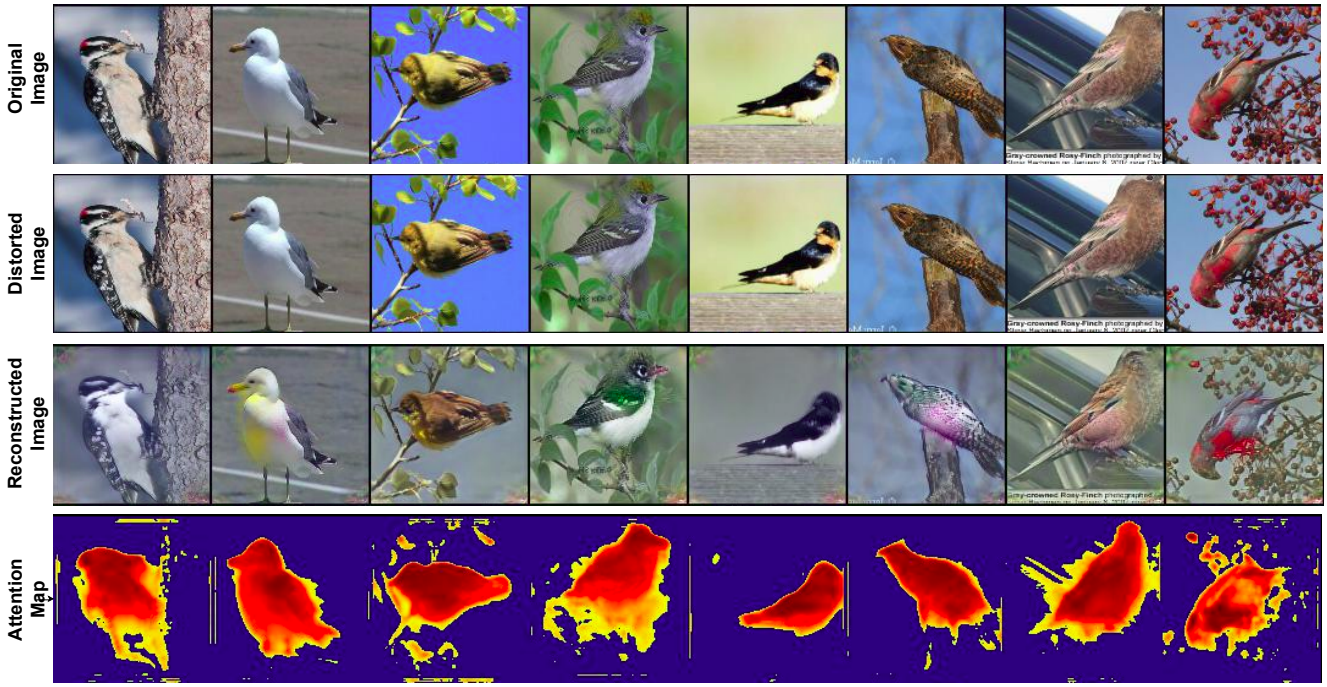


Gaussian Noise

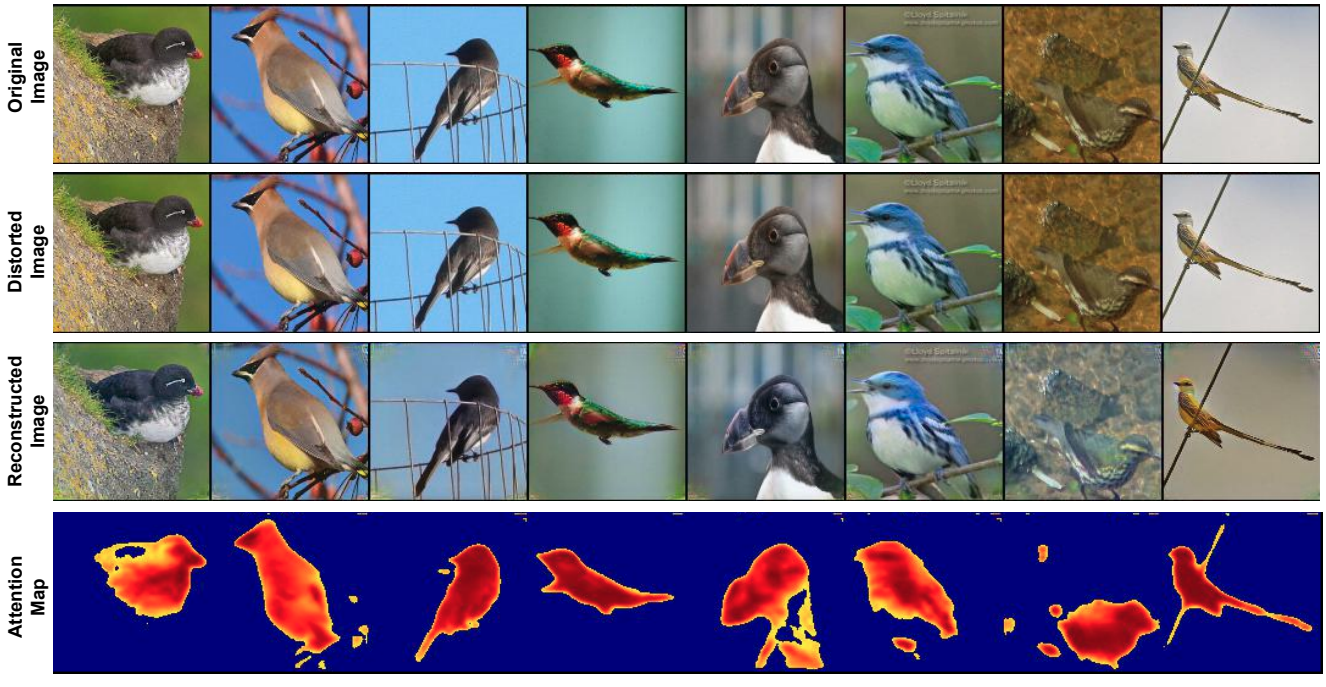


Salt and Pepper

Figure 6. On observing the outputs produced by our network, the impact of different image distortion approaches on generating inputs is evident. These outputs are obtained from inputs distorted using Gaussian Noise and Salt and Pepper techniques.



Colour Jitter



Diffusion Noise

Figure 7. On observing the outputs produced by our network, the impact of different image distortion approaches on generating inputs is evident. These outputs are obtained from inputs distorted using Colour Jitter and our proposed Diffusion based techniques.

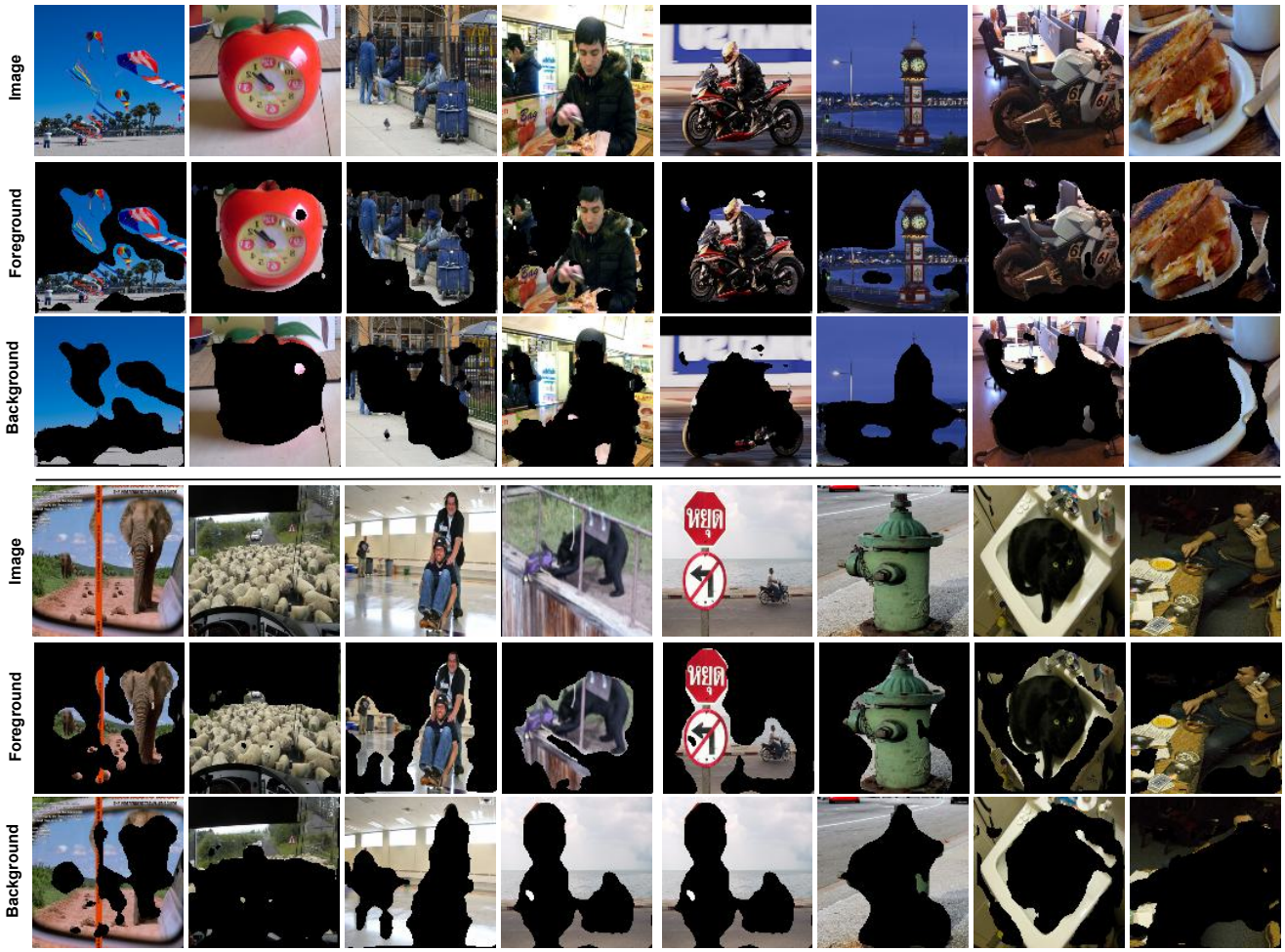


Figure 8. Foreground-Background Mask generated using our approach on COCO dataset [3].

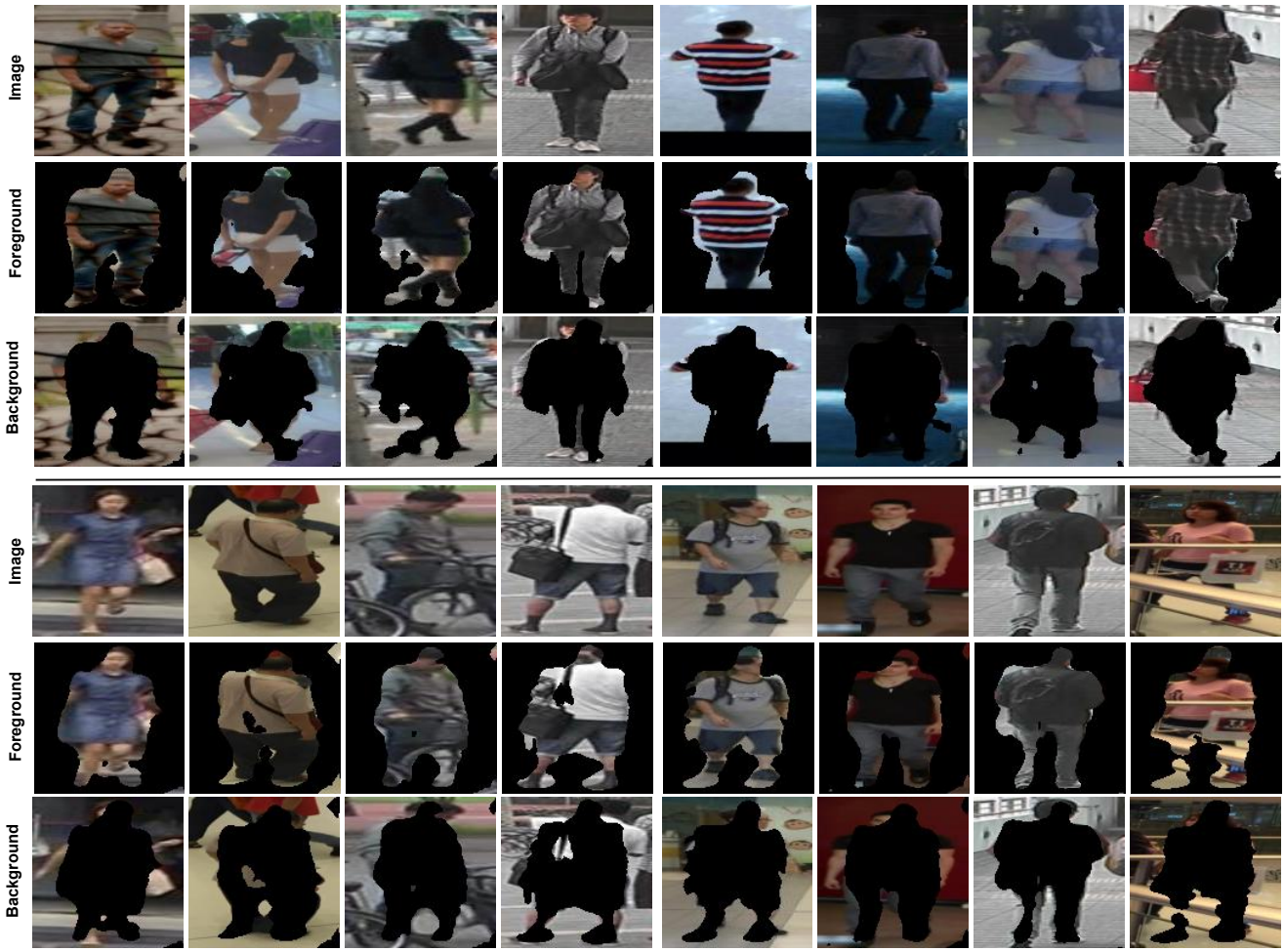


Figure 9. Foreground-Background Mask generated using our approach on CUHK-PEDES dataset [7].

References

[1] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 1

[2] 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. 2

[3] 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. 9

[4] 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, 3

[5] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 1

[6] 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

[7] Shengyu Zhang, Donghui Wang, Zhou Zhao, Siliang Tang, Di Xie, and Fei Wu. Mgd-gan: Text-to-pedestrian generation through multi-grained discrimination, 2020. 2, 10