Face Detection and Facial Attribute Editing

Mohammadamin Ahantab[1], Mohammadamin Barekatain[1],
Azade Farshad[1], and Yasaman Rajaee[1]

[1]Technical University of Munich

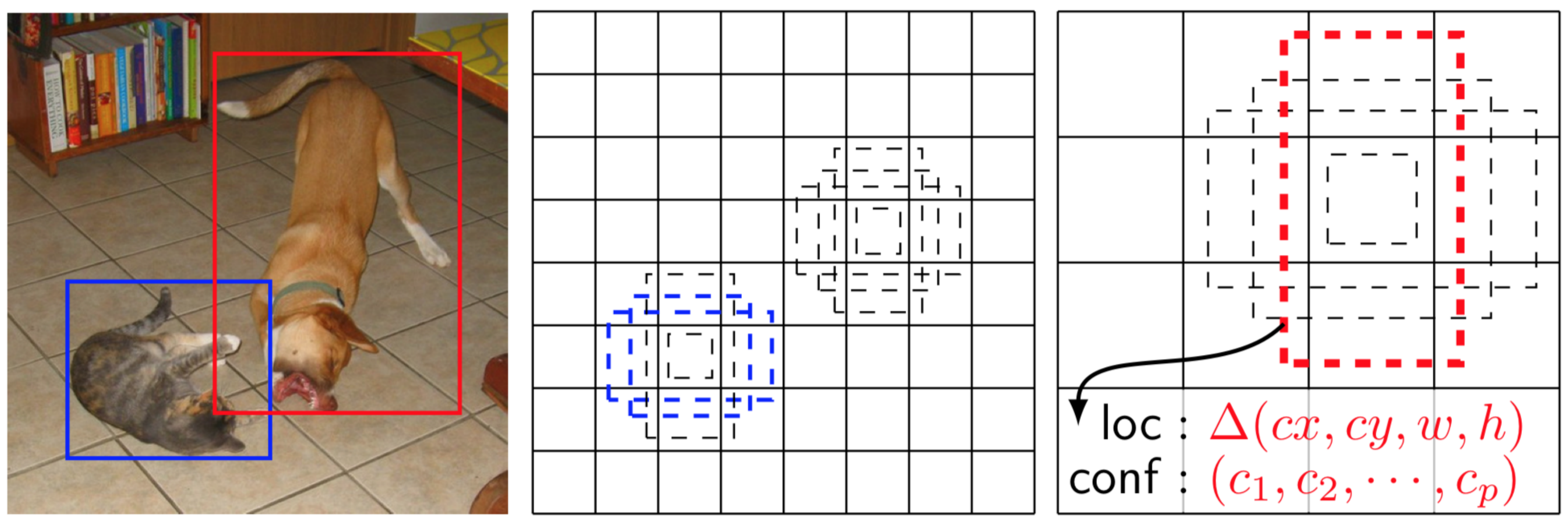
Access our code on Github! 😎

Overview

- It is practically infeasible to collect images with arbitrarily specified attributes for faces of each person.
- We are interested in the problem of manipulating natural images of faces by controlling some attributes of interest.
- Existing approaches take as an input a cropped face and then perform facial attribute editing task. However, in the real-world scenario, the input is not a cropped face.
- We developed a pipeline which first detects faces in the image and then changes the specified attributes for each of the detected faces.

Face Detection

To detect the faces we fine-tuned a Single Shot MultiBox Detector (SSD) with MobileNets as its base network. SSD, discretizes the output space of bounding boxes into a set of default boxes over different aspect ratios and scales per feature map location. At prediction time, the network generates scores for the presence of each object category in each default box and produces adjustments to the box to better match the object shape.


loc : $\Delta(cx, cy, w, h)$
conf : $(c_1, c_2, \cdots, c_p)$

(a) Image with GT boxes (b) 8×8 feature map (c) 4×4 feature map

Figure 1: SSD framework. (a) SSD only needs an input image and ground truth boxes for each object during training. In a convolutional fashion, SSD evaluates a small set (e.g. 4) of default boxes of different aspect ratios at each location in several feature maps with different scales (e.g. 8 x 8 and 4 x 4 in (b) and (c)). For each default box, SSD predicts both the shape offsets and the confidences.

Facial Attribute Editing

In this project we use StarGAN, an image to image translation generative model which is utilized to modify attributes of the detected face, such as hair color, gender and age. StarGAN is a conditional GAN which is trained to output target images from source images conditioned on target domain labels. The target and source images are fed to the **D** (Discriminator), which has the task to distinguish between the two domains. Three different loss objectives are used in StarGAN:

- Adversarial Loss: used to train **G** (Generator) and **D** for generating fake images which look like real images
- Domain classification loss: by minimizing this objective, **D** learns to classify a real image to its corresponding original domain.
- Reconstruction loss: $L1$ loss between the original image and reconstructed image from target to source.

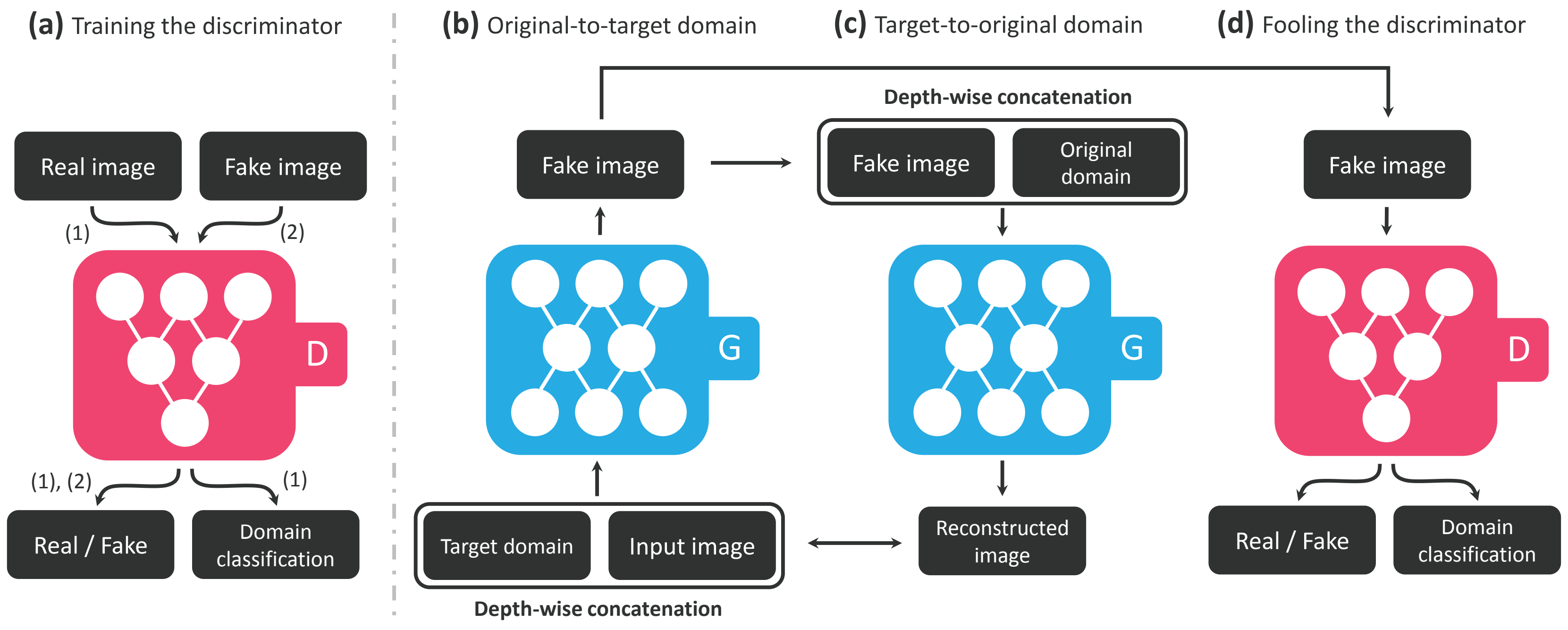


Figure 2: Overview of steps to train the StarGAN network

Approach

We propose the following pipeline:

1. **Face detection:** faces in the image are extracted using a fine-tuned SSD network.
2. Extracted face regions are extended by a factor of 1.5 to have a better context.
3. **Facial attribute editing:** extended face regions with their corresponding target attributes are fed to StarGAN to perform attribute editing.

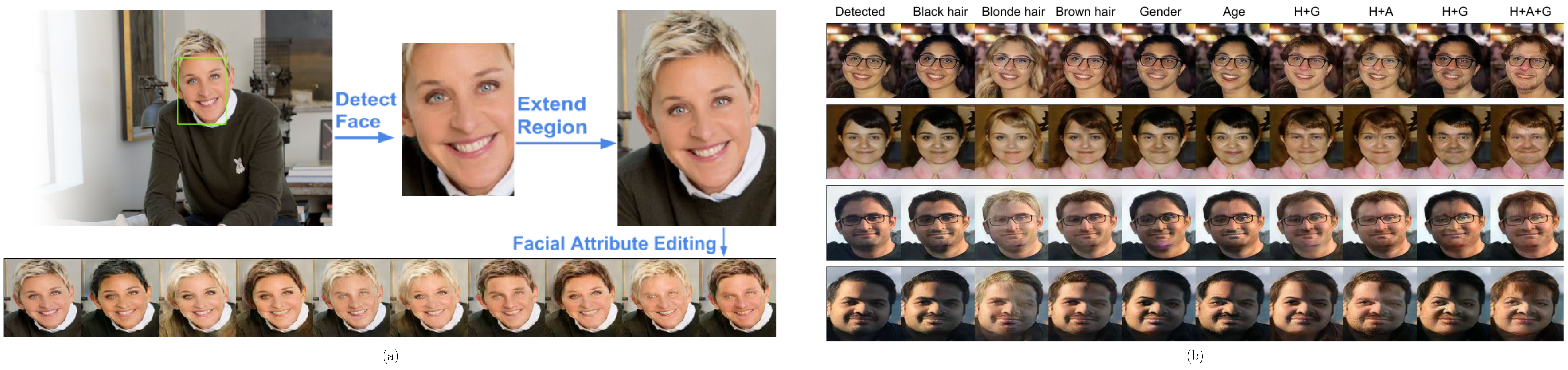
Experiments

Dataset For training StarGAN and SSD *CelebA* dataset is used, which consists of $\sim 200k$ images of celebrity faces with annotated attributes and position of the faces.

Training Table below shows the hyper-parameters used for training each part of the pipeline.

Model	Learning Rate	LR Decay Policy	Optimizer	Optimizer Parameters	Batch Size	Data Augmentation
SSD	0.004	Exponential	RMS-Prop	$decay = 0.95$, $\gamma = 0.9$, $\epsilon = 1.0$	24	random h-flip + crop
StarGAN	10^{-4}	Linear	Adam	$\beta_1 = 0.5$, $\beta_2 = 0.999$	16	random h-flip

Results Our trained SSD network achieved 87.12% average precision on CelebA validation set - we did not find any reported detection performance on this dataset in literature as it is mostly used for other tasks such as smile detection and gender recognition. Following figures illustrate some qualitative results of our pipeline.



(a) (b)

Figure 3: Our proposed face detection and facial attribute editing pipeline on **unseen** images. In (b) the original images are not shown due to the lack of space.