CS230

Unsupervised Semantic Face Synthesis

Jack Lee
jack9766@stanford.edu

1 Introduction

We propose a novel application of unsupervised image-to-image translation networks in the domain of semantic image synthesis. Given images sampled from marginal distribution $p(x_2)$ and semantic masks sampled from conditional distribution $p(x_1|x_2)$, our goal is to estimate the two conditionals $p(x_2|x_1)$ which is a many-to-one mapping and $p(x_1|x_2)$ which is a one-to-many mapping. We achieve this by training MUNIT [1] on an unpaired semantic mask to image dataset.

2 Related Works

2.1 Generative Adversarial Networks

The GAN framework [2] is commonly used to align generated images to the target domain. This is achieved by training a generator to fool a discriminator whose is trying to distinguish between real and generated images.

2.2 Semantic Image Synthesis

Park *et al.* [3] proposed spatially-adaptive normalization and GauGAN achieved high-fidelity semantic image synthesis using paired datasets; outperforming state-of-the-art supervised image-to-image translation networks such as pix2pixHD [4].

2.3 Unsupervised Image-to-Image Translation

Various constraints and assumption can be enforced during training to achieve image-to-image translation with an unpaired dataset. CycleGAN [5] enforces cycle consistency loss where the same images must be generated after being translated to the target domain and back. BicycleGAN [6] further improved output diversity by enforcing bijective consistency between latent encoding and output modes. UNIT [7] assumed corresponding images from different domains can be encoded into and decoded from the same latent space. MUNIT generated multi-modal outputs by assuming images can be encoded into content and style encoding pairs and then reconstructed or translated by decoding the same or different encoding pairs.

3 Dataset

We use CelebAMaskHQ [8] as our training dataset. The dataset consists of 30000 paired semantic masks and images where each image corresponds to up to 19 different classes of semantic masks (eyes, brow, ears, etc). The dataset is preprocessed by normalizing the images to a mean of 0.5 and standard deviation of 0.5, fusing semantic maskes of different classes into a 19 channel one-hot tensor; all inputs are resized to a resolution of 256 by 256.

CS230: Deep Learning, Winter 2018, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

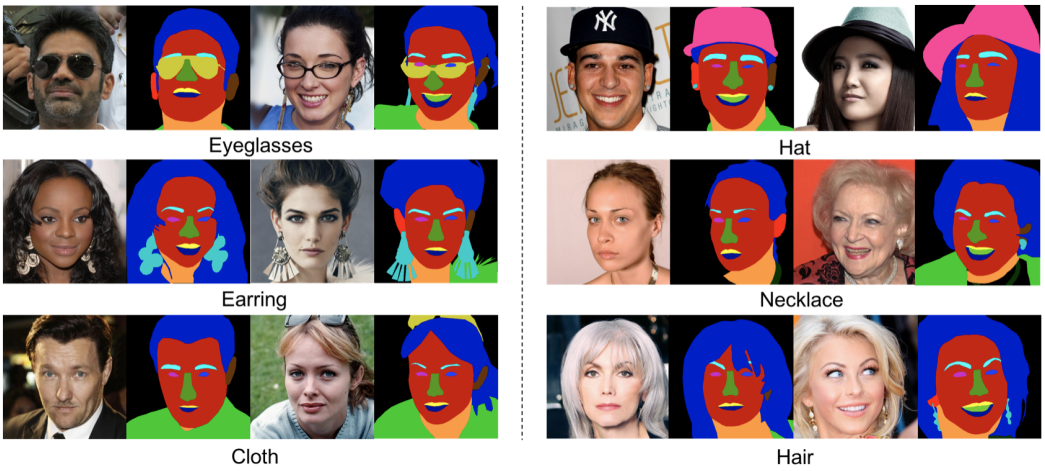


Figure 1: Samples from the CelebAMaskHQ dataset.

4 Approach

Given an unpaired mask-to-face dataset, let $x_1 \in X_1$ be a semantic mask sample and $x_2 \in X_2$ be a face sample. We make a partially shared latent space assumption where the corresponding x_1 and x_2 can be generated from a shared a content encoding $c \in C$ and distinct style encodings $s_1 \in S_1$ and $s_2 \in S_2$. Given the above assumption, we can utilize the MUNIT architecture to achieve unsupervised image translation. The architecture consists of a pair of autoencoders:

$$\hat{x}_1 = G_1(E_1^c(x_1), E_1^s(x_1))$$
$$\hat{x}_2 = G_2(E_2^c(x_2), E_2^s(x_2))$$

To perform cross domain translation we rearrange the setup as shown below where s_1 and s_2 are style encodings sampled from a guassian prior:

$$x_{1\to2} = G_2(E_1^c(x_1), s_2)$$
$$x_{2\to1} = G_1(E_2^c(x_2), s_1)$$

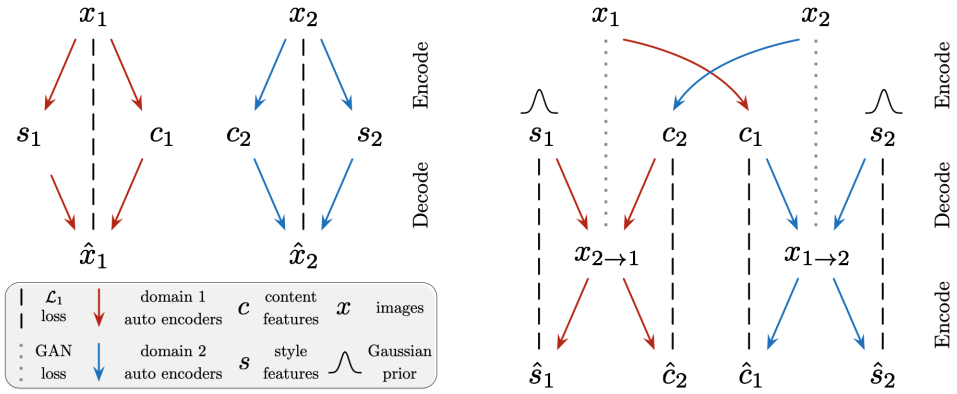
To train the autoencoders we enforce image reconstruction losses denoted as dashed lines in Figure 2 part (a) and latent reconstruction losses denoted as dashed lines in Figure 2 part (b):

$$\mathcal{L}_{recon}^{x_1} = \|\hat{x}_1 - x_1\|_1$$
$$\mathcal{L}_{recon}^{x_2} = \|\hat{x}_2 - x_2\|_1$$
$$\mathcal{L}_{recon}^{c_1} = \|E_2^c(x_{1\to2}) - c_1\|_1$$
$$\mathcal{L}_{recon}^{c_2} = \|E_1^c(x_{2\to1}) - c_2\|_1$$
$$\mathcal{L}_{recon}^{s_1} = \|E_1^s(x_{2\to1}) - s_1\|_1$$
$$\mathcal{L}_{recon}^{s_2} = \|E_2^s(x_{1\to2}) - s_2\|_1$$

To align the translated images with the target domain we introduce discriminators $D_1(x_{2\to1})$ and $D_2(x_{1\to2})$ for each domain; and, enforce adversarial losses $\mathcal{L}_{GAN}^{x_1}$ and $\mathcal{L}_{GAN}^{x_2}$ denoted as dotted line in Figure 2 part (b). In this case, we used MSE loss for adversarial loss.

The total loss is denoted is shown below where $\lambda_x, \lambda_c, \lambda_s$ are scalar weights controlling the importance of each reconstruction terms:

$$\min_{E_1,E_2,G_1,G_2} \max_{D_1,D_2} L(E_1, E_2, G_1, G_2, D_1, D_2) = \mathcal{L}_{GAN}^{x_1} + \mathcal{L}_{GAN}^{x_2} + \lambda_x(\mathcal{L}_{recon}^{x_1} + \mathcal{L}_{recon}^{x_2}) + \lambda_c(\mathcal{L}_{recon}^{c_1} + \mathcal{L}_{recon}^{c_2}) + \lambda_s(\mathcal{L}_{recon}^{s_1} + \mathcal{L}_{recon}^{s_2})$$



(a) Within-domain reconstruction

(b) Cross-domain translation

Figure 2: Model overview.

	pix2pix	MUNIT
FID	67.99	39.49

Table 1: FID scores.

5 Results

We trained a pix2pix model and MUNIT model on the CelebAMaskHQ dataset using a NVIDIA Tesla V100 GPU for 8 epochs and 70 epochs respectively (20 hours each). We observed the quality of pix2pix outputs remained similar after the first epoch while the quality of MUNIT outputs fluctuated drastically and is still improving at the end of the training session.

We selected the best checkpoints for each model measured by their FID scores [9]; the images generated using those checkpoints are shown in Figure 3. We observed MUNIT producing images with higher diversity and more realistic textures of hair and skin compared to pix2pix's smoothed out and unimodal outputs. However, there are obvious flaws in the MUNIT outputs such as hats having the same texture and color as hair and badly formed glasses. In addition, MUNIT seems to be more robust to input noises; the semantic mask on the second last row of Figure 3 is slightly mislabelled (black bar across the chin), the pix2pix model generated abnormal noise around the mislabeled area whereas MUNIT largely ignored this anomaly.

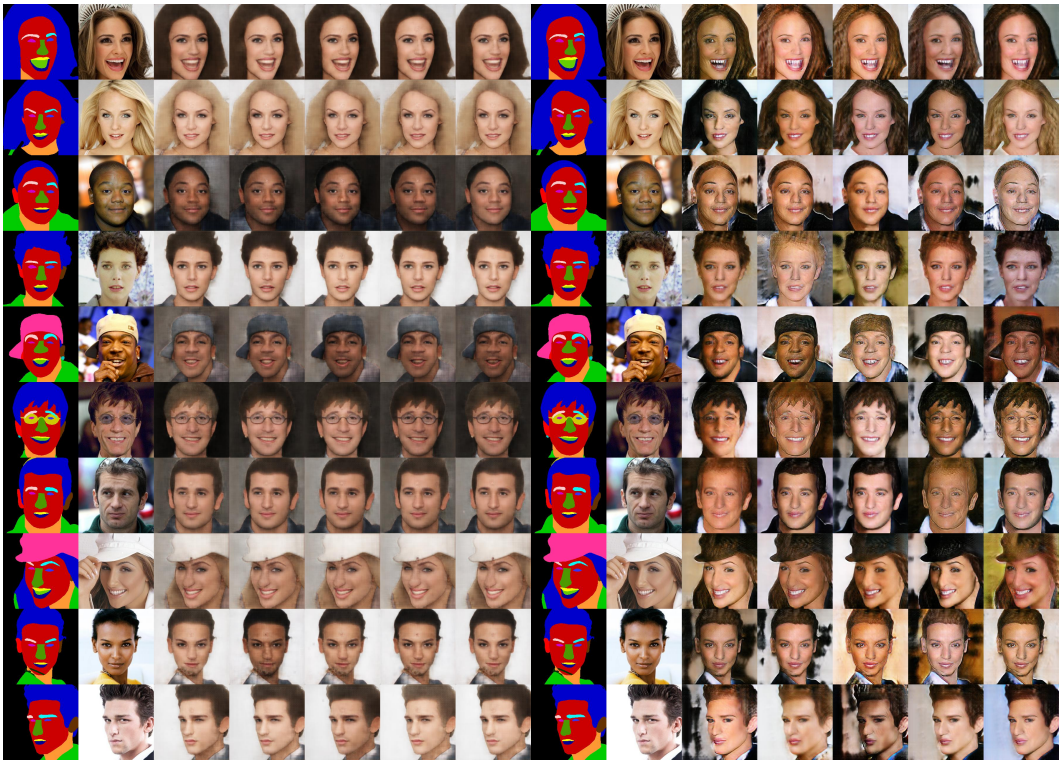
We measured the FID scores of the models' outputs by randomly sampling 10000 semantic masks and 10000 face images in an unpaired fashion; we then measure the distributional distance between the of 10000 images generated from the 10000 semantic masks and 10000 real images using the FID metric. As shown in table 1, MUNIT's multimodal outputs matched the ground truth distribution much closely compared to pix2pix's outputs.

6 Conclusions

In this paper, we presented a novel application of unsupervised image translation in the domain of semantic face synthesis and compared the results both qualitatively and quantitatively with state-of-the-art supervised image translation techniques. Future work includes iterating on the MUNIT architecture and hyperparameters to improve its performance in the domain of semantic face synthesis.

7 Contributions

This project is completed by Jack Lee under the supervision of Sharon Zhou.



(a) pix2pix outputs (b) MUNIT outputs

Figure 3: Images generated by pix2pix (a) and images generated by MUNIT (b). The first column consists of semantic masks, second column consists of real images and the rest are generated images.

References

[1] Xun Huang, Ming-Yu Liu, Serge J. Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. *CoRR*, abs/1804.04732, 2018. URL `http://arxiv.org/abs/1804.04732`.

[2] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[3] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. *CoRR*, abs/1903.07291, 2019. URL `http://arxiv.org/abs/1903.07291`.

[4] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. *CoRR*, abs/1711.11585, 2017. URL `http://arxiv.org/abs/1711.11585`.

[5] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *CoRR*, abs/1703.10593, 2017. URL `http://arxiv.org/abs/1703.10593`.

[6] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A. Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. *CoRR*, abs/1711.11586, 2017. URL `http://arxiv.org/abs/1711.11586`.

[7] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. *CoRR*, abs/1703.00848, 2017. URL `http://arxiv.org/abs/1703.00848`.

[8] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[9] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Günter Klambauer, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a nash equilibrium. *CoRR*, abs/1706.08500, 2017. URL `http://arxiv.org/abs/1706.08500`.