Art and Machine Learning | Prof. Eunsu Kang
CMU Spring 2019
Final Project

# Capricious Portraits

Marisa Lu
Mira Murali
Loki Ravi
Izzy Stephen
Yuhan Xiao

# Concept

Artwork lasts hundreds of years, but identity is ephemeral. Traditionally, painters and illustrators used figuration as a conduit for their own consciousness, and a reflection of the viewer's. This reflects the deep-rooted human desire to see one's own reality reflected as accurately as possible in art.

Despite this, the figure has always been idealized. From the distorted proportions of Mannerist painting to the happy housewives of mid-century advertising, representations of the figure are more likely to reveal cultural beliefs than truths about human consciousness. This discrepancy between art and reality points towards a deeper significance in every human face an artist renders, a 'hidden layer' beneath the painted surface. The artwork and advertising we selected as testing images for our model was chosen based on what it revealed about cultural ideals, like beauty standards or the glorification of consumerism.

Just as a painting disintegrates over time, the human psyche disintegrates as well due to age or mental illness. In the era of AI, which is conceptualized as immortal and unfailingly logical, the inherent superiority of human selfhood is being called into question. Inspired by SPADE (Park et al) and Fine Style Transfer (Kerras et al), we exploit AI's ability to iterate endlessly on a human face to manipulate culturally significant imagery.

# Literature Review

## *Pix2Pix in Art*

Since Pix2Pix was released to the world in 2016, artists working with machine learning tools have used it extensively. Pix2Pix projects range from the lighthearted and comical to the serious and political. One early implementation, Christopher Hesse's innovative Edges2Cats project, takes in a line drawing (series of edges) of any form, and generates an odd, squashed yet convincingly photorealistic cat in that shape.

Finnish artist Memo Akten works with Pix2Pix to explore philosophical questions. He sees making art with machine learning as a way to "reflect on how we make sense of the world" (Artnome), proving that jumping to conclusions about an object's potential is a limitation of human perception that need not necessarily be built into AI. Akten's video Learning to See demonstrates a Pix2Pix model trained on beautiful landscape images, and tested on mundane desktop scenes involving cords, office supplies, and other everyday items. As the artist notes in the caption, "It can only see what it already knows, just like us." Our perception of daily life is limited by what we know and what we've experienced. Certain settings may always incite boredom and despair, but as Akten's video shows, there is potential to find beauty and inspiration even in the winding shape of an electrical cord.

In ML artist Gene Kogan's project Invisible Cities, Kogan and his collaborators used Pix2Pix to create a series of photorealistic satellite maps generated from the input of a multicolored graphic denoting buildings, bodies of water, roads, and the like. Though the training data relied on graphics paired with the corresponding real-life satellite imagery, the final result could take in any sketch using the designated colors and create a convincing 'map' of this imagined landscape. Kogan cites a passage from Italo Calvino as inspiration: "Cities, like dreams, are made of desires and fears… everything conceals something else". Cities, accumulations of human dwellings and infrastructure, encode information about how humans live, and the highly regulated architecture of modern cities is reflected back in Kogan's imaginary ones.

Another project by Kogan, informally known as Meat Puppet, pointedly questions the phenomenon of fake news and political figureheads by applying Pix2Pix to video input from a webcam. Specifically, Kogan trained a model on footage of Donald Trump, and applied it to webcam input of himself as he made silly faces, generating a shockingly lifelike, somewhat grotesque video of Trump smiling, laughing, and grimacing in time. As a predecessor to controversial 'deepfake' videos, Meat Puppet shows how crucial it will soon become for media consumers to understand the power of machine learning, and identify works made with it, 'artistic' or otherwise.

### *Image Segmentation*

Conventional semantic segmentation methods make use of simple image processing techniques such as edge detection, mean shift filtering and region growing. With the birth of AI, convolutional neural networks (CNNs) outperformed traditional image processing techniques. Some semantic segmentation networks include SegNet (an encoder-decoder architecture for pixel-wise segmentation), Fully Convolutional Networks (FCNs) and U-Net, an extension of the FCN originally developed for biomedical image segmentation. Due to the lack of large datasets in the medical field, U-Net compensates for this by data augmentation and works surprisingly well with small datasets.

### *Image Synthesis from Semantic Segmentation Masks*

Generating realistic images given just a simple semantic layout of a scene has been gaining a lot of popularity lately due to its application in many areas of image processing, which includes content generation and image editing. However, most deep learning methods use CNNs which consist of normalization layers that tend to degrade the semantic meaning present in the layout (Pix2Pix, cascade refinement networks are some examples). Hence, in this project we use one of the state-of-the-art algorithms, SPatially Adapative DEnormalization (SPADE) to mitigate this effect and allow semantic information to propagate across layers.

# Technique

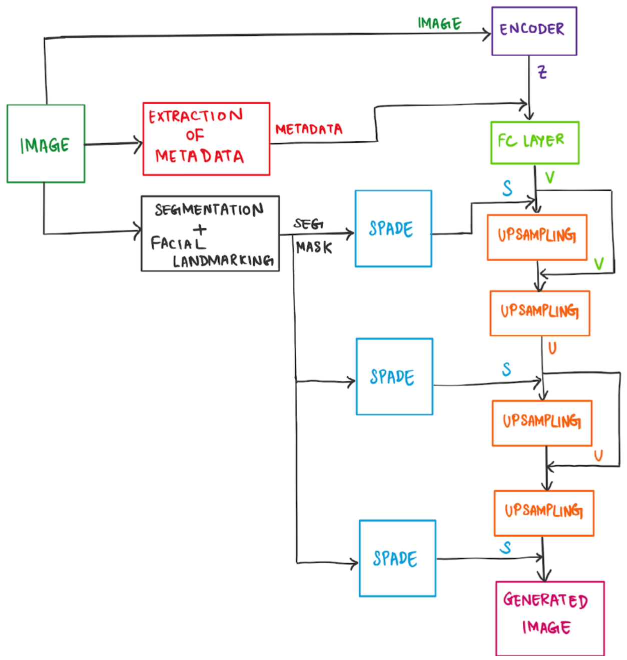

*Figure 1: Training Pipeline*

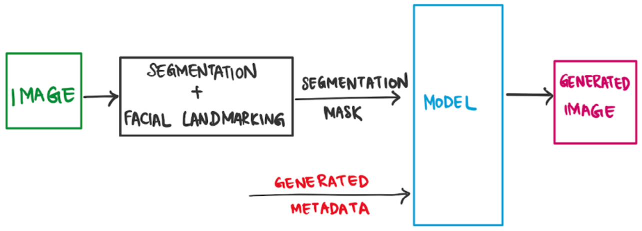

*Figure 2: Testing Pipeline*

The goal of this project is to be able to control the generation process in the GAN by conditioning on only the global features. Building on our previous work in the GAN assignment, we discovered that fine grained control of local features such as the color of the eyes or lips requires decorrelation of the feature space over which the generation is conditioned. This is because when conditioned on the entire input image or some representation of it, the GAN essentially learns the conditional probability of the generated image, given the joint probability of the input features. So changing a single input feature such as hair color would result in a cascade of changes unintended. Since decorrelating this input space requires a complex encoder, this is beyond the scope of our work. Instead we hope to obtain coarse grained control of the GAN by conditioning the GAN as a whole on the global features that every upsampling layer requires, such as age. The local feature generation is also conditioned on the semantic segmentation labels that are passed in to every upsampling layer. By doing this, we have effectively, separated the feature space affecting only local feature generation such as size and color of eyes, from the global features that affect all the layers such as the age which affects the texture of skin as well as the texture of hair.

Our basic pipeline for training and testing is depicted in Figures 1 and 2 respectively. During training, we take the input image of a face and pass it through our segmentation model to extract a segmentation mask, which segments the face, hair and background as three different classes. Then, we use facial landmarking to detect the points on the face that correspond to the certain facial parts such as the eyes, eyebrows, nose and mouth. We use these points to draw bounding boxes over the aforementioned facial parts and form our final segmentation mask, which consists of seven classes: eyes, eyebrows, nose, mouth, hair, skin, and background. Here, we explicitly avoid tight freeform masks to allow the model some degree of freedom to condition the exact shape of local features on global metadata.

Next, we extract global metadata using Microsoft Face API. This metadata consisted of the age, the gender and the presence/absence of glasses in the given image of a face. To allow for the semantic information across the model, we use the spatially adaptive normalization algorithm, which learns modulation parameters to normalize the activation parameters in the previous layer, thereby allowing the semantic information to pass undeterred through the current layer.

This normalization method, abbreviated as SPADE, appends the segmentation mask to each layer in the model, thus getting rid of the encoder of the pix2pix model. However, since semantic maps are color agnostic with respect to the color of a person's hair/eyes we need an additional style encoder to understand color. It also allows us to concatenate the metadata (age, gender, etc.) to the latent vector we get from the output of the encoder before passing it through a fully connected layer and eventually the generator. Thus the deep learning model used here consists of a Style Encoder, the Upsampling portion of U-Net with SPADE for normalization. During testing, we created our own metadata information which was passed along with the segmentation mask to the model, allowing us to control the age of the person.

*Dataset*

Our data is comprised of the Flickr Faces High Quality (FFHQ) dataset which consists of 70,000 high quality images of faces.  The AffectNet dataset, which includes 420,000 images annotated categorically by emotion displayed (i.e. happy, sad, angry, surprise, etc.) and by intensity of emotion was used initially to train the model to understand facial expressions. We trained on 128x128 images to produce images of the same dimensions. This was majorly due to the lack of training resources. We believe the method can be applied to 1024x1024 images as well.

*Model*

The generator of our model consists of a Style Encoder and the upsampling half of the U-Net. The output of the Style Encoder is constrained to be 128 dimensional latent vector carrying only the essential information from the input image such as the color of the skin and hair. The latent vector concatenated with the global input features are passed through a Dense network with a ReLU activation that converts the (128+G) dimensional vector into a 128 dimensional vector. This serves as input to the U-Net Upsampling Generator with 4 upsampling blocks. These blocks each contain a Residual connection popular in ResNet models and their output is normalized using SPADE which utilizes semantic segmentation labels. The output from this generator is a 128x128 image.

In order to train this generator we employed 2 discriminators based on the patchGAN discriminator used in Pix2Pix. We call these multi-scale discriminators since they classify the images based on “patches” of increasing size.

The loss consisted of the traditional GAN loss, as well as the loss from 2 discriminators. Additionally we also used a KL Divergence loss to train the Style Encoder. In order to improve the overall quality of the image, we used a pretrained VGG network to implement the traditional perception loss.

*Training*

We trained the network on 128x128 images for 22 epochs. The training time was nearly 2.5 hours per epoch. We noticed that when the training process was destabilizing, updating the generator for every 2 updates of the discriminators helped establish a better baseline for the generator to follow. We were unable to train for more than 22 epochs due to resource constraints. However, we do recommend the model be trained for at least 50-75 epochs for good quality results.

*Testing*

During testing, we passed in the segmentation masks of the image and created our own metadata to observe how the output image varied with respect to the metadata attributes. Thus, we were able to control the generated image by varying the attributes we sent in as additional information.

# Method

*Segmentation and Landmarking*

We use a combination of the MobileNetV2 and the U-Net architecture to segment the images into three categories: face, hair and background. We then use DLib's pretrained facial landmark detector to draw bounding boxes around specific parts of the face, such as the eyes, nose, eyebrows and mouth. The final segmented image now constitutes our segmentation mask, which is passed through the SPADE algorithm built on the Pix2Pix Model. This is shown in Fig. 3

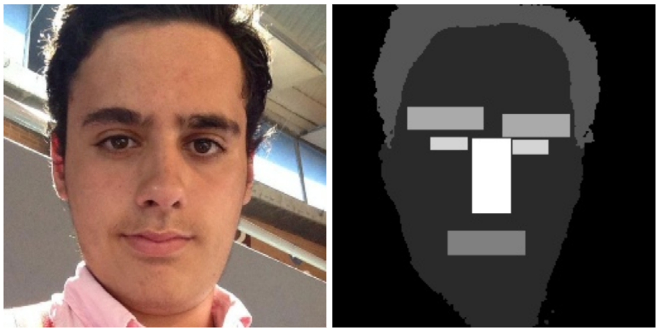

Fig. 3: Original Image (left) and the final segmentation mask (right)

*SPADE*

As mentioned before, traditional convolutional neural networks used for image synthesis make use of normalization which results in loss of semantic information. The spatially adaptive normalization trick only applies normalization to activation in the previous layer, thereby allowing the flow of semantic information across layers.

*Metadata*

We extracted some metadata from the images, thanks to the Microsoft Face API. From this we handpicked some global features such as "Glasses?", "Age", "Gender" and concatenated these along with the latent vector from the encoder and fed it to the Generator network. This allowed the Generator to be conditioned on the joint distribution of the metadata.

# Results

The following results (Figure 4 and Figure 5) depict the variation in the generated image based on the variation in the metadata passed through the trained model. By specifying the age and whether or not the the resultant face should have spectacles through the metadata, we have some control over the generated image.

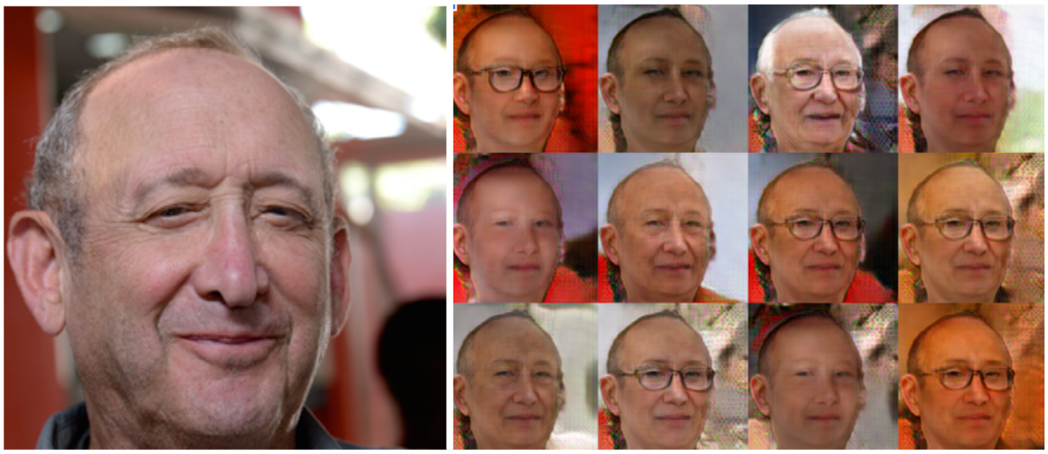

Fig 4. Input image (left) and output images (right)

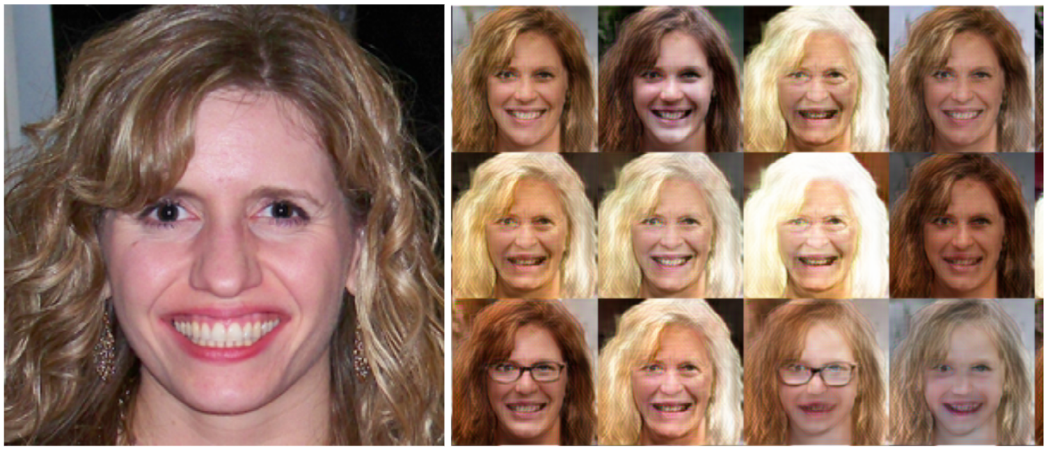

Fig 5. Input image (left) and output images (right)

# Code

https://github.com/mira-murali/emote-ai

# Contributions

Artistic concept, poster preparation: Izzy Stephen
Video documentation: Marisa Lu, Izzy Stephen
Result curation and presentation: Marisa Lu, Yuhan Xiao
Metadata extraction, model architecture: Loki Ravi
Image segmentation, facial landmarking: Mira Murali
Training and testing: Loki Ravi, Mira Murali
Report preparation: All

# References

**AffectNet:** Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor, "AffectNet: A New Database for Facial Expression, Valence, and Arousal Computation in the Wild", IEEE Transactions on Affective Computing, 2017.

**FFHQ:** Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. CoRR, abs/1812.04948, 2018b. URL http://arxiv.org/abs/1812.04948.

**Fine Style Transfer:** Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. CoRR, abs/1812.04948, 2018b. URL http://arxiv.org/abs/1812. 04948.

**Pix2Pix:** Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2016). Image-to-image translation with conditional adversarial networks. arXiv preprint arXiv:1611.07004 .

**SPADE:** Park, T., Liu, M. Y., Wang, T. C., & Zhu, J. Y. (2019). Semantic Image Synthesis with Spatially-Adaptive Normalization. arXiv preprint arXiv:1903.07291.

**MobileNetV2:** Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4510-4520).

**U-Net:** Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham.