End to End 3D Face Model Synthesis Using Textual Descriptions

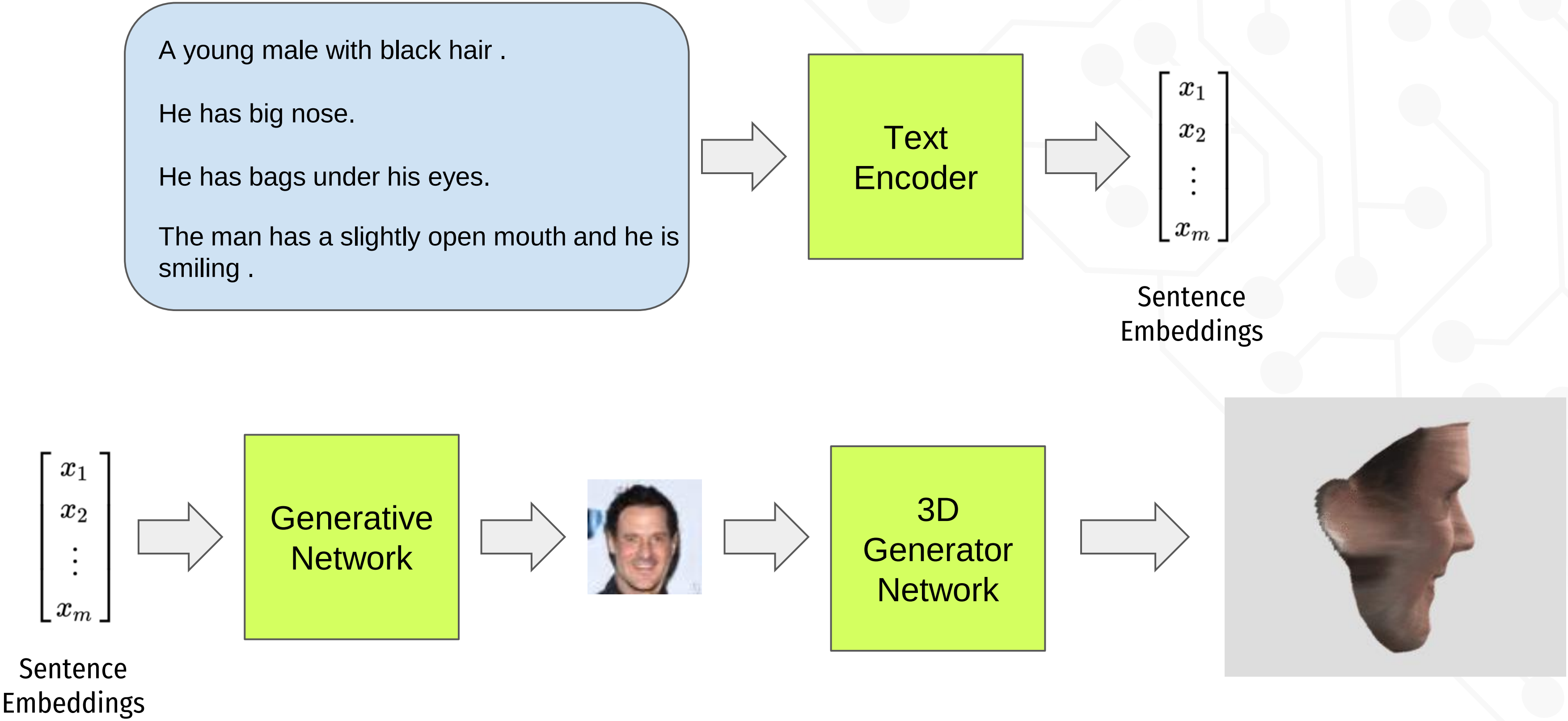
Mehmet Uluç Şahin

Advisor: Asst. Prof. M. Furkan Kıraç

Department of Computer Science
Özyeğin University

İstanbul, June 10th, 2020

A Quick Demo


A young male with black hair .
He has big nose.
He has bags under his eyes.
The man has a slightly open mouth and he is smiling .
Text Encoder
$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}$
Sentence Embeddings
$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}$
Sentence Embeddings
Generative Network
3D Generator Network

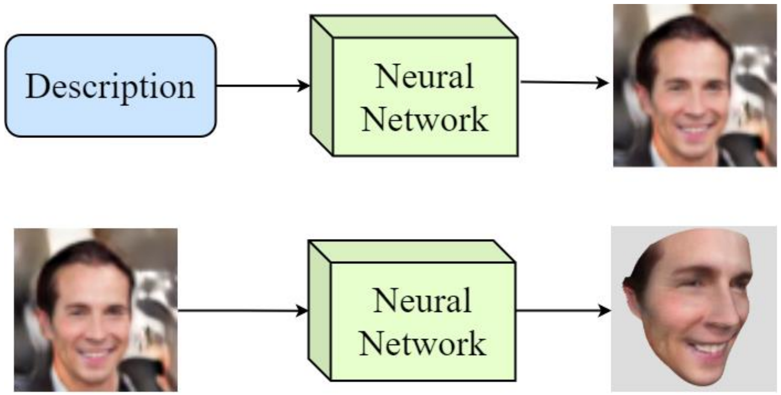
Outline

- Introduction
- Background
    - Generative Adversarial Networks (GANs) and Conditional GANs
    - Word Embeddings as Conditioning Variable
    - StyleGAN
    - Position Map Regression Network (PRN)
    - Learned Perceptual Image Patch Similarity (LPIPS)
- Proposed Methods
- Experiments
- Results
- Conclusion

Introduction

In this thesis:

- We improve currently available Text to Face results.
- We propose a measure for evaluating generative models.
- We are providing end to end system for generating 3D face models from given textual descriptions



Introduction

Goal of this thesis:

- Providing realistic 3D face models from textual description.
- Combining advantages of Conditional GANs with StyleGAN architecture on human face generation task.
- Exploring GAN evaluation measures.

Introduction

Why?

- Visualising a human face important applications
- 3D modeling is difficult, time consuming and expensive
- 3D models can transmit more information
- Automating this process can be very impactful

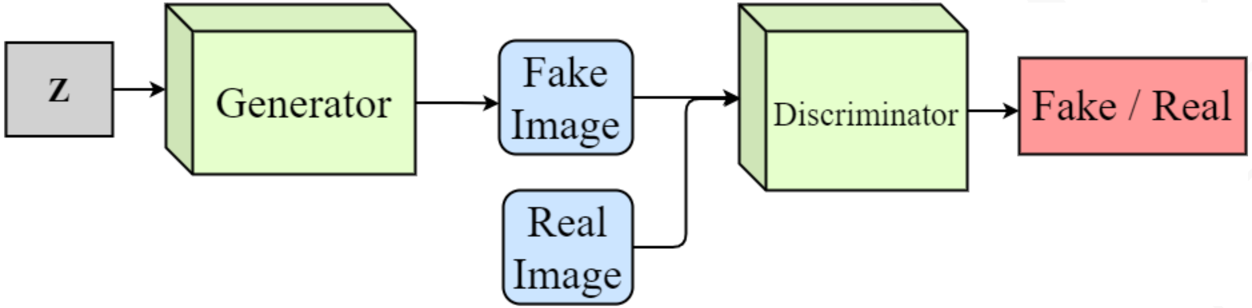
*Image Taken From: https://www.chron.com/news/houston-texas/houston/article/Police-sketch-vs-mugshots-How-accurate-are-they-11297369.php

Introduction

- Generative NNs can generate realistic results.
- With the development of GANs, results are getting better, StyleGAN results that are mostly indistinguishable from real images.
- However, controlling the output is a difficult task.
- Conditional GANs aim to solve this problem.

Background - GANs

- GANs are composed of two main components: **generator** G and **discriminator** D.

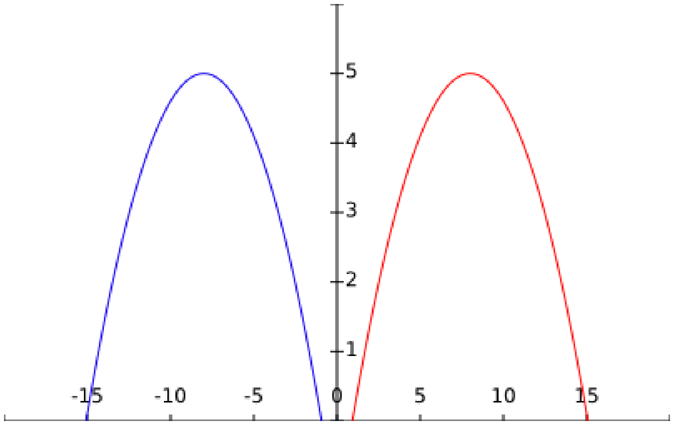


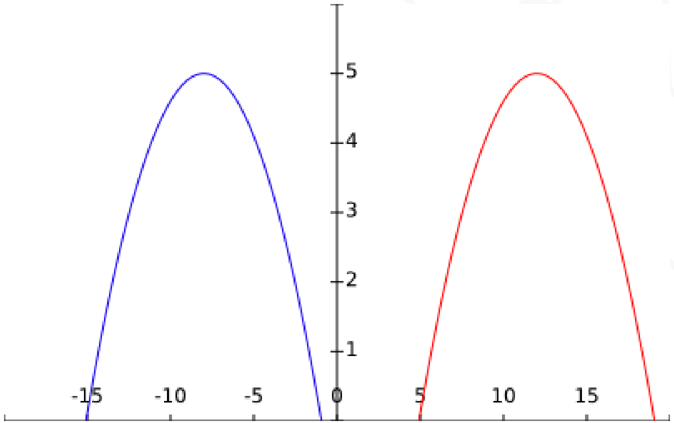
$$\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}[logD(x)] + E_{z\sim p_z(z)}[log(1 - D(G(z)))]$$

Background - GAN Loss Function

Wasserstein GANs:

- Naive GANs use **Jensen Shannon Divergence** (JSD) as loss function, which is based on **Kullback-Leibler Divergence**.
- JSD falls short in taking the distance between two probability distributions into consideration if there is no overlapping (log2).
- **Wasserstein Distance** works well even with non-overlapping probability distributions.

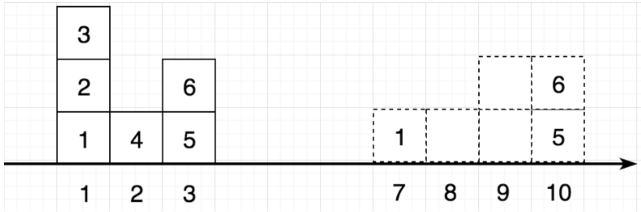




Background - GAN Loss Function

Wasserstein Distance:

- Also called Earth Mover's Distance
- Defined as minimum work required to convert one distribution into another

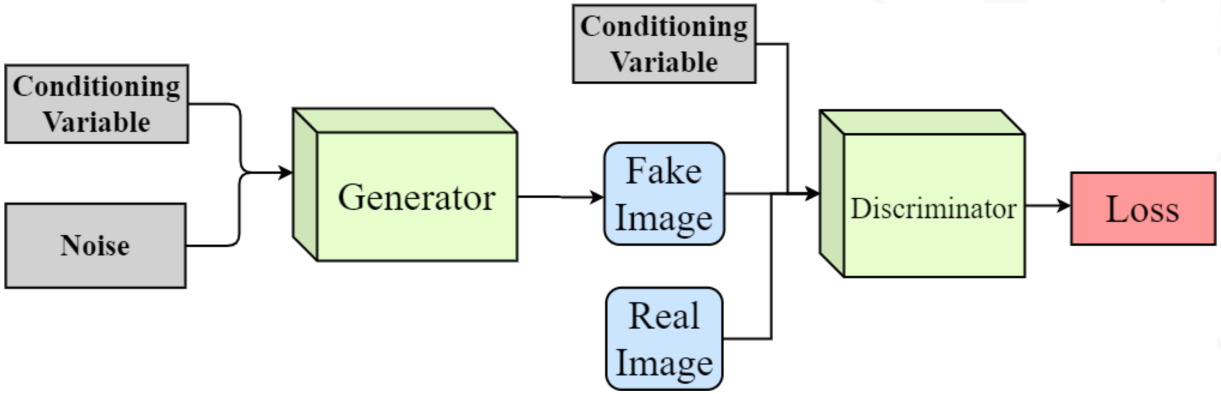


Wasserstein Distance is defined as:

$$W(\mathbb{P}_r, \mathbb{P}_g) = \inf_{\gamma \in \Pi(\mathbb{P}_r, \mathbb{P}_g)} \mathbb{E}_{(x,y)\sim\gamma}\big[\, \|x - y\| \,\big]$$

Background - Conditional GANs

- We cannot control output of Naive GANs.
- Conditional GANs (cGANs) allows controlling output with Conditioning Variable (CV).



$$\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}[logD(x|y)] + E_{z\sim p_z(z)}[log(1 - D(G(z|y)))]$$

Background - Conditional GANs

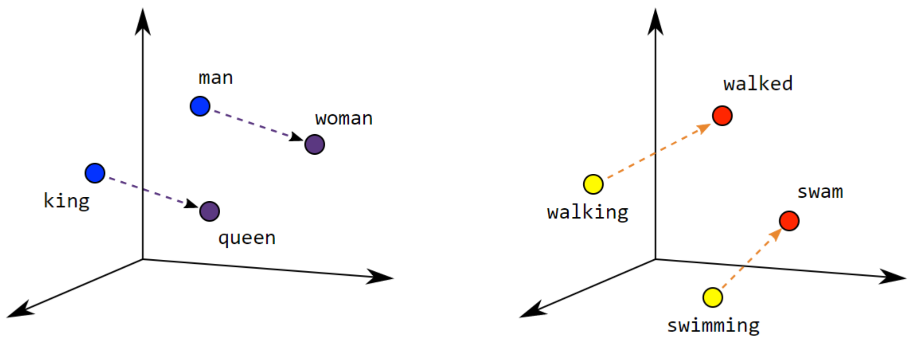
GAN-CLS Loss:

- In addition to input with correct CV, an incorrect CV is given.
- Discriminator learns to penalize if generated samples are not matching with CV.
- Discriminator should classify samples with wrong CV as fake: increased loss for generator
- This type of discriminators are called **Matching-aware Discriminator**.

Background - Word Embeddings

Word Embeddings:

- We need numerical representation for words.
- Mapping of words in vector space that contain information about word.
- Words with similar meanings have similar representations.



* Image taken from: https://developers.google.com/machine-learning/crash-course/images/linear-relationships.svg

Background - Word Embeddings

SkipGram Model:

- For each word, network also takes words surrounding the current word in a range.
- Network can predict probability of a word appearing in the window of word that is selected.
- We are interested in **values in middle layers**, which we call **embeddings**

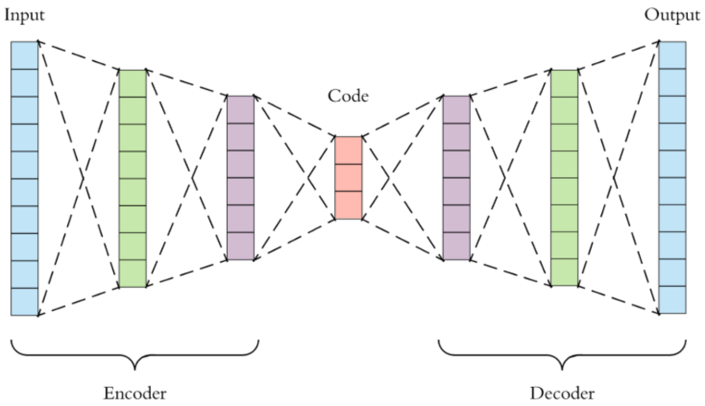


*Image taken from: https://towardsdatascience.com/applied-deep-learning-part-3-autoencoders-1c083af4d798

Background - Word Embeddings

- SkipGram Minimizes the average negative log likelihood:

$$-\frac{1}{T}\sum_{i=1}^{T}\sum_{-m\leq j\leq m, j\neq 0} log(P(w_{i+j}|w_i))$$

- Probability of a word **o** (outside word) given a center word **c:**

$$P(o|c) = \frac{e^{u_o^T v_c}}{\sum_{w\in V} e^{u_w^T v_c}}$$

Background - Word Embeddings

FastText:

- Used in this thesis for its ability to compute embeddings for unknown words.
- Computes embeddings for each character n-gram using SkipGram method.
- Allows computing representations for words that are not available in the corpus.

Background - Word Embeddings

Sentence Embeddings:

- Two problems in word embeddings: **dimension of input** and **longer training times**.
- Sentence embeddings solve both problems: fixed, smaller dimensions.
- Computed by summing embeddings of all words in the sentence
- However, may cause loss of information.

Background - StyleGAN

- Proposed by Karras et al. (2018) with a novel generator architecture.
- StyleGAN is a **Wasserstein** GAN based on **Progressively Growing GAN**.
- Training starts with only a few layers, more layers are added as training progresses.
- Each block is responsible from different level of details in resulting output.

Background - StyleGAN

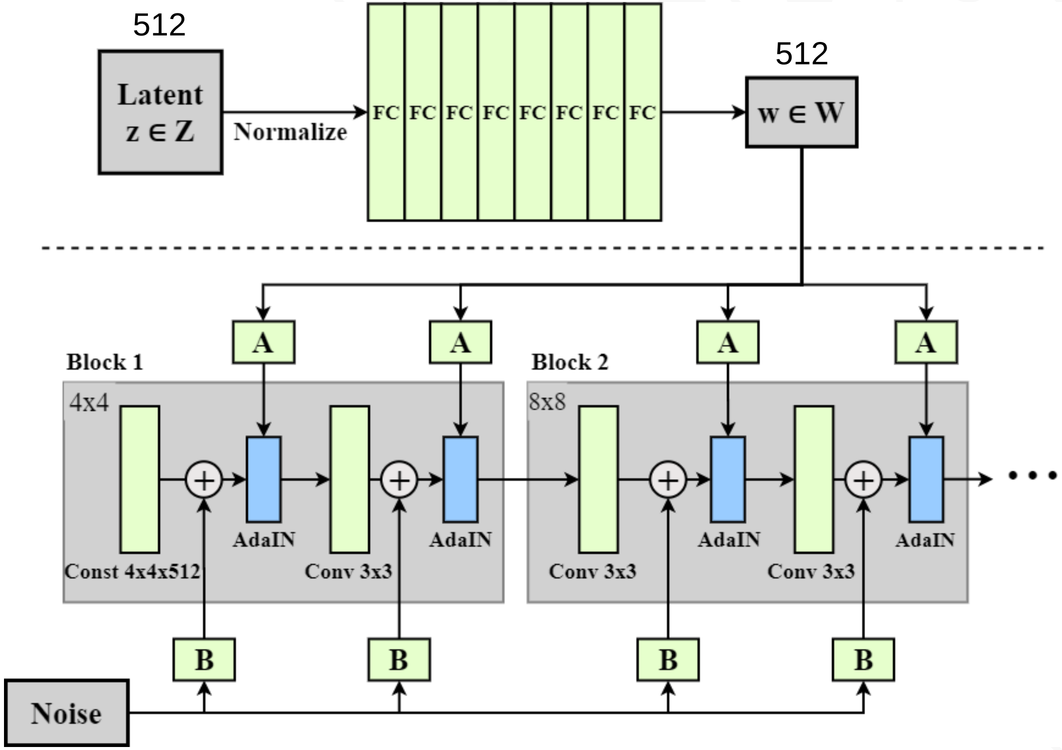
Generator of StyleGAN is composed of two parts:

1) Mapping Network

2) Synthesis Network

A : Learned Affine Transform

B : Learned per-channel scaling factor



Background - StyleGAN

Adaptive Instance Normalization (AdaIN) is an extension to Instance Normalization proposed by Ulyanov et al. (2016).

Instance Normalization:

$$y_{tijk} = \frac{x_{tijk} - \mu_{ti}}{\sqrt{\sigma_{ti}^2 + \epsilon}}, \quad \mu_{ti} = \frac{1}{HW}\sum_{l=1}^{W}\sum_{m=1}^{H} x_{tilm}, \quad \sigma_{ti}^2 = \frac{1}{HW}\sum_{l=1}^{W}\sum_{m=1}^{H}(x_{tilm} - mu_{ti})^2$$

Adaptive Instance Normalization:

$$\mathrm{AdaIN}(x, y) = \sigma(y)\left(\frac{x - \mu(x)}{\sigma(x)}\right) + \mu(y)$$

Background - Position Map Regression Network

Based on 3D Morphable Model, proposed by Feng et al. (2018) as an end to end architecture that does:

- Face Alignment,
- Regression of 3D Facial Geometry.

Previous methods that predict parameters instead of coordinates directly are harder to train and achieving good results require special care.

Background - Position Map Regression Network

- Proposes **UV Position Map** to represent 3D facial structure.
- Encoder has one convolutional layer, followed by 10 residual blocks.
- Decoder contains 17 transposed convolution layers.
- The Green rectangles represent the residual blocks (encoder), and the blue ones represent the transposed convolutional layers (decoder).

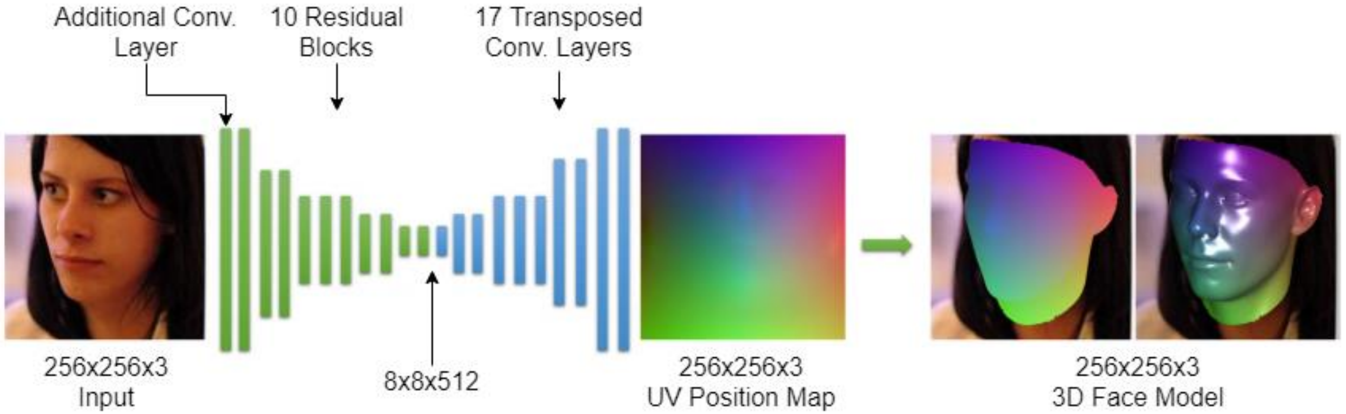
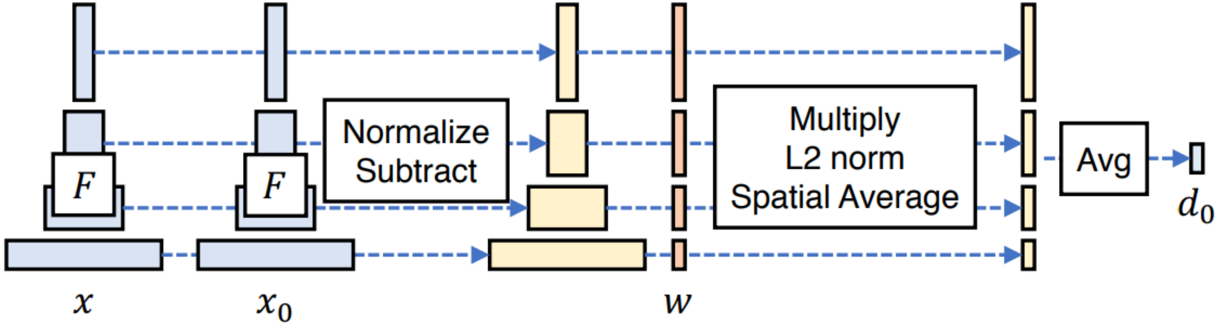


*Figure taken and modified from [13]

Background - LPIPS

- Comparing two images directly with Euclidean Distance is not useful in our case.
- Learned Perceptual Image Patch Similarity (LPIPS) focuses on perceptive similarity.
- Zhang et al. (2018) used deep features that are obtained from intermediate layers of neural networks for comparing image similarity.
- Pretrained AlexNet model is used for computing deep embeddings.

Background - LPIPS



Computing Distance

Similarity between two images:

$$d(x, x_0) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} ||w_l \odot (\hat{y}^l_{hw} - \hat{y}^l_{0hw})||_2^2$$

*https://arxiv.org/pdf/1801.03924.pdf

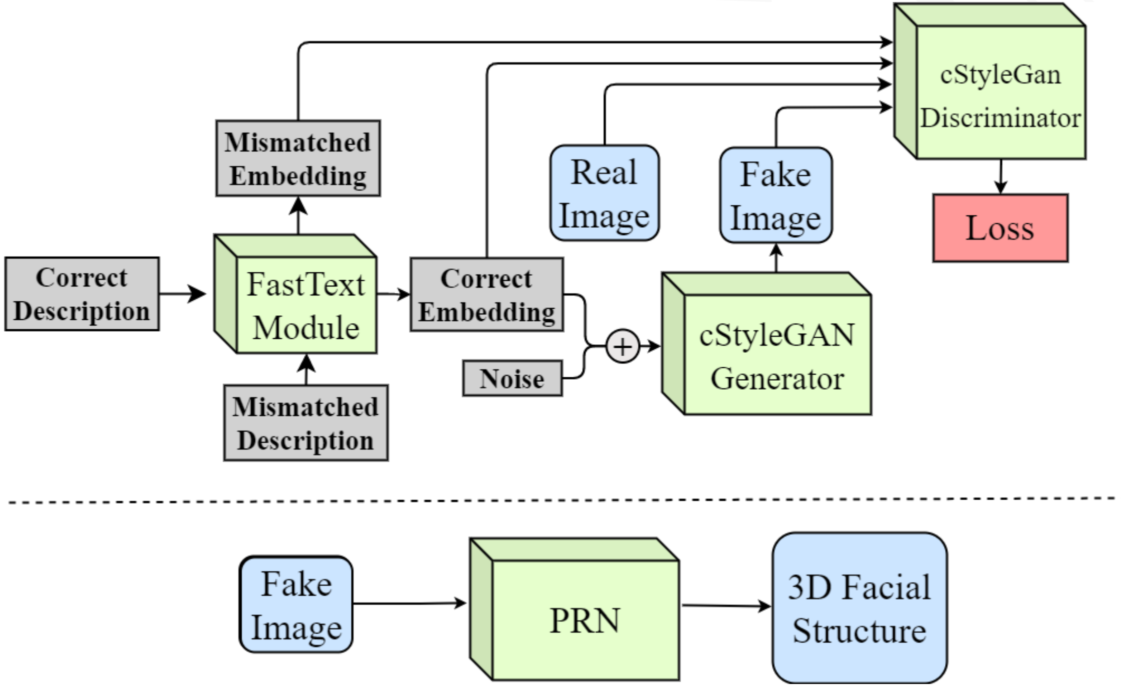
Proposed Methods

Our methods can be divided as follows:

- Our proposed Conditional StyleGAN (cStyleGAN) which uses sentence embeddings.
- Our GAN evaluation measure based on LPIPS: Perceptual Quality Distance (PQD).
- Improving 3D models by applying post-process onto 3D models.
- Extending CelebA dataset by providing descriptions generated by using annotations that are available in CelebA dataset.

Proposed Methods

System Pipeline:



Proposed Methods

cStyleGAN:

- Achieved by conditioning both generator and discriminator of StyleGAN on conditioning variable.
- Input is 300 dimension sentence embeddings obtained from FastText module + 100 dimension noise.
- Trained on extended CelebA and Face2Text datasets.
- Generates 64x64 resolution realistic images aligned with given descriptions.

Proposed Methods

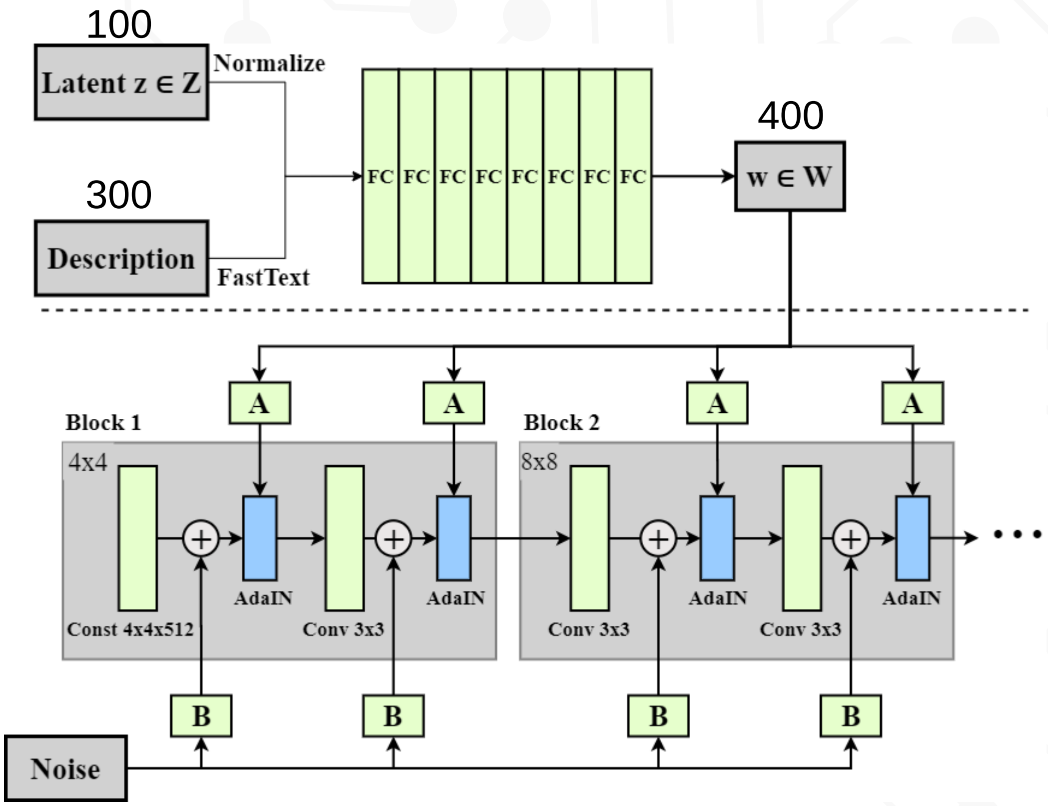
Generator of cStyleGAN:

Mapping Network:

- Input is sentence embeddings concatenated with noise.

Synthesis Network:

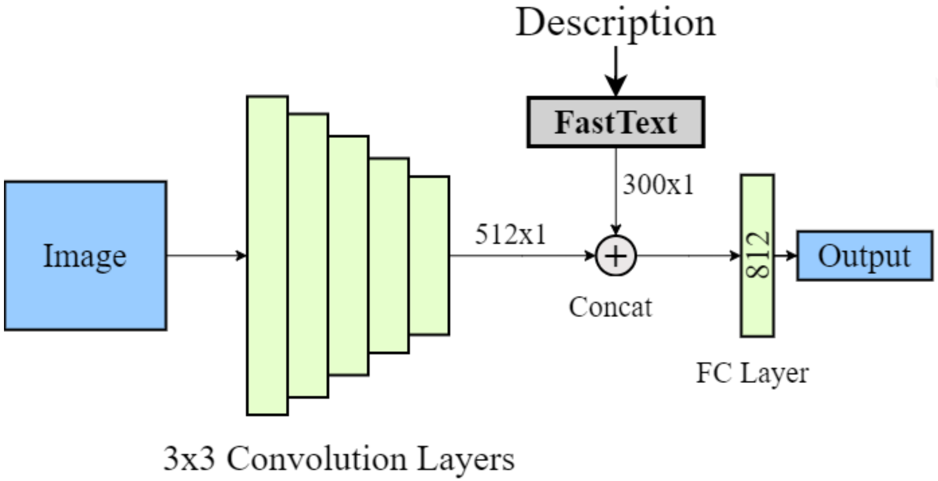
- Output is 2D face image with changing size.
- Best results are achieved at 64x64.
- Noise sampled from *N*(0, 0.07)



Proposed Methods

Discriminator Type 1 of cStyleGAN:

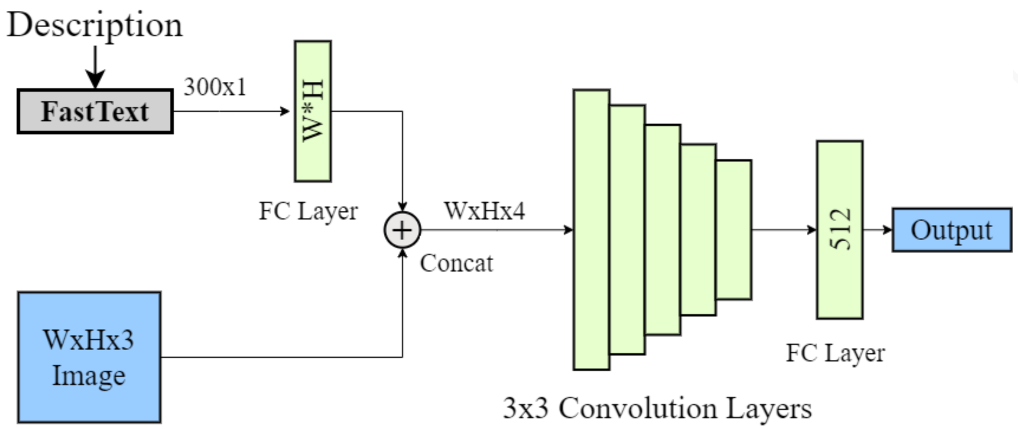
- Convolution layers are not utilizing information of word embeddings
- Outputs are not successfully conditioned



Proposed Methods

Discriminator Type 2 of cStyleGAN:

- Convolution layers are utilizing information of word embeddings as well as FC layer.
- Outputs are successfully conditioned.
- Additional learned FC Layer for reshaping embeddings.



Proposed Methods

Perceptual Quality Distance (PQD):

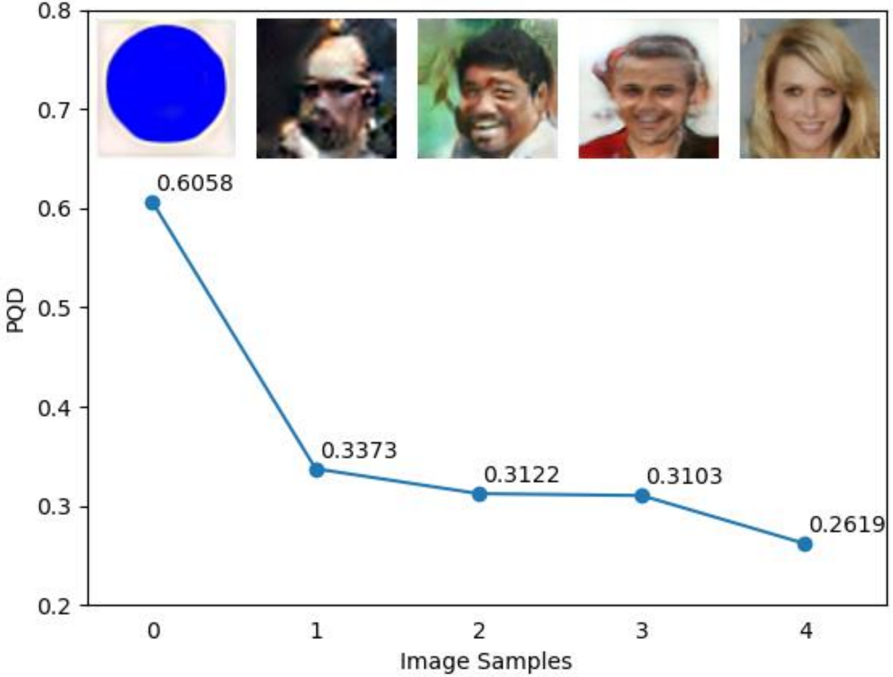
- Based on LPIPS.
- Compares distance between generated images and dataset.

PQD between two sets of images are calculated as follows:

$$PQD(x_g, x_r) = \sum_{m}^{M} t_m \sum_{n}^{N} t_n F(x_{gn}, x_{rm}) \frac{1}{N} \frac{1}{M}$$

Proposed Methods

PQD Scores:



Proposed Methods

Improvements on 3D pipeline:

- 3D model smoothed in a window of 7.
- RGB image converted to grayscale, normalized to range [-0.5, 0.5]
- Values outside of range [-0.49, 0.49] and values between range [-0.35, 0.26] eliminated.
- Estimated height-map is applied to Z channel of 3D model for enhanced depth information.
- Textures of 3D models are upscaled by using pre-trained ESRGAN.

Proposed Methods

Improvements on 3D pipeline:

Proposed Methods

Extending CelebA dataset with textual descriptions:

- Originally, there are 40 annotations for each 202599 image in CelebA dataset.
- We have generated randomised sentences using annotations.
- Low quality annotations are eliminated.
- **48069** descriptions are generated after eliminating low quality annotations.

Proposed Methods

Annotation: ['-1', '-1', '1', '-1', '-1', '-1', '-1', '-1', '-1', '1', '-1', '-1', '-1', '-1', '-1', '-1', '-1', '-1', '1', '1', '-1', '1', '-1', '-1', '1', '-1', '-1', '1', '-1', '-1', '-1', '-1', '-1', '1', '1', '-1', '1', '-1', '-1', '-1']

Attractive: 1
Blond_Hair: 1
Heavy_Makeup: 1
Male: -1
Pointy_Nose: 1
Wavy_Hair: 1
Young: -1

Generated Description: An old female with blond wavy hair. She has pointy nose. The attractive woman is wearing heavy make up.

Annotation Index	Meaning
0	5_o_Clock$shadow$
1	Arched_Eyebrows
2	Attractive
3	Bags_Under_Eyes
4	Bald
5	Bangs
6	Big_Lips
7	Big_Nose
8	Black_Hair
9	Blond_Hair
10	Blurry
11	Brown_Hair
12	Bushy_Eyebrows
13	Chubby
14	Double_Chin
15	Eyeglasses
16	Goatee
17	Gray_Hair
18	Heavy_Makeup
19	High_Cheekbones
20	Male
21	Mouth_Slightly_Open
22	Mustache
23	Narrow_Eyes
24	No_Beard
25	Oval_Face
26	Pale_Skin
27	Pointy_Nose
28	Receding_Hairline
29	Rosy_Cheeks
30	Sideburns
31	Smiling
32	Straight_Hair
33	Wavy_Hair
34	Wearing_Earrings
35	Wearing_Hat
36	Wearing_Lipstick
37	Wearing_Necklace
38	Wearing_Necktie
39	Young

Proposed Methods

Numbers of Eliminated Annotations:

Reason	Number of images eliminated
Have less than specified amount (8) of annotations marked as "1"	84251
Conflicting annotations marked as "1"	77236
Skipped for preserving male to female ratio	70138
Generated description amount	48069

Experiments

Datasets:

- **Face2Text Dataset**: has 4076 images from CelebA dataset with descriptions.
- **CelebA Dataset**: has 202599 images, each with 40 annotations.
- **Extended Dataset**: Combined Face2Text and additional images from CelebA with generated descriptions, resulting in **52145 images with descriptions**.
- A toy dataset of colored shapes for testing conditioning performance.

Experiments

cStyleGAN trained only on Face2Text Dataset:

A young man with short brown hair and small eyes. His eyebrows are thick. His nose is small and his lips are thin. A stubble is growing on his face. He has got a well - defined jawline.

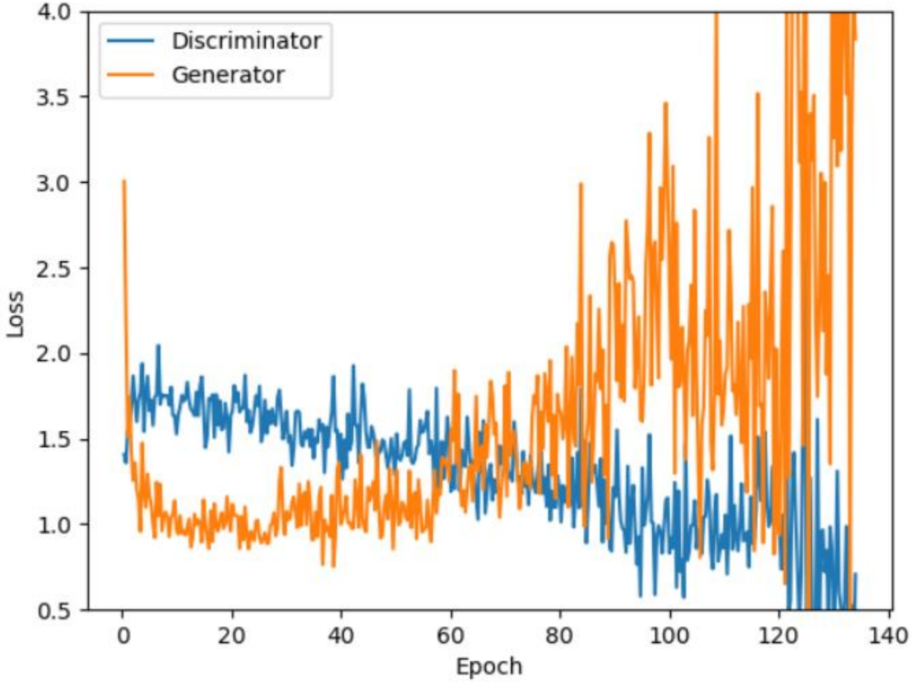
A man with short, brunette hair, thick eyebrows, light eyes, a long nose and a stubble beard and moustache. He has thin, smiling lips and a prominent chin.

A woman with curtained, blond hair, a beaky nose, dark eyes, thin eyebrows and a wide smile.

A woman with long blonde hair, big green eyes and full pouting lips. She is wearing makeup.

Experiments

Exploding losses after 64x64 resolution (Epoch 120):



Experiments

Effects of Noise:

- Noise is sampled from ***N*** **(0, 0.07)**.
- Fixed descriptions with different noise values result in different images.

A serious looking woman with straight blond hair. She has arched eyebrows. Her eyes are brown and big and her lips are thin. She has got a heavy lower lip.

Experiments

Data Augmentation:

- Augmenting data caused unwanted results.
- Model learned augmentations as well.

Applied Transform	Probability of Being Applied
Random Horizontal Flip	50%
5 degrees of random rotation	100%
Adding Gaussian Noise $\mathcal{N}(0, 0.07)$	30%
Erasing random patch from image	1%

Experiments

Ablation Study:

- We have disabled certain convolution blocks in synthesis network.
- Subsequent blocks are dependent on previous blocks.
- **Disabling Block 2**: Loss of general facial structure details and image composition
- **Disabling Block 3**: More detailed, but still a blurry face image
- **Disabling Block 2 and Block 3**: Similar result of only disabling Block 3; Block 3 is dependent on the output of Block 2

(a) Original Output

(b) Block 2 Disabled

(c) Block 3 Disabled

(d) Block 2 and 3 Disabled

Results

Results of cStyleGAN trained on Face2Text Dataset (4K data instances):

- **Left:** Mode collapsed with 1.5m iterations per each resolution.
- **Right:** Reducing iterations to 600K resulted in better outputs.

Results

Results of cStyleGAN trained on Extended Dataset:

- Best results so far.
- No Mode Collapse experienced.
- Trained with 1.5m iterations per each resolution.
- Generated images are 64x64 resolution.

Results

A woman with wavy black hair. She has arched eyebrows. Her eyes are brown and small and her lips are thin. She has got a heavy lower lip. She looks serious.

A woman with straight black hair. She has arched eyebrows. Her eyes are brown and small and her lips are thin. She has got a heavy lower lip. She is smiling.

A woman with wavy black hair with bangs. She has arched eyebrows. Her eyes are brown and small and her lips are thin. She has got a heavy lower lip. She is smiling.

A woman with straight blond hair with bangs. She has arched eyebrows. Her eyes are brown and small and her lips are thin. She has got a heavy lower lip. She is smiling.

A pale skinned woman with straight blond hair with bangs. She has arched bangs. She has arched eyebrows. Her eyes are brown and small and her lips are thin. She has got a heavy lower lip. She is smiling.

A serious looking woman with straight blond hair. She has arched eyebrows. Her eyes are brown and big and her lips are thin. She has got a heavy lower lip.

A male with brown hair. The man has a serious look on his face. He has big nose. He has bags under his eyes. He has no beard.

A male with brown hair. The man is smiling. He has big nose. He has bags under his eyes. He has no beard.

A male with brown hair. He has oval face. The man is smiling. He has small, pointy nose. He has bags under his eyes. He has a goatee.

Results

Comparison with Existing Work:

(a) Pro-StackGAN on Face2Text

(b) Pro-StackGAN on Extended

(c) Text2FaceGan

(d) cStyleGAN on Face2Text (600K iterations)

(e) cStyleGAN on Extended

Model	Dataset	PQD
Pro-StackGAN	Face2Text	0.36279
Pro-StackGAN	Extended	0.31031
cStyleGAN (ours)	Face2Text	0.28584
cStyleGAN (ours)	Extended	0.26478

Inception Score	
Text2FaceGAN	1.4 ± 0.7
Ours (Face2Text Dataset)	1.9 ± 0.2
Ours (Extended Dataset)	2.4 ± 0.1

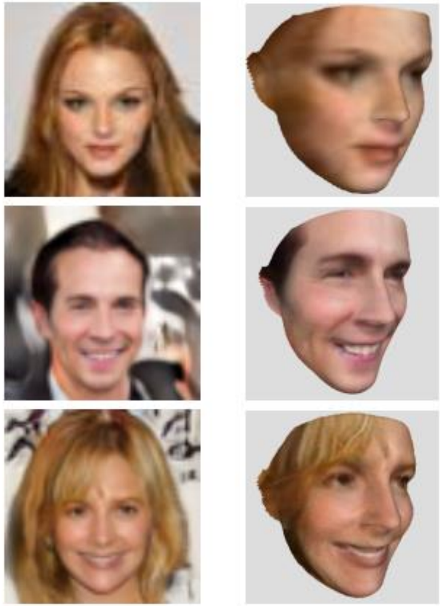
Results

3D Generation:

A serious looking woman with straight blond hair. She has arched eyebrows. Her eyes are brown and big and her lips are thin. She has got a heavy lower lip.

A young male with black straight hair. He has bags under his eyes. He has no beard. The man is attractive. He has big nose. The man is smiling.

A young female with blond hair with bangs. The woman is attractive. She has bags under her eyes. The woman is smiling. She has pointy and big nose.



Input Resolution	Time
8x8	N/A
16x16	N/A
32x32	N/A
64x64	1.471 seconds
128x128	3.713 seconds
256x256	12.517 seconds

Conclusion

In conclusion:

- Existing Face to Text implementations do not achieve high quality results.
- Existing methods do not utilize big datasets such as CelebA, limited to Face2Text Dataset.
- Results can be improved by utilizing synthetic descriptions.
- There is no existing work that generates end to end 3D facial models from text data.
- GAN evaluation methods are lacking.

Conclusion

In this thesis:

- We have improved existing Text to Face results using Conditional StyleGAN.
- We extend CelebA Dataset by providing synthetic descriptions and utilize in our training.
- We propose our evaluation measure: Perceptual Quality Distance.
- Finally, we provide end to end 3D face generation pipeline from textual descriptions.

Conclusion

In this thesis, we show that:

- StyleGAN can be used for achieving state of the art text 2 face results.
- Pre-trained word embedders such as FastText can be used in Text to Face domain without additional training.
- Synthetic sentences can be used for improving results.

Thank you for listening!

References

[9] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4401–4410, 2019.

[12] A. Gatt, M. Tanti, A. Muscat, P. Paggio, R. A. Farrugia, C. Borg, K. P. Camilleri, M. Rosner, and L. Van der Plas, “Face2text: collecting an annotated image description corpus for the generation of rich face descriptions,” arXiv preprint arXiv:1803.03827, 2018.

[13] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou, “Joint 3d face reconstruction and dense alignment with position map regression network,” in Proceedings of the European Conference on Computer Vision (ECCV), pp. 534–551, 2018.

[20] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, “Generative adversarial text to image synthesis,” arXiv preprint arXiv:1605.05396, 2016.

[32] D. Ulyanov, A. Vedaldi, and V. Lempitsky, “Instance normalization: The missing ingredient for fast stylization,” arXiv preprint arXiv:1607.08022, 2016.

[37] V. Blanz and T. Vetter, “A morphable model for the synthesis of 3d faces,” in Proceedings of the 26th annual conference on Computer graphics and interactive techniques, pp. 187–194, 1999.

[40] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 586–595, 2018.

References

[42] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Advances in neural information processing systems, pp. 1097–1105, 2012.

[44] Akanimax, “Akanimax.t2f: text to face generation using deep learning,” GitHub repository, https://github.com/akanimax/T2F, 2018.

[45] O. R. Nasir, S. K. Jha, M. S. Grover, Y. Yu, A. Kumar, and R. R. Shah, “Text2facegan: Face generation from fine grained textual descriptions,” in 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM), pp. 58– 67, IEEE, 2019.