Deep Neural Networks for Conditional Visual Synthesis

Minglun Gong
School of Computer Science, University of Guelph

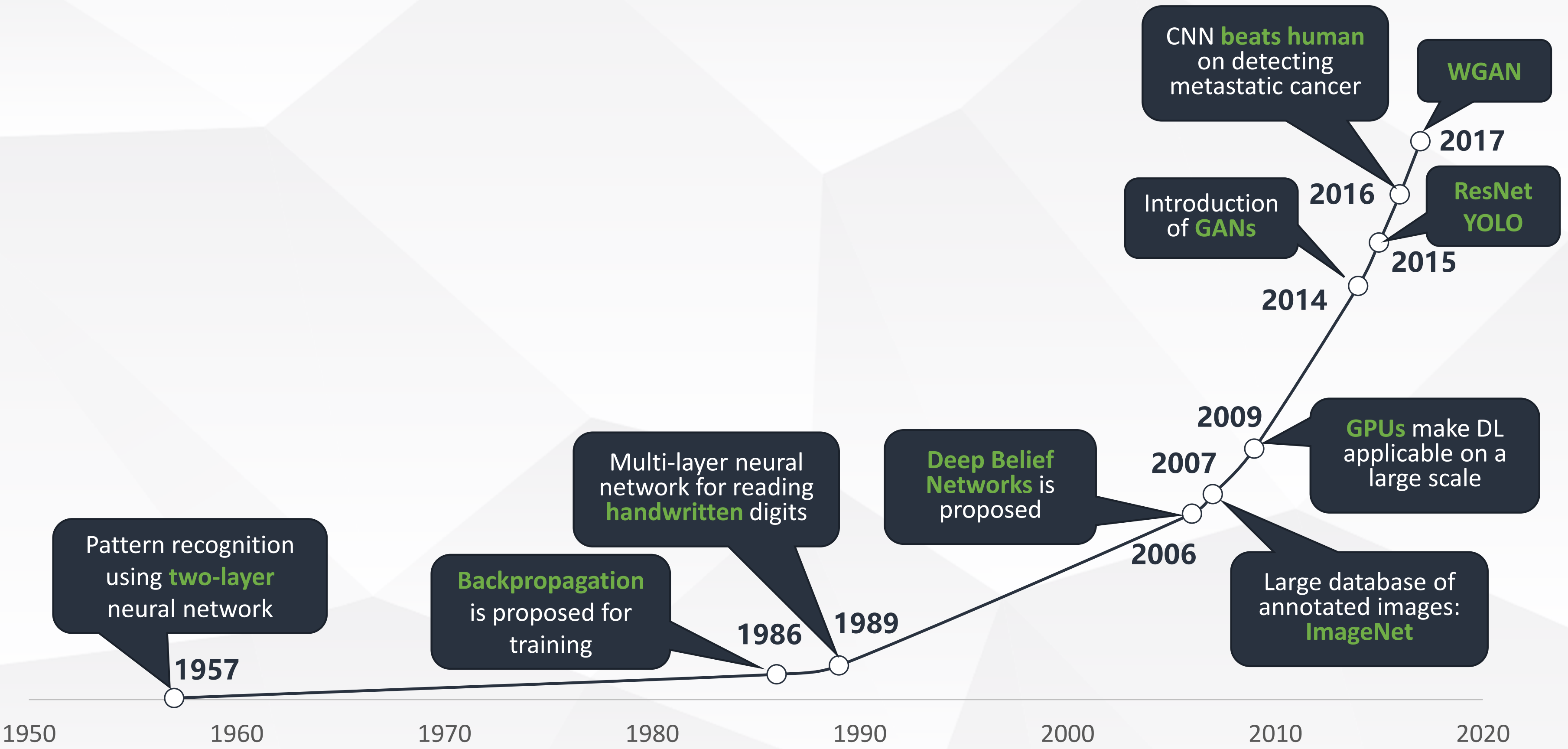
Based on the thesis work of Xin Huang
PhD student at Memorial University

UNIVERSITY of GUELPH

UNIVERSITY of GUELPH

Deep Learning in Computer Vision

- **Introduction**
- **Overview**
- **Task 1**
- **Task 2**
- **Task 3**
- **Task 4**
- **Summary**



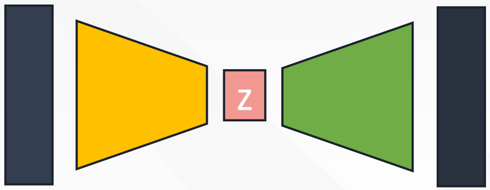
Visual Synthesis

Which images are fake?

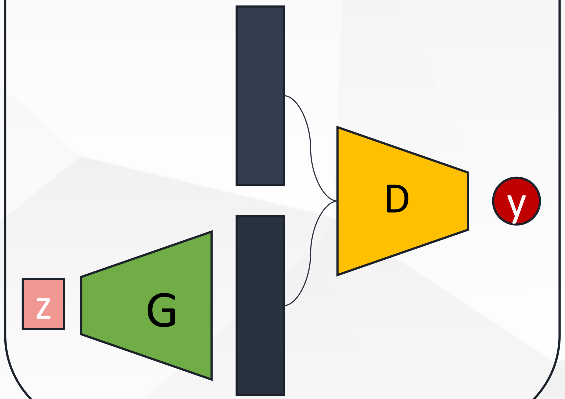
Unconditional vs. Conditional Synthesis

Latent variable models

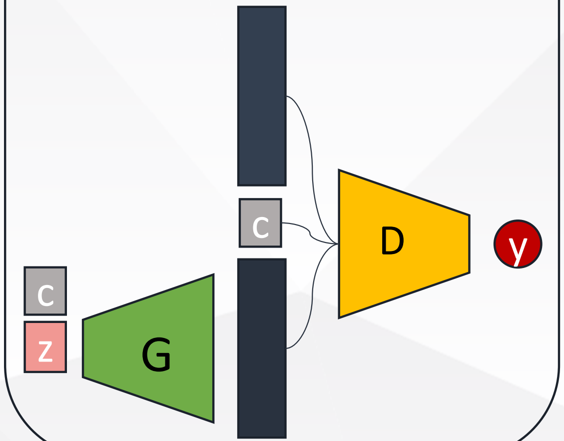
Autoencoders and Variational Autoencoders (VAEs)



Generative Adversarial Networks (GANs)



Conditional Generative Adversarial Networks (cGANs)



Objectives of This Work

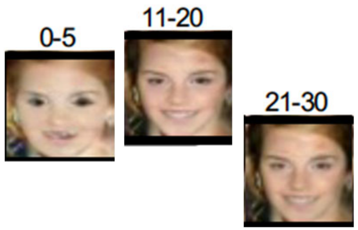
Text, age, audio, strokes

Text to Image Synthesis

Face Aging

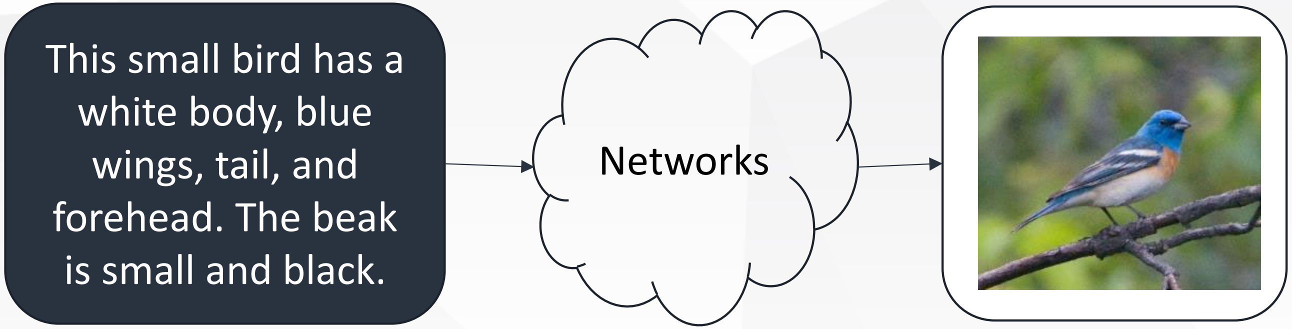
Talking Face Generation

Neural Painting



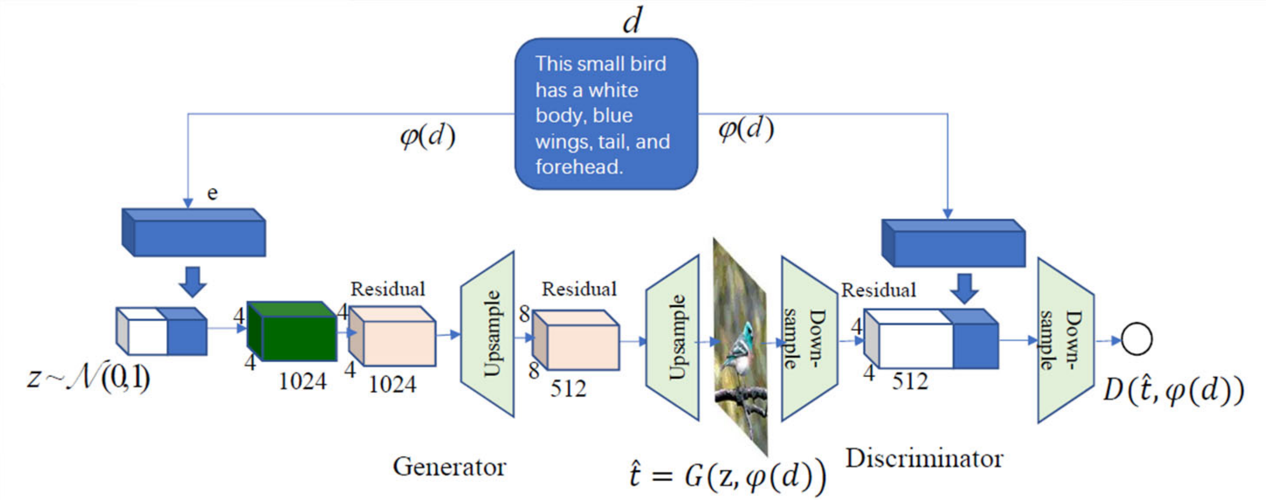
Text-to-Image Synthesis

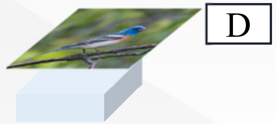
Introduction
Overview
Task 1
Task 2
Task 3
Task 4
Summary



Given an input text description or a description sequence, automatically produce an image or image sequence that semantically matches the input description.

Previous Work: GAN-CLS



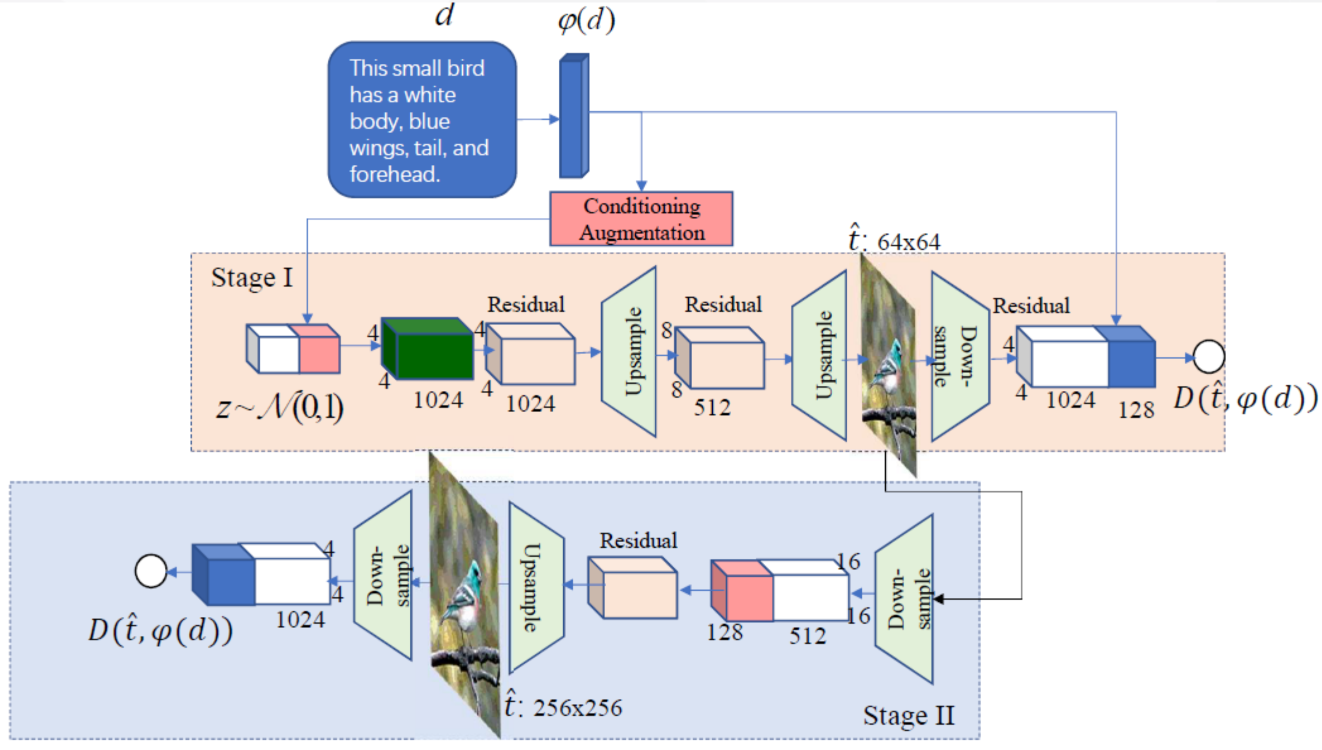


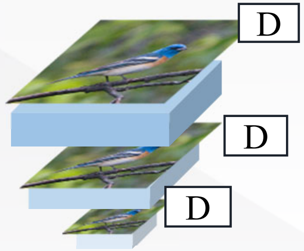
(a)

G: $\mathbb{R}^Z \times \mathbb{R}^T \rightarrow \mathbb{R}^D$ D: $\mathbb{R}^D \times \mathbb{R}^T \rightarrow \{0, 1\}$

[1] 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.

Previous Work: StackGAN

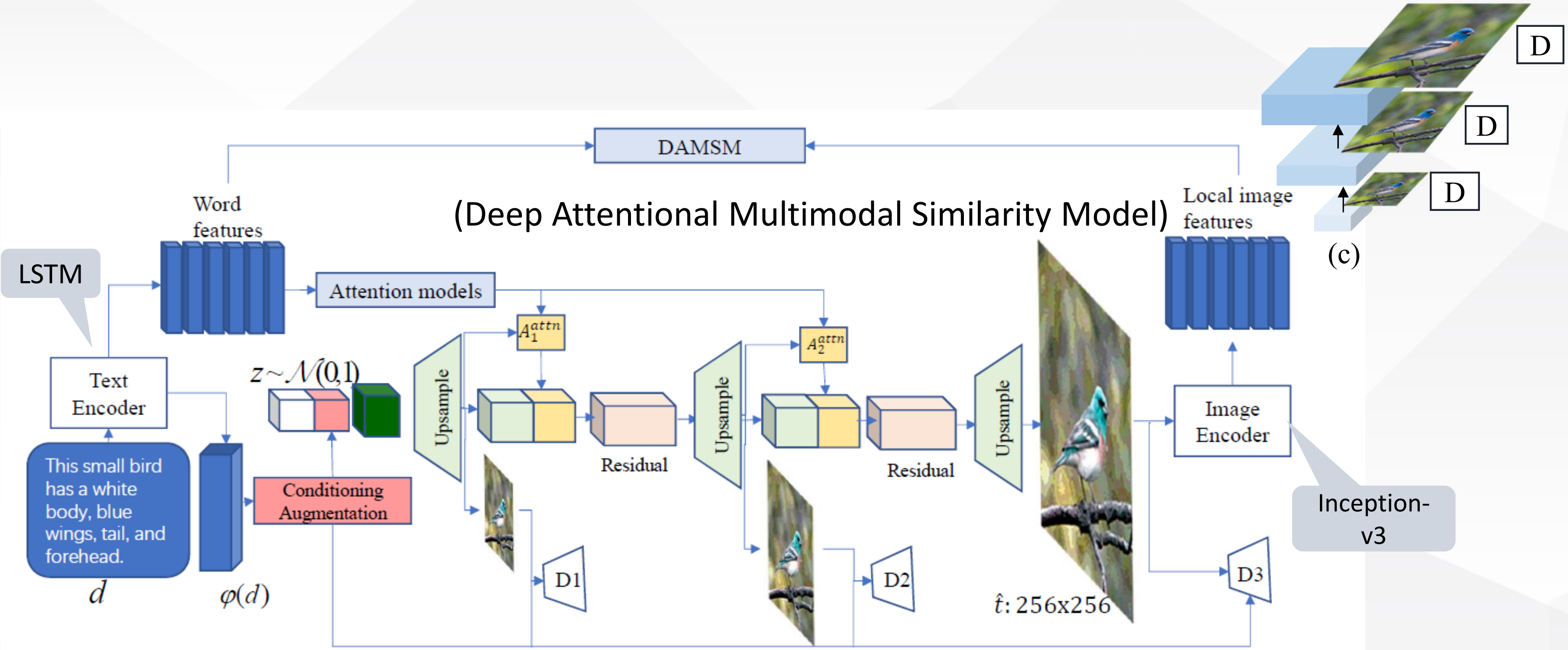




(b)

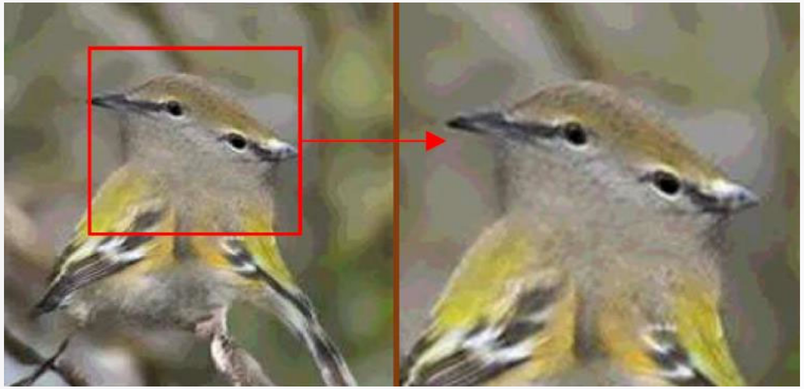
[2] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, “Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks,” in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5907–5915.

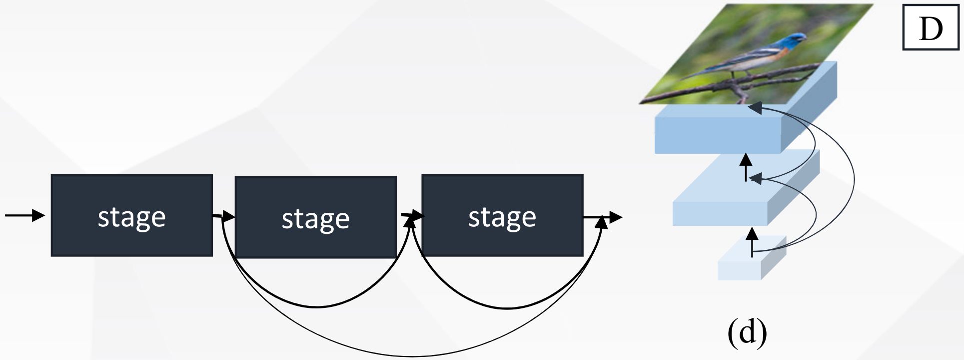
Previous Work: AttnGAN

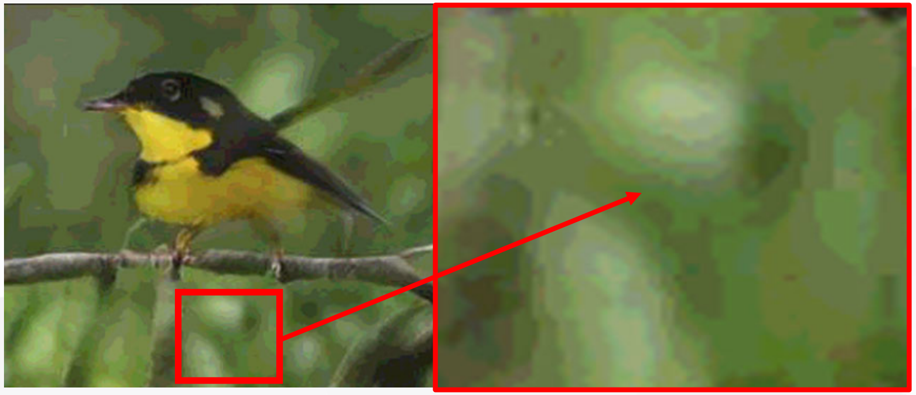


[3] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, “Attngan: Fine-grained text to image generation with attentional generative adversarial networks,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1316–1324.

Motivation



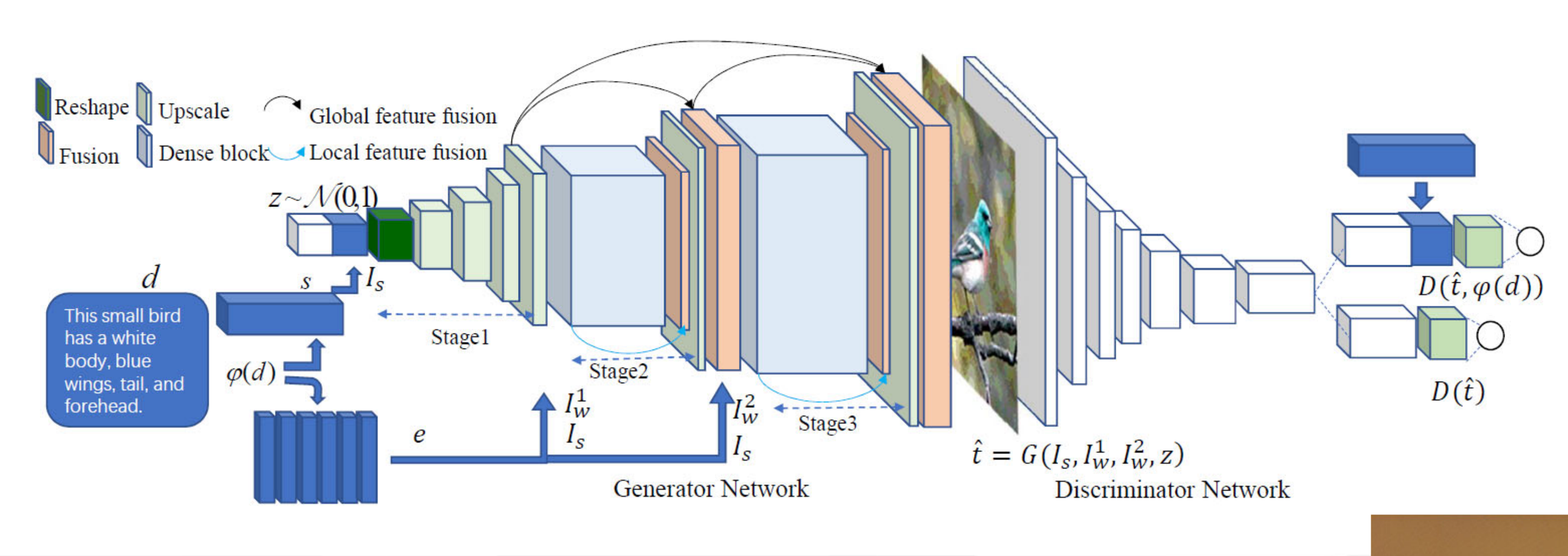




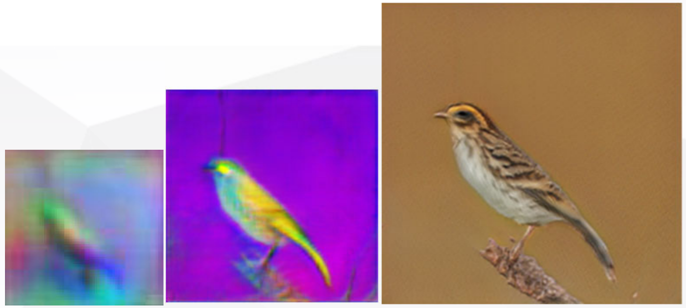
Use only **one pair** of generator and discriminator to:

1) Allow end-to-end training
2) Synthesize high-resolution and **consistent** images

Hierarchically-fused GAN

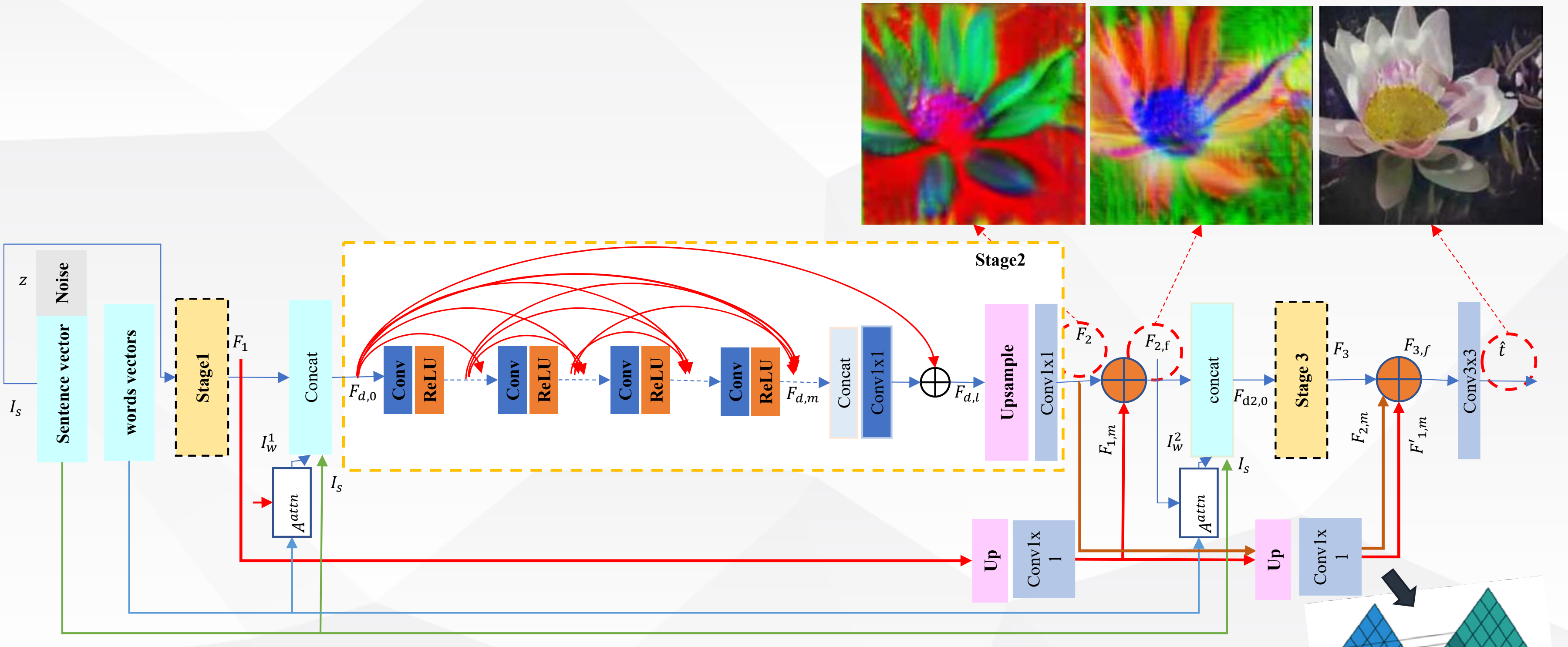


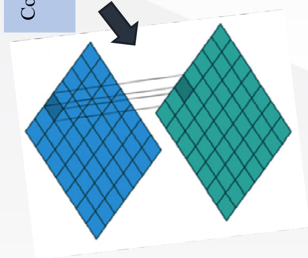
$$\mathcal{L} = arg \min_{G} \max_{D} V(D, G, I_s, I_w, z) = \mathcal{L}_G + \lambda \mathcal{L}_{DAMSM}$$



UNIVERSITY of GUELPH

Hierarchically-fused Generator





$$F_{3,f} = F_3 + H_{\mathrm{sup2}}(H_{\mathrm{sup1}}(F_1)) + H_{\mathrm{sup2}}(F_2)$$

Experiments

Datasets

	Descriptions/image	Images
Oxford-102 (flowers)	10	8,189
CUB (birds)	10	11,788
COCO (general)	5	82,783

Evaluation Metrics

Inception Score

VS similarity

Learning rate:

0~150 iteration 0.001

150~300 iteration 0.0005

300~600 iteration 0.0002

UNIVERSITY of GUELPH

Results on Bird Dataset

Introduction

Overview

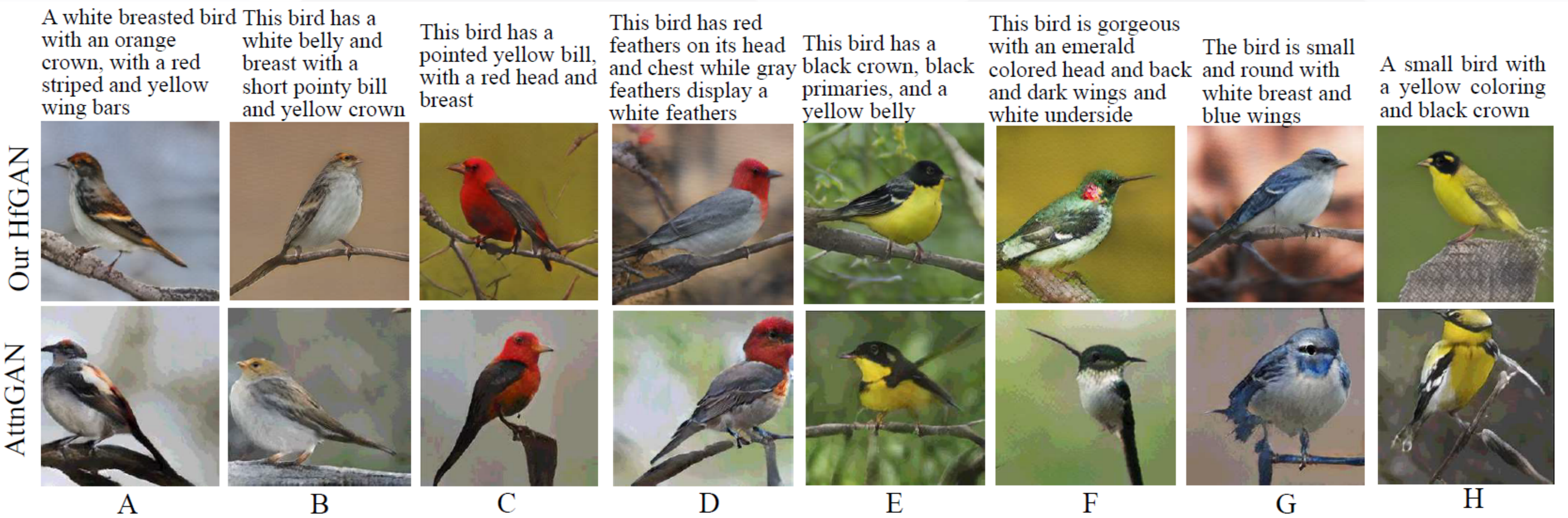
Task 1

Task 2

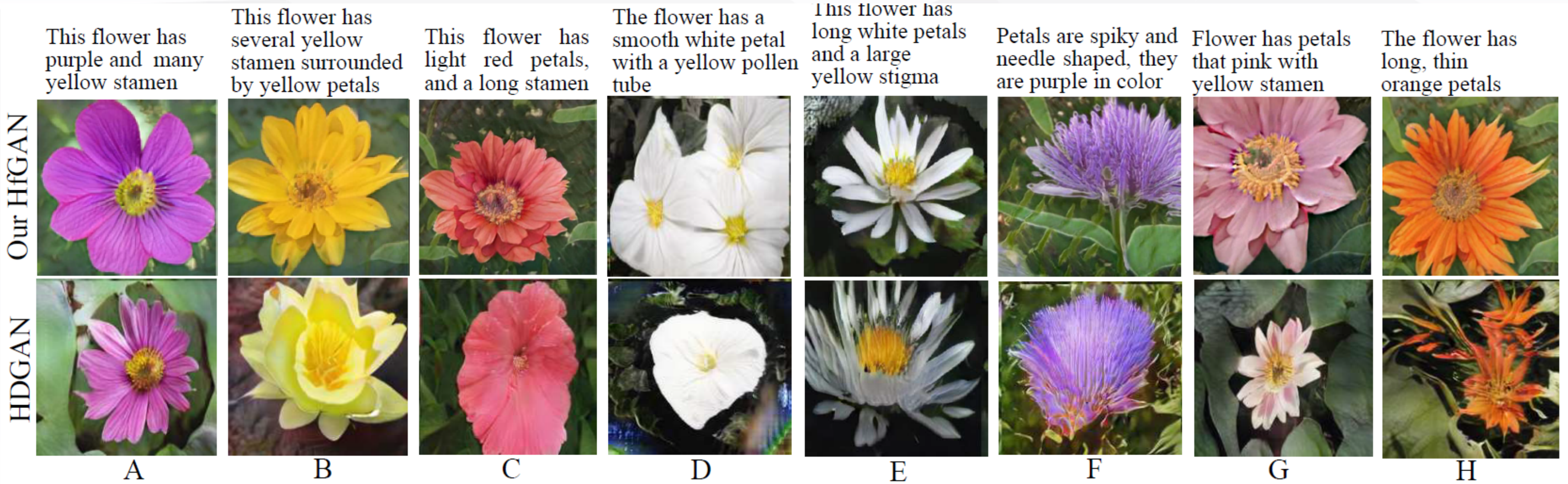
Task 3

Task 4

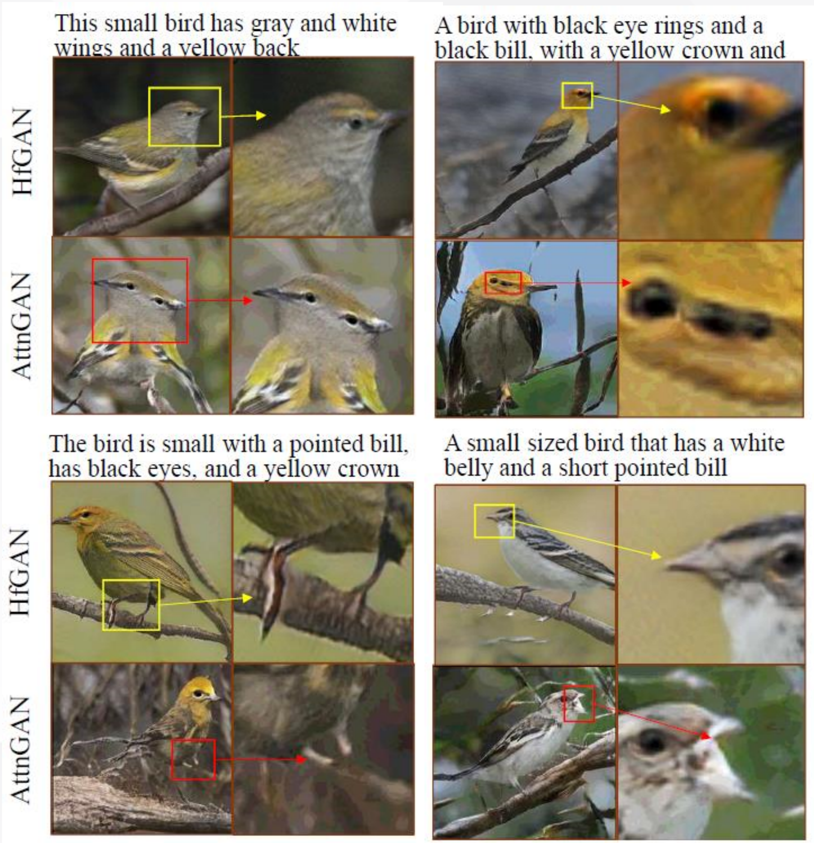
Summary

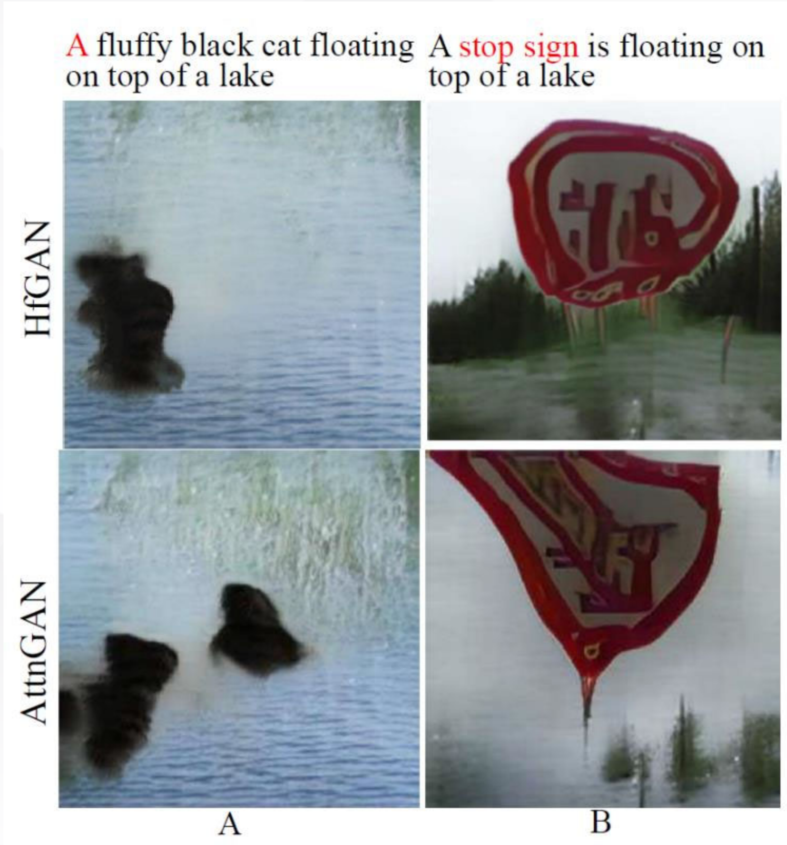


Results on Flower Dataset



More Comparisons & Failure Cases





UNIVERSITY of GUELPH

Quantitative Evaluations

- Introduction
- Overview
- Task 1
- Task 2
- Task 3
- Task 4
- Summary

INCEPTION SCORES ON THE THREE DATASETS OBTAINED BY PREVIOUS TEXT-TO-IMAGE MODELS AND OUR HFGAN. THE SCORES OF EXISTING APPROACHES ARE REPORTED IN THE RESPECTIVE PAPERS. THE HIGHEST SCORES ARE SHOWN IN **BOLD**.

Method	Dataset		
	Oxford-102	CUB	COCO
GAN-INT-CLS [1]	$2.66 \pm .03$	$2.88 \pm .04$	$7.88 \pm .07$
GAWWN [19]	/	$3.60 \pm .07$	/
StackGAN [2]	$3.20 \pm .01$	$3.70 \pm .04$	$8.45 \pm .03$
StackGAN++ [3]	/	$3.84 \pm .06$	/
TAC-GAN [29]	$3.45 \pm .05$	/	/
AttenGAN [4]	/	$4.36 \pm .03$	$25.89 \pm .47$
HDGAN [18]	$3.45 \pm .07$	$4.15 \pm .05$	$11.86 \pm .18$
Our HfGAN	$\mathbf{3.57 \pm .05}$	$\mathbf{4.48 \pm .04}$	$\mathbf{27.53 \pm .25}$

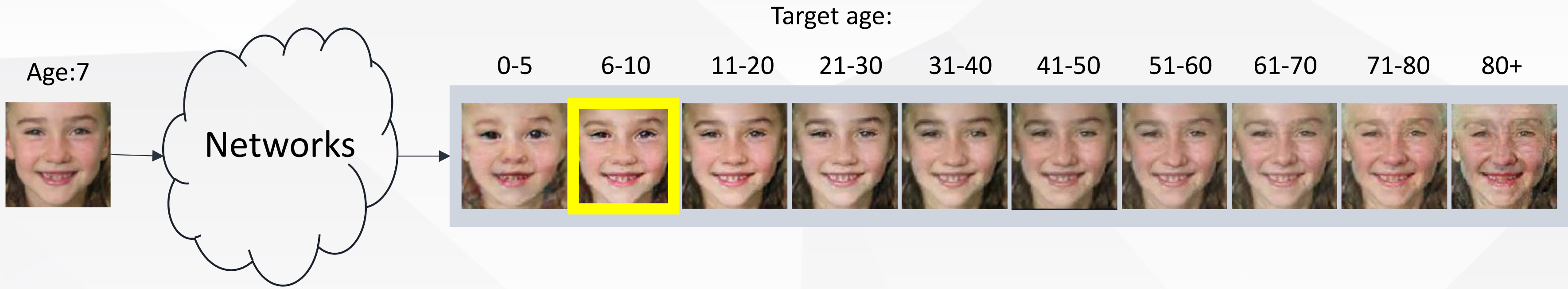
THE VS SIMILARITY SCORE ON THE THREE DATASETS BY PREVIOUS MODEL [2] [18] AND OUR MODEL

Method	Dataset		
	Oxford-102	CUB	COCO
StackGAN [2]	$.278 \pm .134$	$.228 \pm .162$	/
HDGAN [18]	$.296 \pm .131$	$.246 \pm .157$	$.199 \pm .183$
Our HfGAN	$\mathbf{.303 \pm .137}$	$\mathbf{.253 \pm .165}$	$\mathbf{.227 \pm .145}$

TRAINING TIME (S) / EPOCH

Dataset	Oxf-102	CUB
AttnGAN [4]	446.02	6891.54
Our HfGAN	**308.57**	**5614.73**

Face Aging



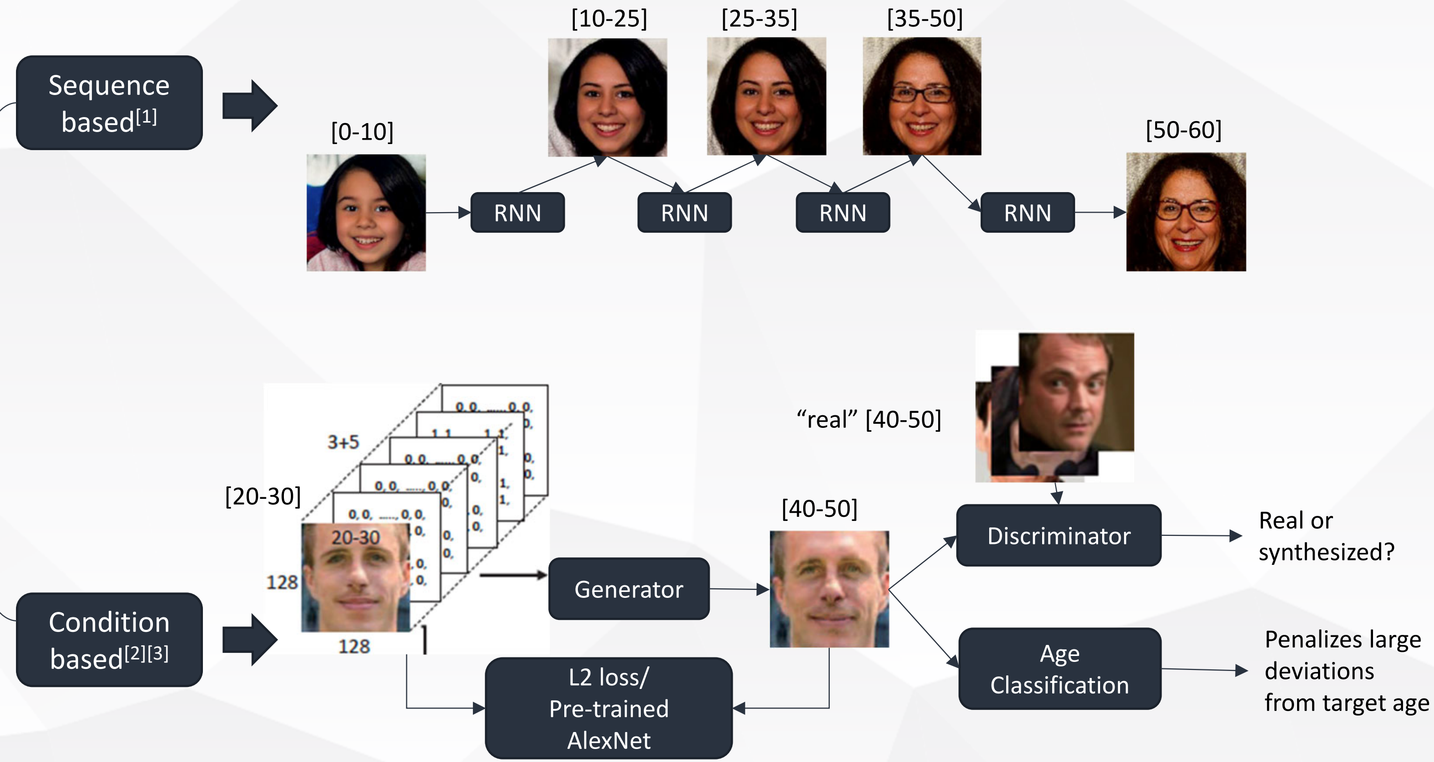
Given an input real face photo, synthesize the appearance of the same person under different ages.

UNIVERSITY of GUELPH

Previous Work

- Introduction
- Overview
- Task 1
- Task 2
- Task 3
- Task 4
- Summary

Age progression

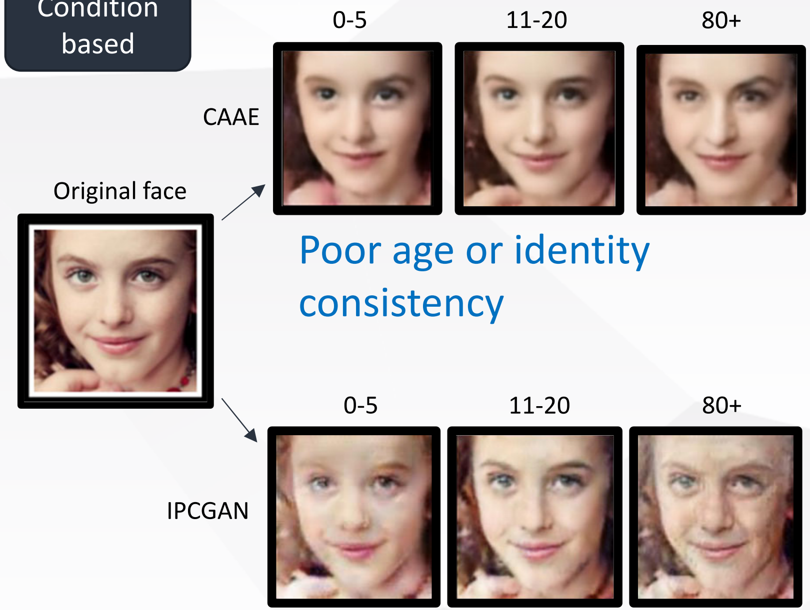


CAAE,IPCGAN

[1] Wang, Wei, et al. "Recurrent face aging." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
[2] Zhang, Zhifei, Yang Song, and Hairong Qi. "Age progression/regression by conditional adversarial autoencoder." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.
[3] Wang, Zongwei, et al. "Face aging with identity-preserved conditional generative adversarial networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

Motivation

Condition based



Poor age or identity consistency

Lack of changes in facial structures between young and old ages

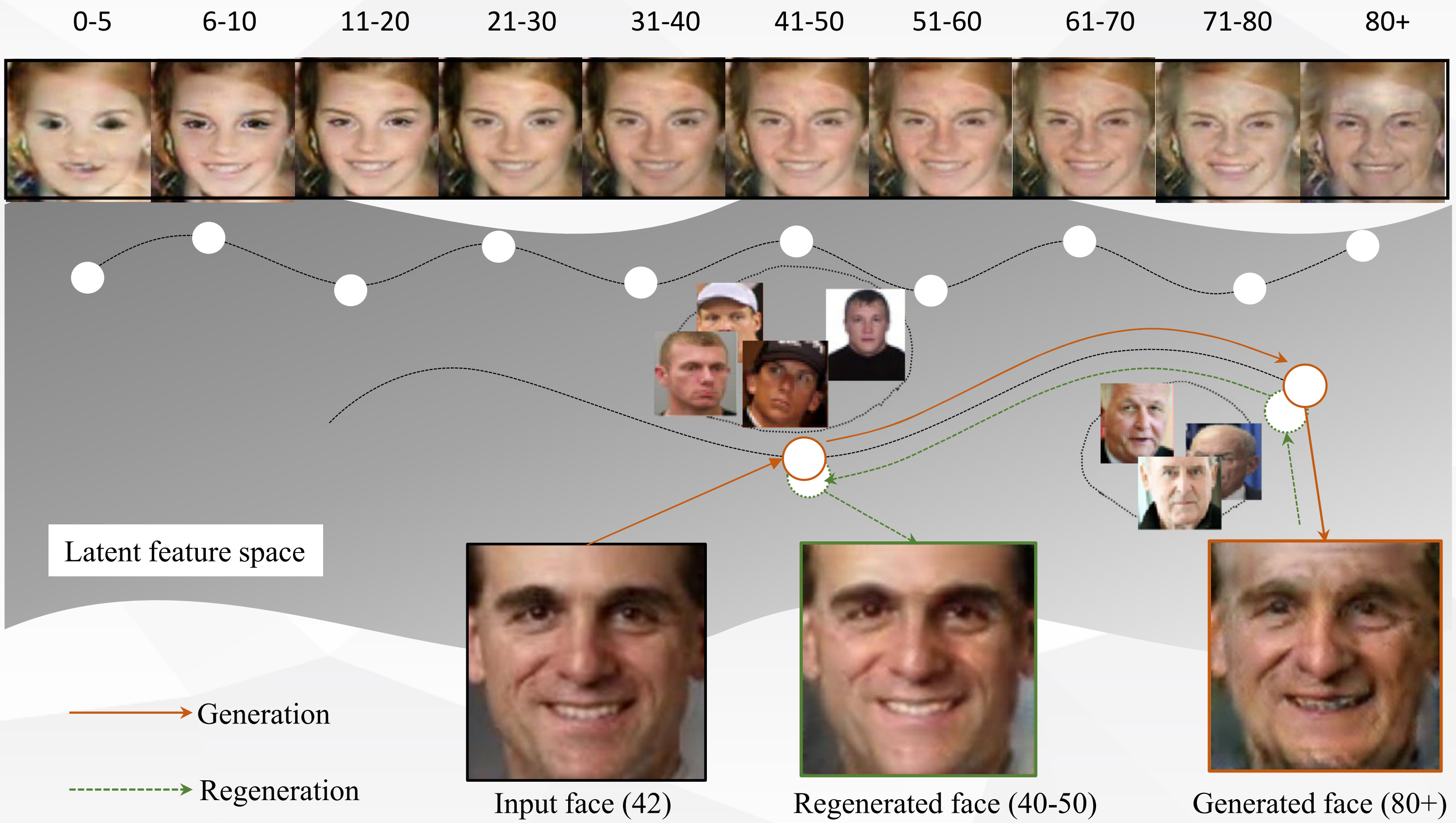
Design a conditional network that:

1) Use both **primal and inverse** tasks to ensure age and identity consistency.
2) Use **landmark attention** to generate face with more precise facial structure and poses.

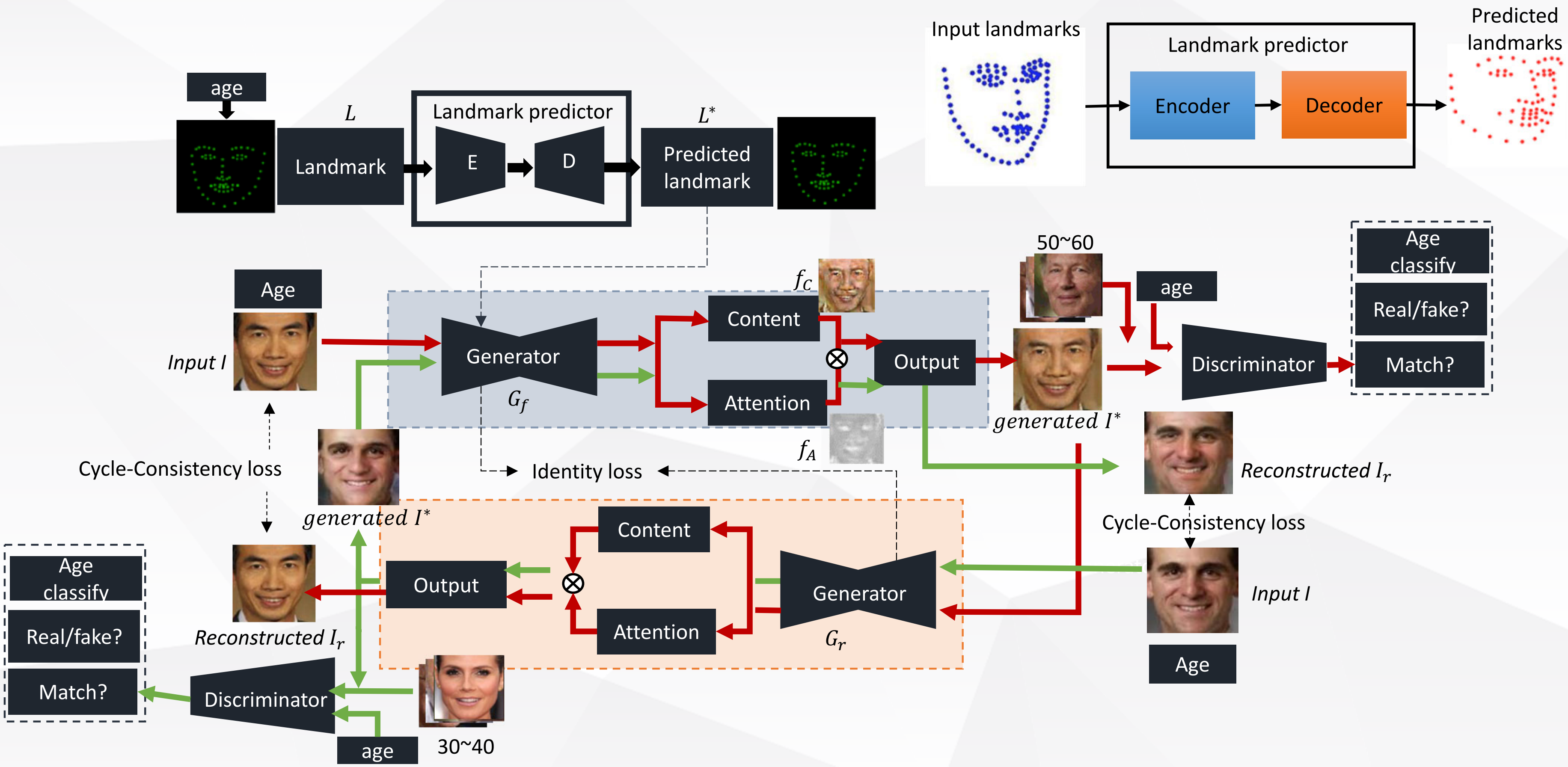
UNIVERSITY of GUELPH

Age Progression & Face Reconstruction

- Introduction
- Overview
- Task 1
- Task 2
- Task 3
- Task 4
- Summary



Landmark-Guided Conditional GAN



UNIVERSITY of GUELPH

Experiments

Datasets

Table 1: Image numbers in each age group of UTKFace.

Age group	0-5	6-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80	80+
Numbers	2204	850	1645	7736	4316	2091	2192	1160	676	530

Implementation Details

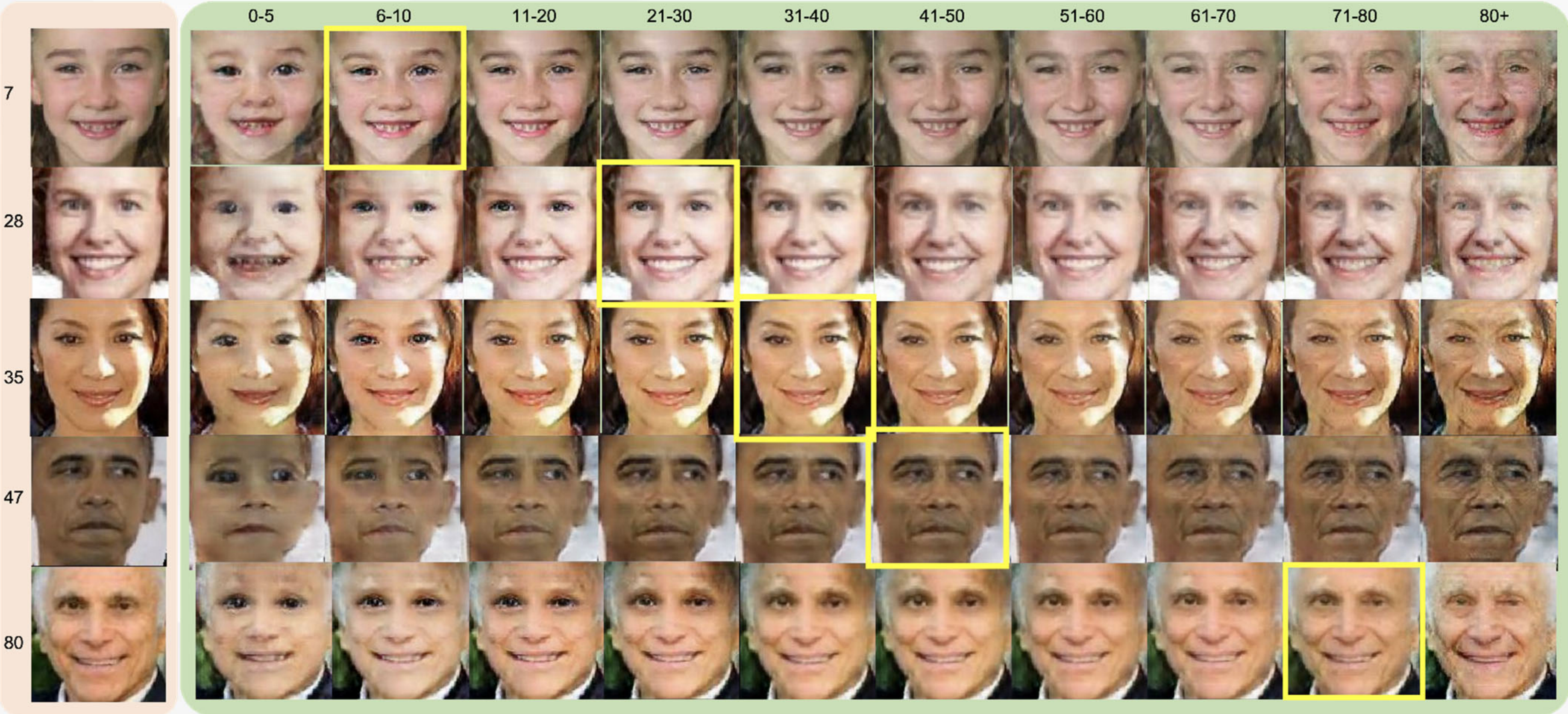
Implemented on PyTorch3

Tested on a single Nvidia GeForce GTX 1080 Ti GPU with 50GiB memory

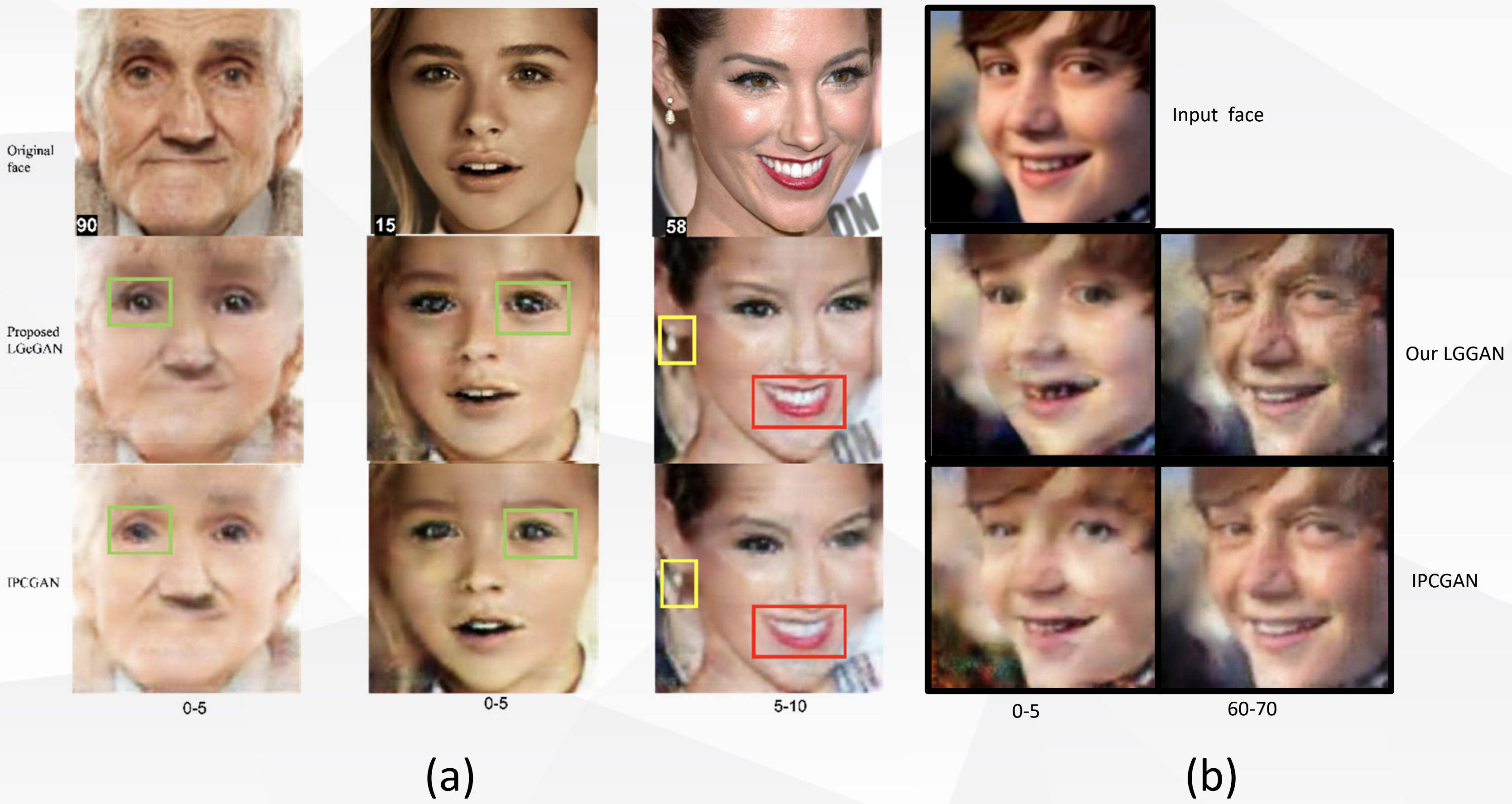
Applied batch normalization

Set a fixed learning rate of 0.0002.

Face Aging Results



Detailed Comparison with IPCGAN



UNIVERSITY of GUELPH

Quantitative Evaluation

Introduction

Overview

Task 1

Task 2

Task 3

Task 4

Summary

Aging Accuracy:

Table 1: Estimated Age Distributions on UTKFace. Generic is the mean value of each age group computed using the ground truth ages. Value in brackets shows the absolute differences from the ground truth mean age. Best value with minimize differences from generic are shown in boldface. ∗ represents the models that we re-trained on 10 age groups.

Age group	21-30	31-40	41-50	51-60	61-70	71-80	80+
Generic	25.03	35.01	45.12	54.63	65.40	73.66	87.29
CAAE* [34]	24.31(**0.72**)	32.43(2.58)	42.21(2.91)	51.49(3.14)	60.17(5.23)	70.57(3.09)	82.68(4.61)
IPCGAN* [31]	22.74(2.29)	31.74(3.27)	39.93(5.19)	50.04(4.59)	58.32(7.08)	68.42(5.24)	80.33(6.96)
Ours	26.18(**1.15**)	36.91(1.10)	44.68(**1.44**)	51.79(**2.84**)	62.52(**2.88**)	71.05(**2.61**)	88.24(**0.95**)

Identity preservation:

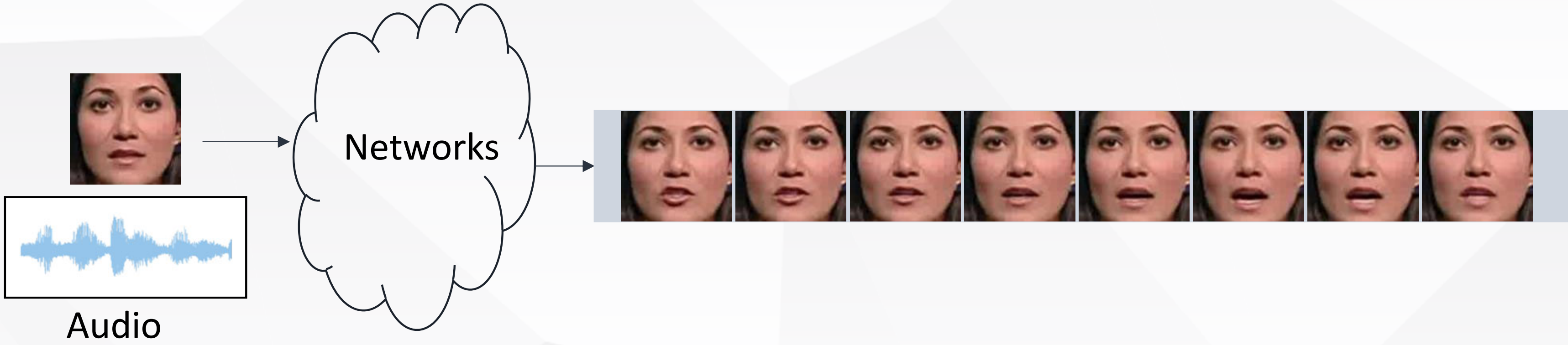
Table 2: Face verification results on UTKFace. The top is the Verification Confidence by our LGcGAN and the bottom is the verification rate for three methods. Best values for Verication Rate are indicated in bold.

Age group	21-30	31-40	41-50	51+
Verification Confidence				
10-20	95.76	94.78	94.65	93.28
21-30	-	95.74	94.54	93.77
31-40	-	-	95.12	94.32
41-50	-	-	-	94.64
Verication Rate				
CAAE [34]	87.05	81.07	73.36	60.25
IPCGAN [31]	**100**	**100**	**100**	**100**
ours	**100**	**100**	**100**	**100**

UNIVERSITY of GUELPH

Talking Face Generation

- Introduction
- Overview
- Task 1
- Task 2
- Task 3
- Task 4
- Summary

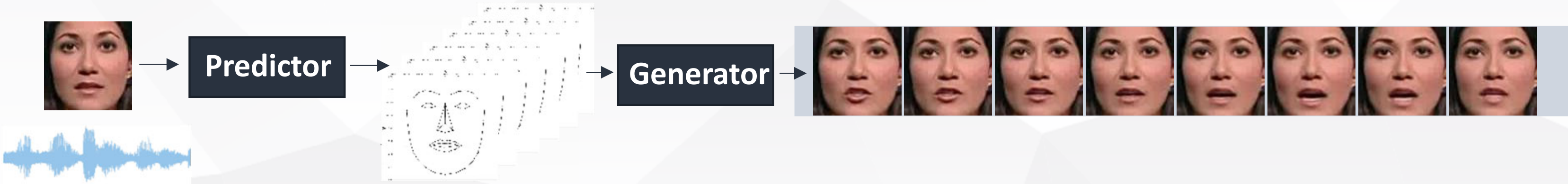


Given an audio clip and an arbitrary face image, automatically produce a talking face video with lip movements synchronizing with the input audio.

Previous Work

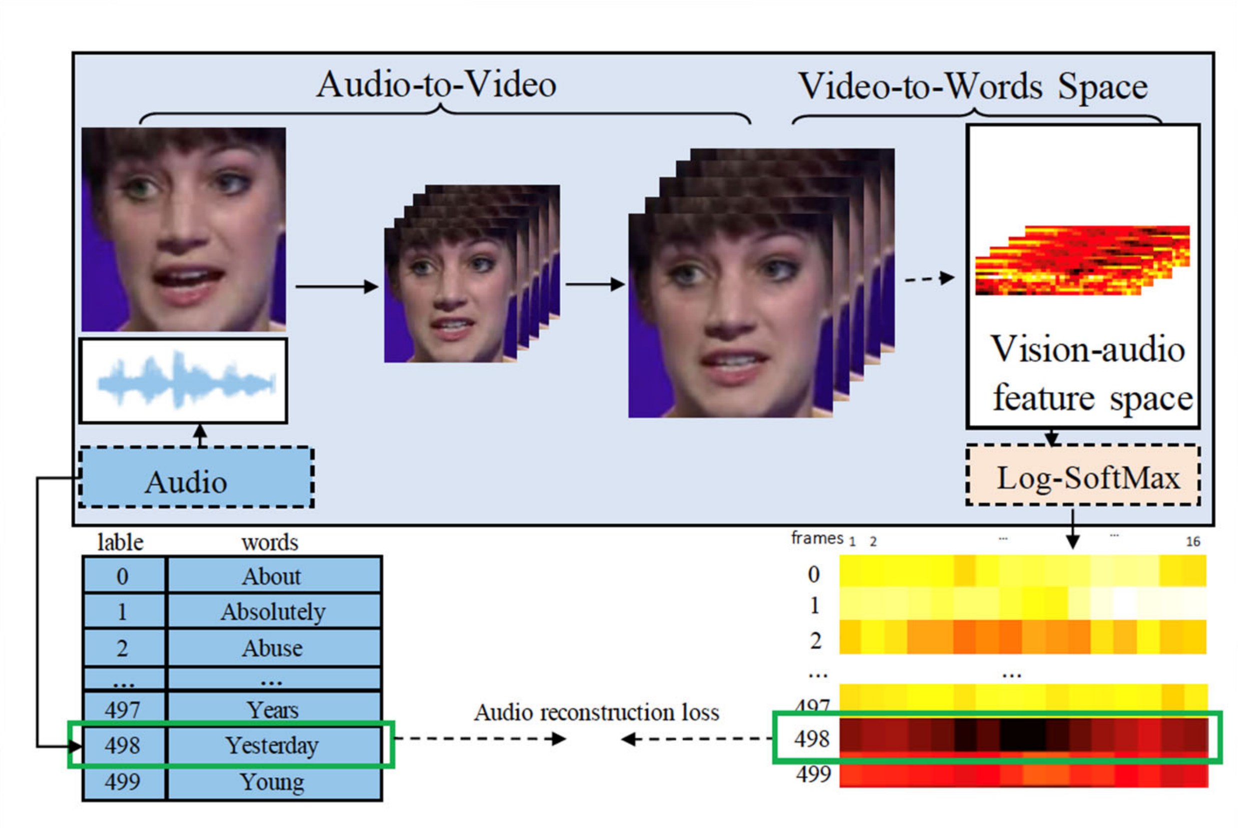
Audio-face[1][2]

Audio-landmark-face[3]



[1] Vougioukas, Konstantinos, Stavros Petridis, and Maja Pantic. "End-to-end speech-driven facial animation with temporal gans." *arXiv preprint arXiv:1805.09313* (2018).
[2] Chung, Joon Son, Amir Jamaludin, and Andrew Zisserman. "You said that?." arXiv preprint arXiv:1705.02966 (2017).
[3] Chen, Lele, et al. "Hierarchical cross-modal talking face generation with dynamic pixel-wise loss." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

Motivation

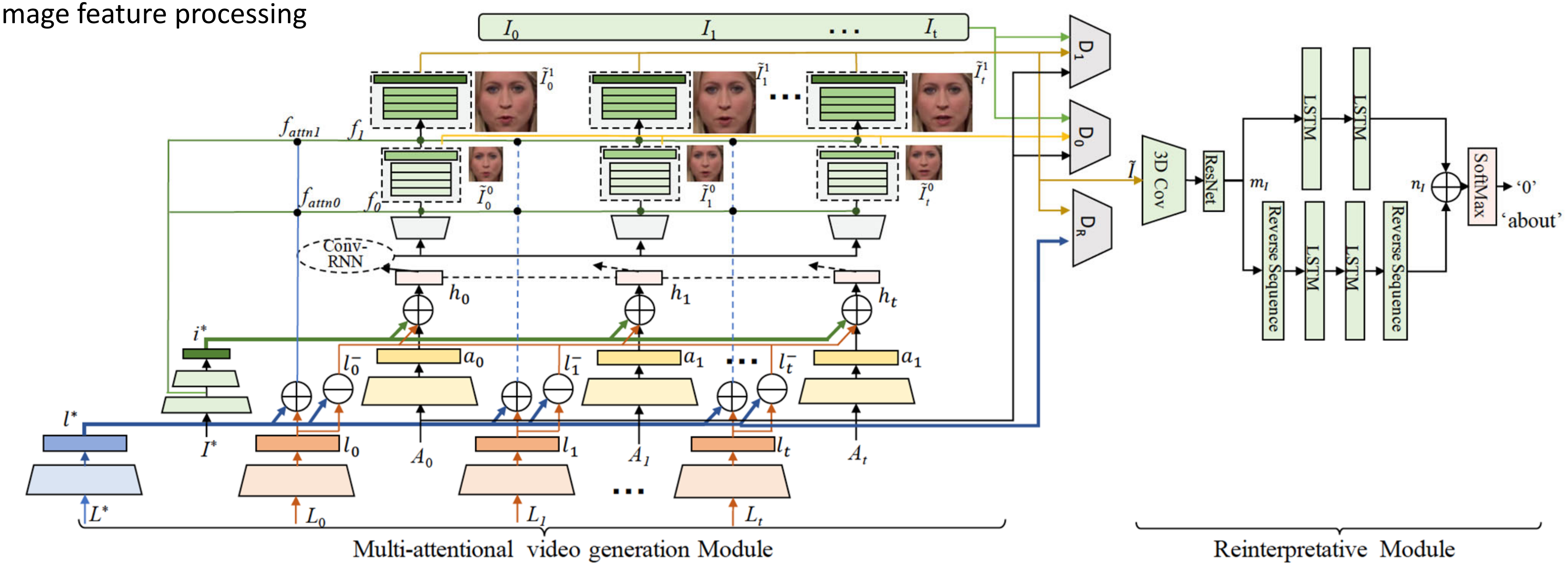


Generate **high-resolution** talking face videos that are:

1) **synchronous** with the input audio
2) Maintain **visual details** from the input face images

Audio-to-video-to-words Network

Inputs:
Audio feature processing
Landmark feature processing
Image feature processing



Fine-Grained Generation with Coarse-to-fine GAN

Introduction

Overview

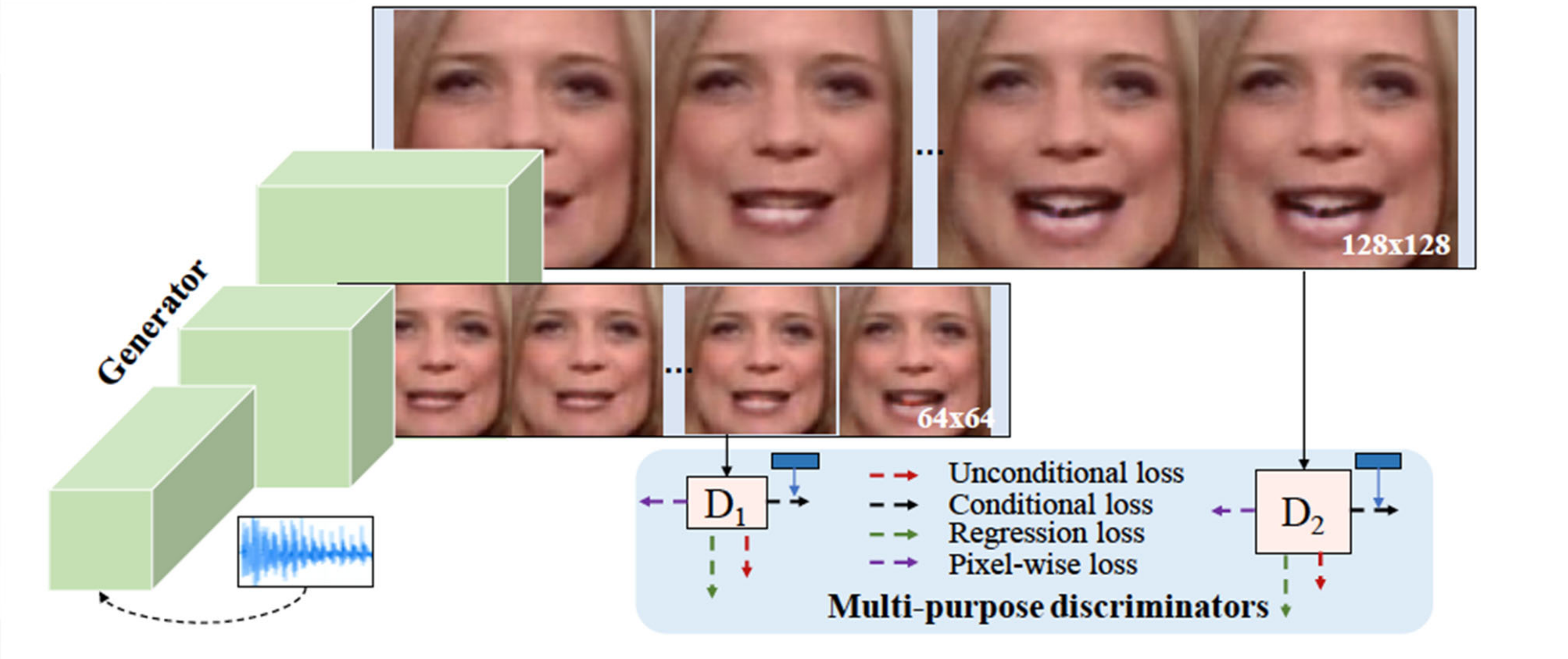
Task 1

Task 2

Task 3

Task 4

Summary



UNIVERSITY of GUELPH

Experiments

Introduction
Overview
Task 1
Task 2
Task 3
Task 4
Summary

Datasets

Training: LRW
500 words
In each word class, there are 1,000 training video samples, 50 test samples and 50 validation samples.

Test: LRW and GRID
GRID contains 33 speakers, each uttering 1,000 short phrases

Evaluation Metrics

Structural Similarity Index (SSIM)
Peak Signal-to-Noise Ratio (PSNR)
Landmark Distance Error (LMD)

Nvidia GeForce GTX 1080 Ti
50GiB memory
Batch normalization
Learning rate: 0.0002
Adam algorithm

UNIVERSITY of GUELPH

Talking Face Results

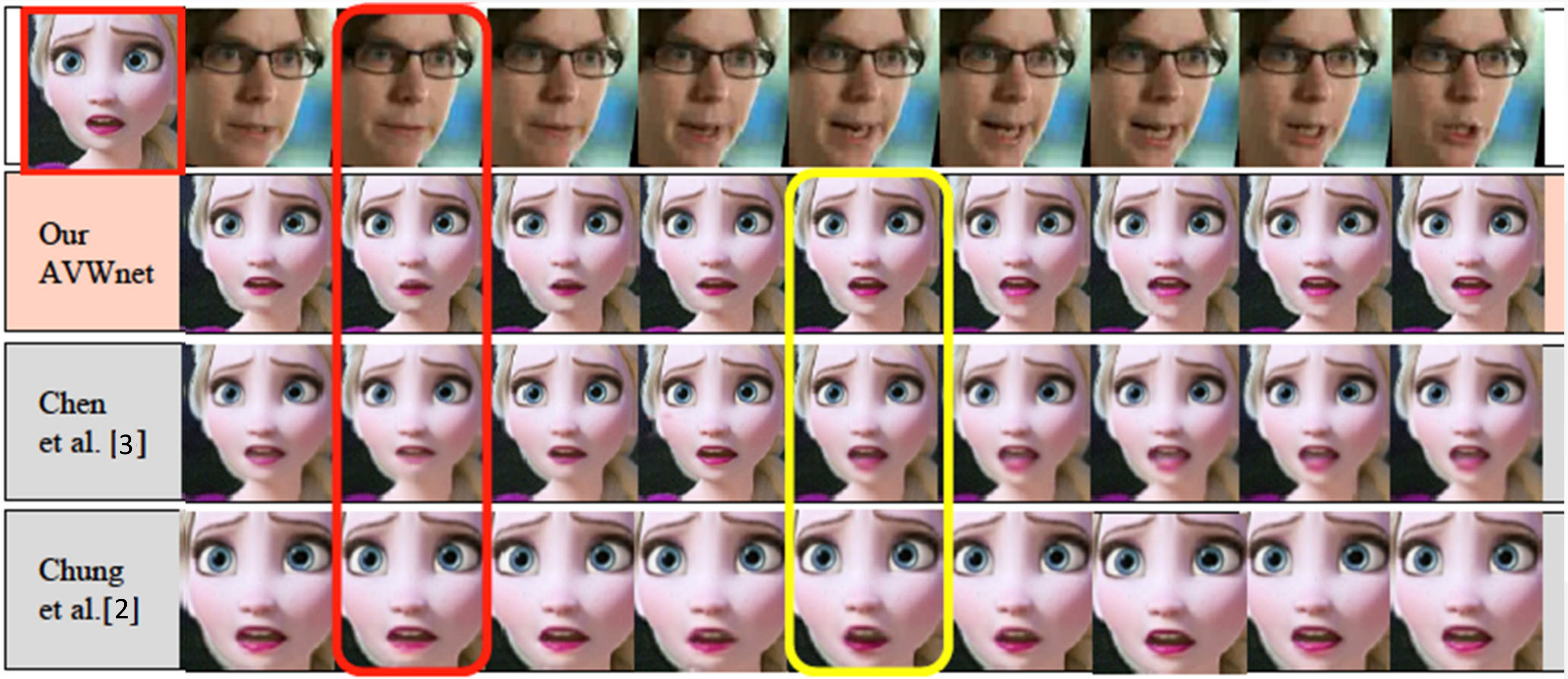
UNIVERSITY of GUELPH

Two Layers of Attention Maps

- Introduction
- Overview
- Task 1
- Task 2
- Task 3
- Task 4
- Summary

Attention masks generated for the coarse (top) and fine (middle) levels and the resulted image frames (bottom).

Qualitative Comparison



Detailed Comparison

- Introduction
- Overview
- Task 1
- Task 2
- Task 3
- Task 4
- Summary

Our AVWnet Chen et al. [3]

Our AVWnet Chen et al. [3]

Quantitative Evaluation

Introduction

Overview

Task 1

Task 2

Task 3

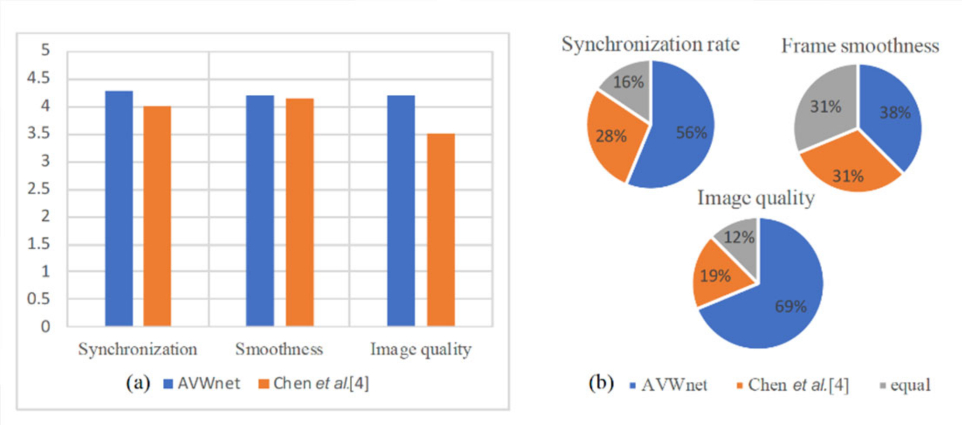
Task 4

Summary

Quantitative evaluation on LRW and GRID testing datasets. Best scores are shown in boldface.

Method	LRW			GRID		
	SSIM	PSNR	LMD	SSIM	PSNR	LMD
Chung [5]	0.71	28.31	3.19	0.74	28.46	3.03
Chen [4]	0.75	30.04	2.97	0.77	31.61	2.88
Our AVWnet	**0.82**	**31.24**	**2.84**	**0.84**	**32.03**	**2.79**

SSIM: Structural Similarity Index
PSNR: Peak Signal-to-Noise Ratio
LMD:Landmark Distance Error

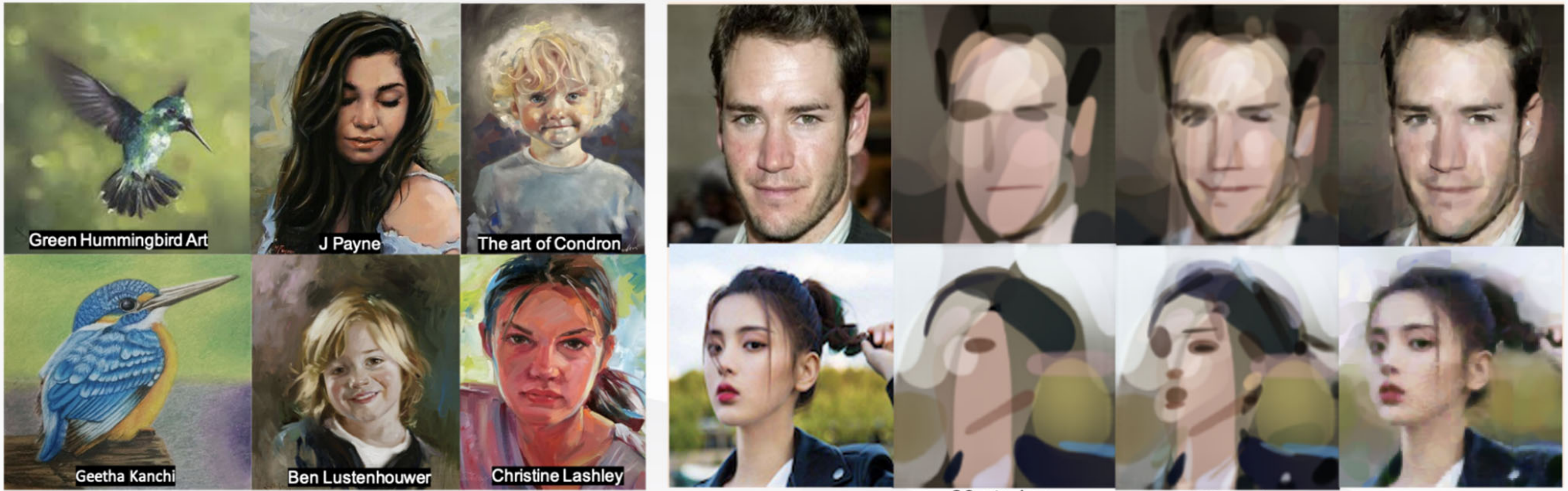


User study on videos generated using the proposed AVWnet and the state of-the-art method.

UNIVERSITY of GUELPH

Neural Painting

Introduction
Overview
Task 1
Task 2
Task 3
Task 4
Summary



20 strokes 100 strokes 200 strokes

(a) (b)

Given an input image, generate a sequence of paining stokes that can be applied to a blank canvas in a stroke-by-stroke manner and reproduce the designed input.

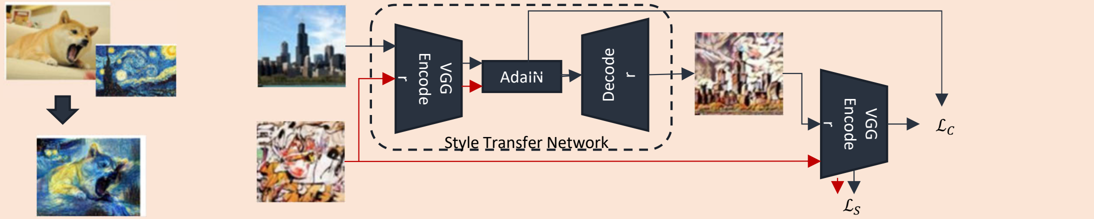
UNIVERSITY of GUELPH

Previous Work

- Introduction
- Overview
- Task 1
- Task 2
- Task 3
- **Task 4**
- Summary

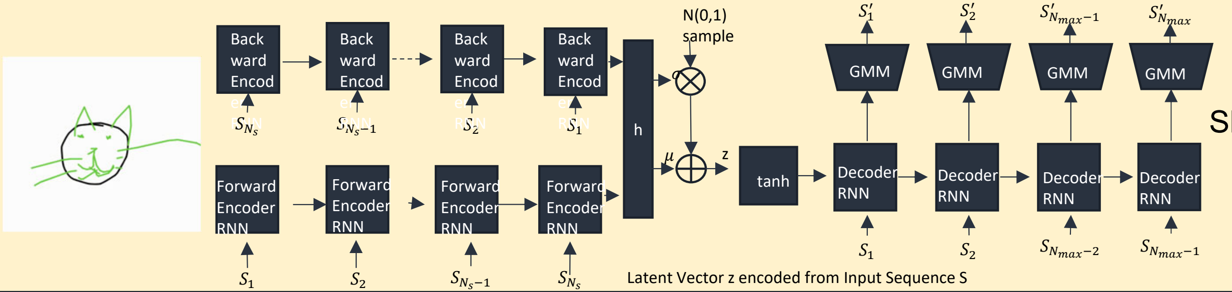
Neural painting

Style transfer[1]



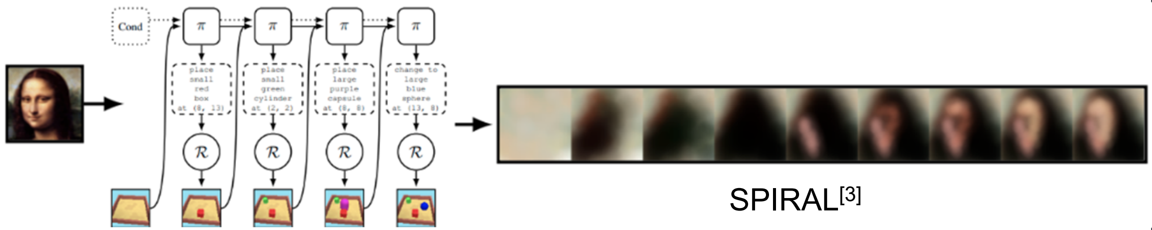
Arbitrary-Style-Transfer[1]

RNN[2]



Sketch-rnn[2]

Deep Reinforcement Learning[3]



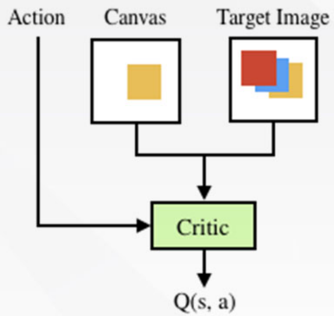
SPIRAL[3]

[1] Huang, Xun, and Serge Belongie. "Arbitrary style transfer in real-time with adaptive instance normalization." Proceedings of the IEEE International Conference on Computer Vision. 2017.
[2] Ha, David, and Douglas Eck. "A neural representation of sketch drawings." arXiv preprint arXiv:1704.03477 (2017).
[3] Ganin, Yaroslav, et al. "Synthesizing programs for images using reinforced adversarial learning." International Conference on Machine Learning. PMLR, 2018.

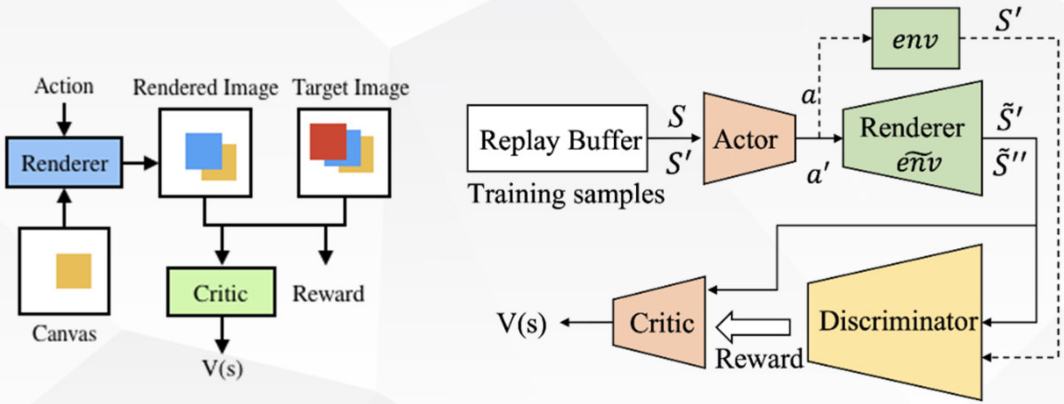
Previous Work

Introduction
Overview
Task 1
Task 2
Task 3
Task 4
Summary

Deep Reinforcement Learning[3]



Original Deep Deterministic Policy Gradients (DDPG)



Model-based DDPG RL algorithms for painting [4]

[4] Z. Huang, W. Heng, and S. Zhou. Learning to paint with model-based deep reinforcement learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 8709–8718, 2019.

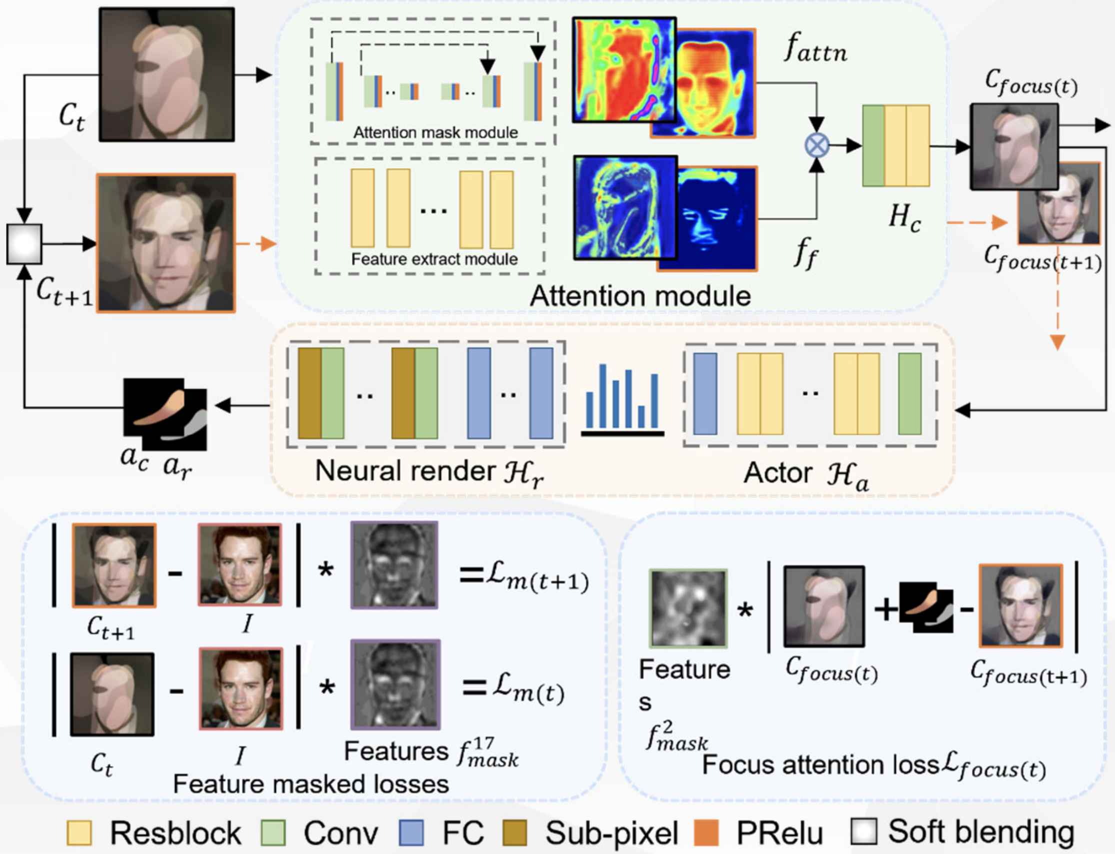
Motivation



Better mimics the painting process used by **human artists**:

1) Pay more attention to **foreground content** rather background details.
2) Paint **important** regions with finer brushes.

Attention-Aware Painting via Deep Reinforcement Learning

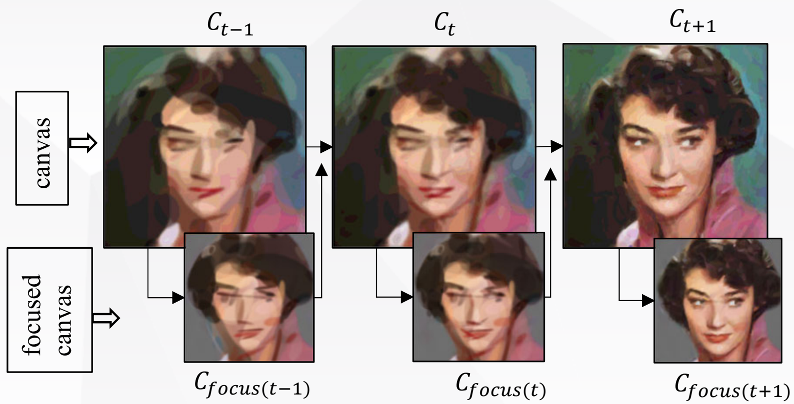


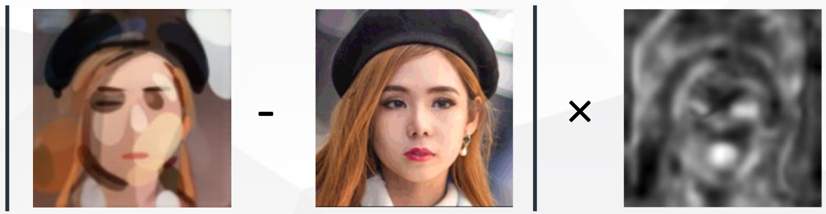
Attention & Feature Mask

- Introduction
- Overview
- Task 1
- Task 2
- Task 3
- Task 4
- Summary

Attention network:
Prioritize areas that contain high saliency foreground subjects.

Feature mask:
Act as a weight for pixel distances between paintings and the target image
Guide the agent to accurately capture the appearances of important and recognition-related regions.





Feature masked loss

UNIVERSITY of GUELPH

Neural Painting Results

Introduction

Overview

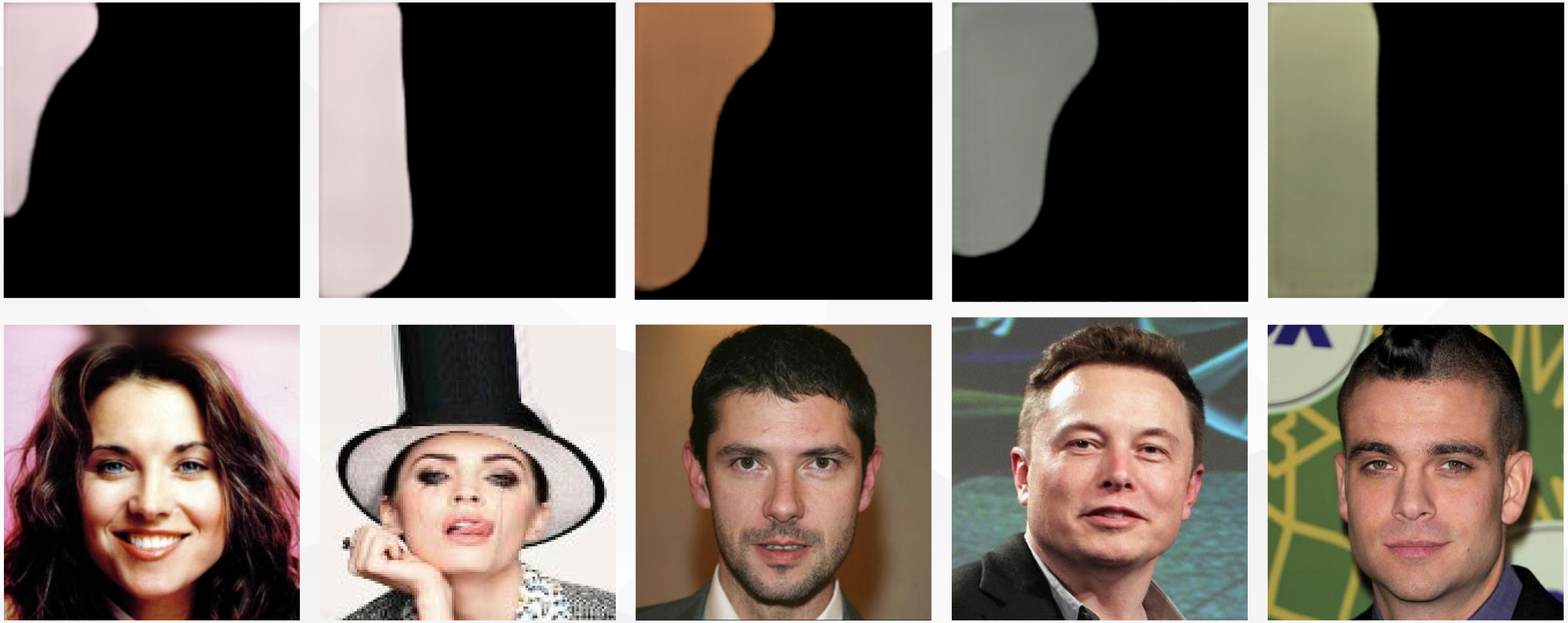
Task 1

Task 2

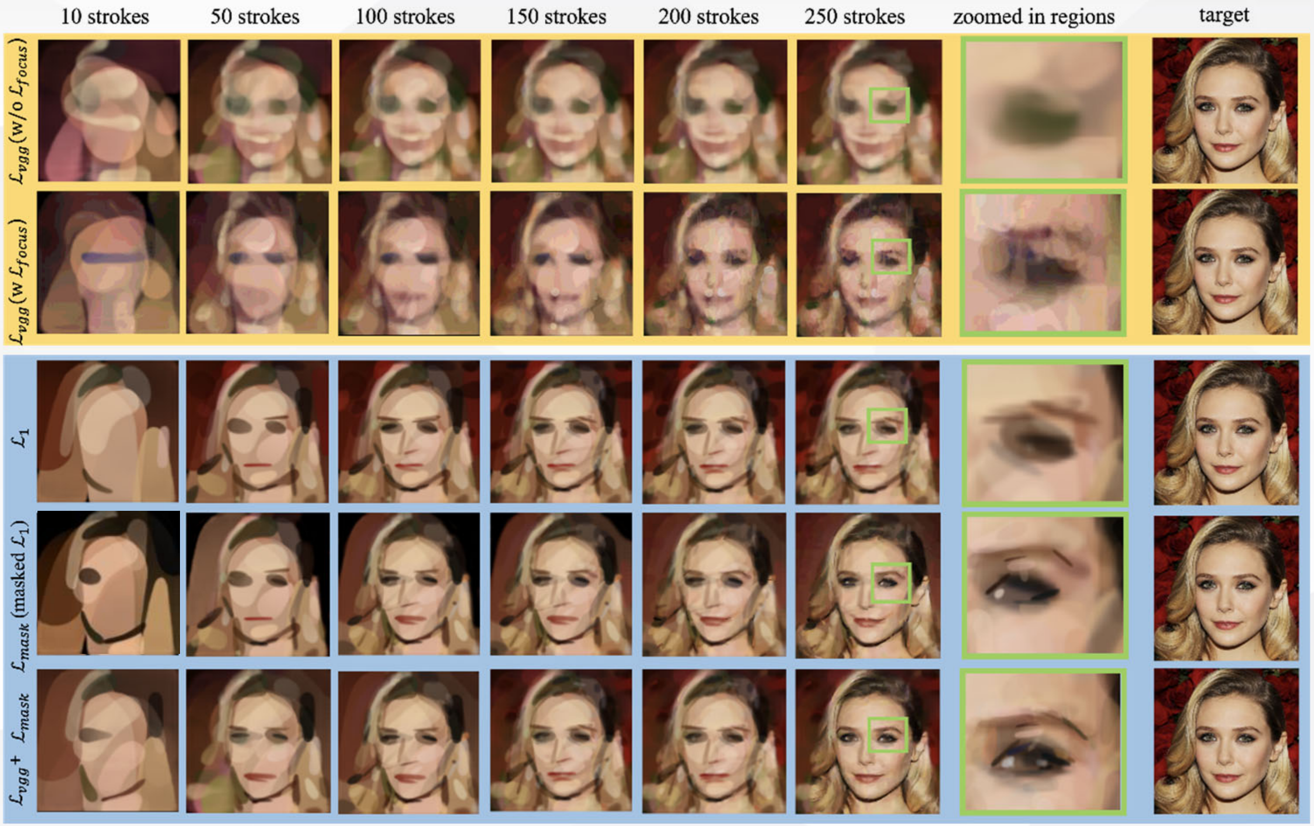
Task 3

Task 4

Summary



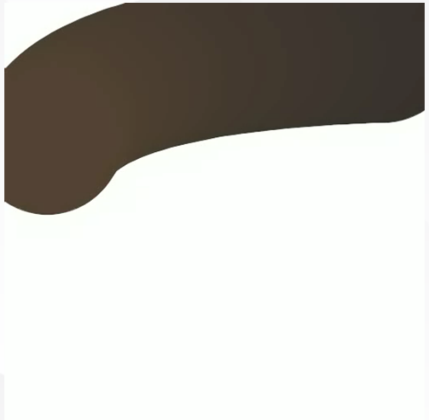
Ablation Study



Qualitative Comparison

Proposed method

Huang et al[4]

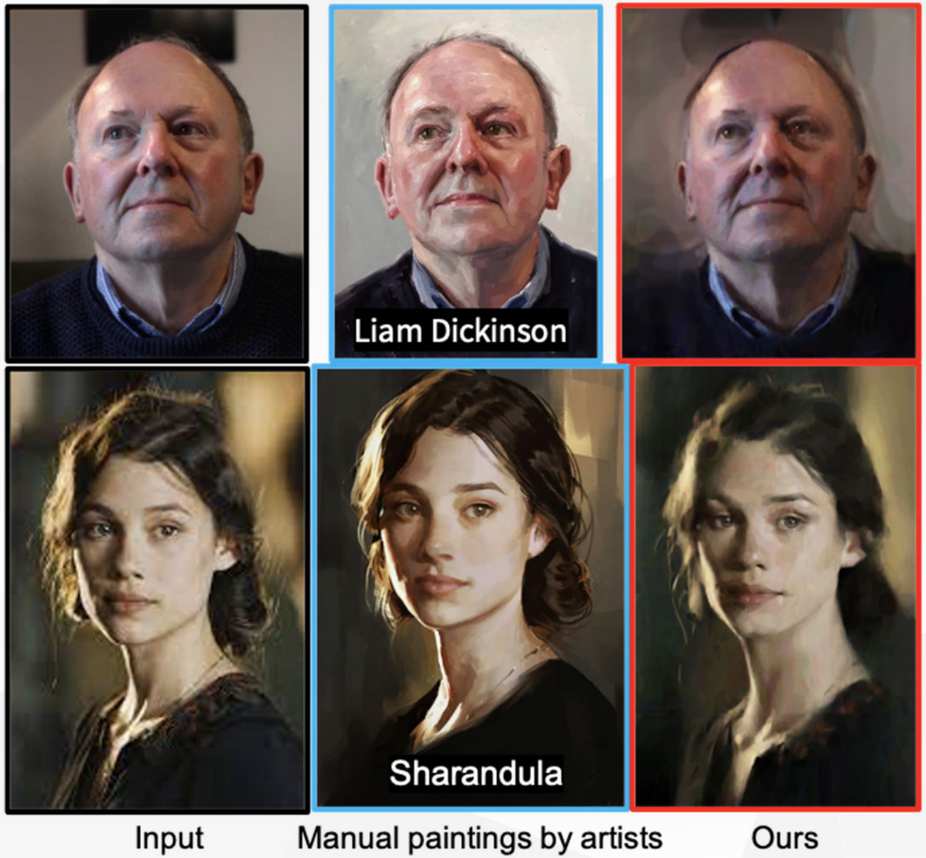


Zou et al[5]

Target

[4] Z. Huang, W. Heng, and S. Zhou. Learning to paint with model-based deep reinforcement learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 8709–8718, 2019.
[5] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE international conference on computer vision, pages 2223–2232, 2017.

Comparison with Human Painting



UNIVERSITY of GUELPH

- Introduction
- Overview
- Task 1
- Task 2
- Task 3
- Task 4
- Summary

Conclusions

- Approaches were proposed to address four different visual synthesis tasks:
  - HfGAN is better in generating consistent and high-quality images because the hierarchical feature maps' fusion can fully extract and utilize the local and global features.
  - LGcGAN the transition pattern at different ages and performs well in preserving personal identity and keeping face-aging consistency.
  - AVWnet can generate talking face videos with better audio-lip consistency and higher frame quality.
  - The proposed end-to-end attention-aware reinforcement learning approach for painting like humans better approximates the target image under small number of strokes and capture finer foreground details in the final results.

UNIVERSITY of GUELPH

Related Publications

- Introduction
- Overview
- Task 1
- Task 2
- Task 3
- Task 4
- Summary

- Xin Huang, Mingjie Wang, & Minglun Gong: Hierarchically-fused generative adversarial network for text to realistic image synthesis. *Conference on Computer and Robot Vision*. Kingston, ON, Canada, May 29-31, 2019. (**Best Paper Award**)
- Xin Huang, Mingjie Wang, & Minglun Gong: Fine-grained talking face generation with video reinterpretation. *The Visual Computer*. October 2020.
- Xin Huang & Minglun Gong: Landmark-guided conditional GANs for face aging. *International Conference on Image Analysis and Processing*. Lecce, Italy, May 23-27, 2022.
- Huang, Xin & Minglun Gong. Attention-Aware Neural Painting via Deep Reinforcement Learning. *Neurocomputing* (under review).

- **Abstract:**
- Conditional visual synthesis is the process of artificially generating images or videos that satisfy desired constraints. Individual visual synthesis tasks, such as high-fidelity natural image generation, artwork creation, and face animation, have many real-world applications. With advances in deep learning, methods for conditional visual synthesis evolve rapidly in recent years, making it one of the hottest research fields in Computer Vision and Graphics. Many of these recent approaches are based on Generative Adversarial Networks (GANs), which has a strong ability to generate samples following almost any implicit distribution, allowing the synthesis of visual content in an unconditional or input-conditional manner. However, GANs still have many limitations, such as difficulty in directly approximating high-resolution image distributions, poor model generalization ability on unpaired datasets, and limited power for mimicking human actions. This talk introduces efforts for tackling these limitations and for handling different conditional visual synthesis tasks.
- The first task is the generation of high-resolution images that are conditioned by text inputs. A novel end-to-end hierarchically-fused GAN is developed, which trains only one generator-discriminator pair to synthesize images from coarse to fine resolution levels. The second task is to simulate facial changes based on desired ages. Facial landmarks are extracted to guide the synthesis and a symmetric framework is employed to enhance both age and identity consistency. The third task aims to synthesize realistic talking face videos that are conditioned by audio inputs. A coarse-to-fine tree-like architecture is designed, which not only ensures synchronization with input audios but also maintains visual details from input face photos. The objective of the final task is to generate painting stroke sequences that can recreate input images. An attention-aware end-to-end deep reinforcement learning framework is developed to better imitate human painting actions. Both qualitative and quantitative validation experiments are conducted for each proposed methods. Comparisons with existing works demonstrate the respective merits of these techniques.