Attention GAN v2

AttentionGAN: Unpaired Image-to-Image Translation using
Attention-Guided Generative Adversarial Networks

석사과정 김 진용

Introduction

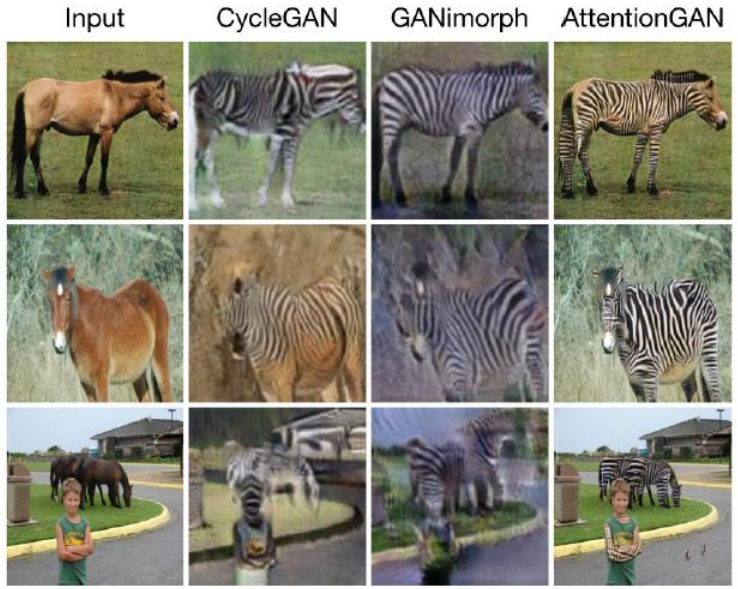


Fig. 1: Comparison with existing image-to-image translation methods (e.g., CycleGAN [3] and GANimorph [6]) with an example of horse to zebra translation. We are interest in transforming horses to zebras. In this case we should be agnostic to the background. However methods such as CycleGAN and GANimorph will transform the background in a nonsensical way, in contrast to our attention-based method.

Image-to-Image translation에서는 흔히 발생되는 "Unwanted part에 대한 변화" 를 다루고자 한다

ConrastGAN

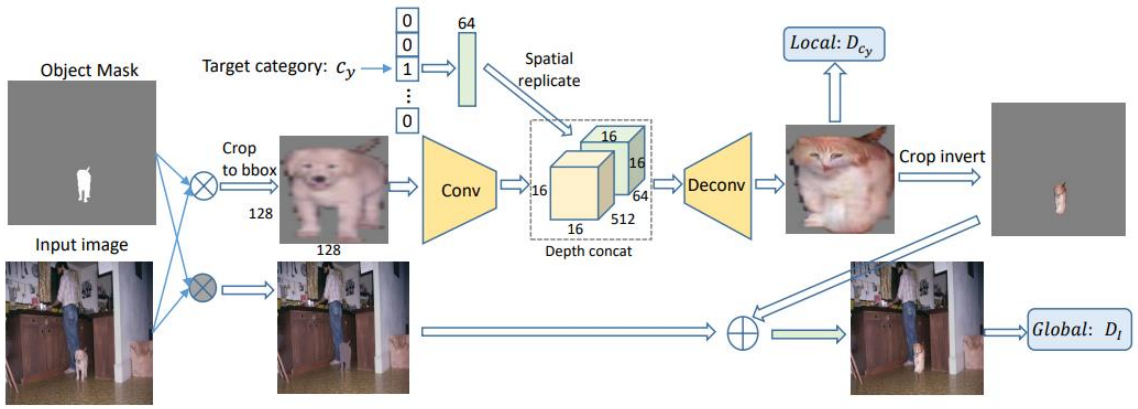
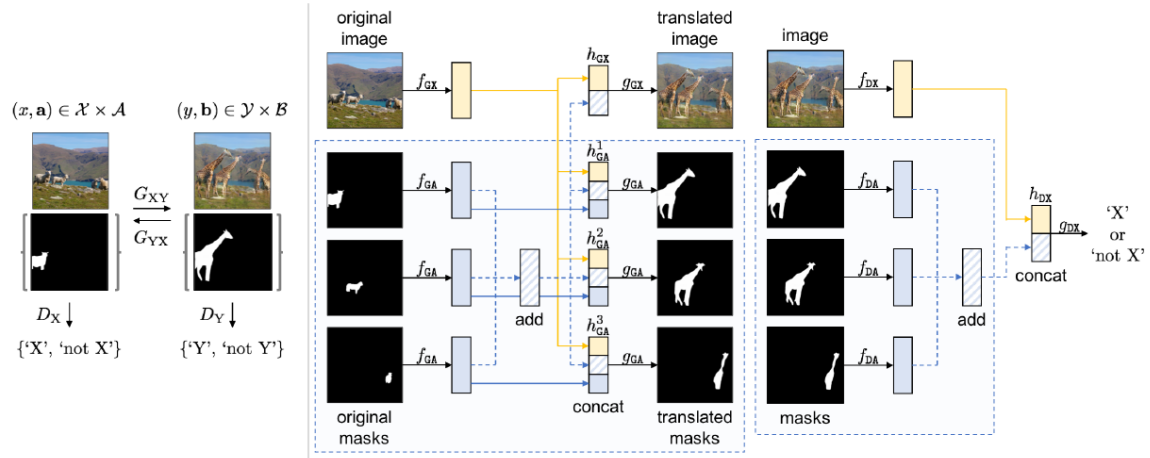


Figure 3: The proposed mask-conditional contrast-GAN for semantic manipulation by taking an input image, an object mask and a target category as input. Please refer more details in Section 3.2.

https://arxiv.org/pdf/1708.00315.pdf

InstaGAN



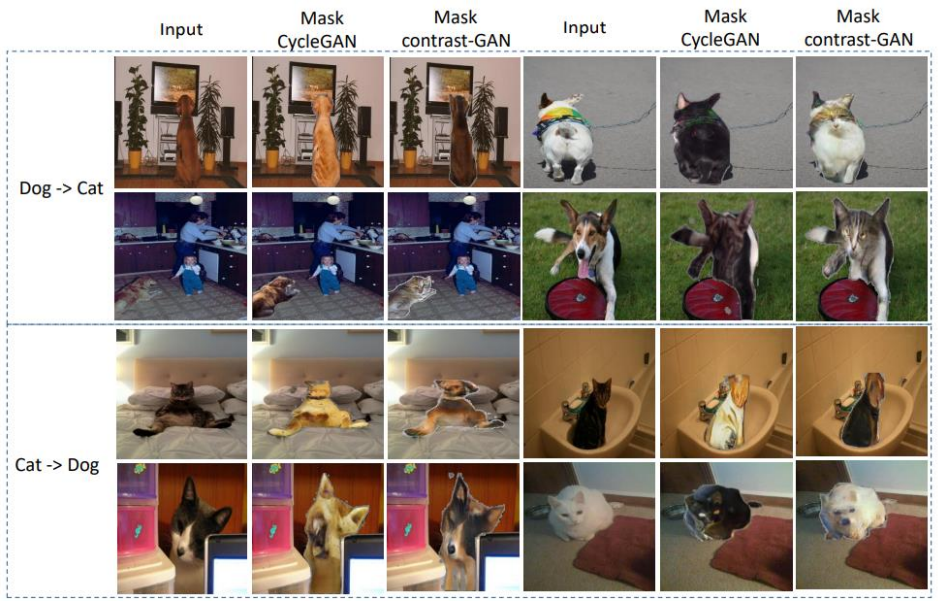
(a) Overview (b) Generator G (c) Discriminator D

Figure 2: (a) Overview of InstaGAN, where generators G_{XY}, G_{YX} and discriminator D_X, D_Y follows the architectures in (b) and (c), respectively. Each network is designed to encode both an image and set of instance masks. G is permutation equivariant, and D is permutation invariant to the set order. To achieve properties, we sum features of all set elements for invariance, and then concatenate it with the identity mapping for equivariance.

https://arxiv.org/abs/1812.10889

이런 문제를 해결하기위한 기존연구들은 annotation masking을 해서 실행하는 방식

ConrastGAN



https://arxiv.org/pdf/1708.00315.pdf

InstaGAN

Figure 7: Results of InstaGAN varying over different input masks.

Figure 8: Translation results on CCP dataset, using predicted mask for inference.

https://arxiv.org/abs/1812.10889

이런 문제를 해결하기위한 기존연구들은 annotation masking을 해서 실행하는 방식
하지만 비용적인 측면에서 training data를 구축하는데 적지않은 비용이 들어감

Attention GAN v1(Scheme I)

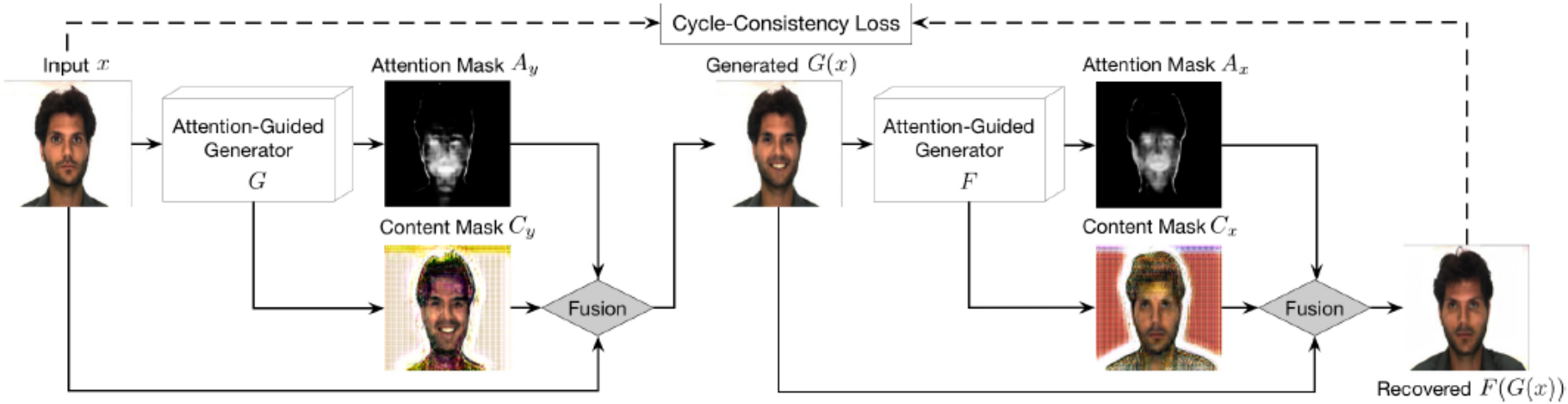


Fig. 2: Framework of the proposed attention-guided generation scheme I, which contains two attention-guided generators G and F. We show one mapping in this figure, i.e., $x \rightarrow G(x) \rightarrow F(G(x)) \approx x$. We also have the other mapping, i.e., $y \rightarrow F(y) \rightarrow G(F(y)) \approx y$. The attention-guided generators have a built-in attention module, which can perceive the most discriminative content between the source and target domains. We fuse the input image, the content mask and the attention mask to synthesize the final result.

이런 문제를 해결하기위해 위와 같은 Architecture를 제안 **했었음**

- Attention-Guided Generator에 통합된 built-in(내장) attention module을 통해 attention mask 생성
- 또한 generator에서 disentanglement를 통해 attention mask 뿐만 아니라 content mask로 분리하여 사용

>> 이 효과로 Unwanted-part에 대한 translation을 막을 수 있다고 생각했음

Attention GAN v1(Scheme I)

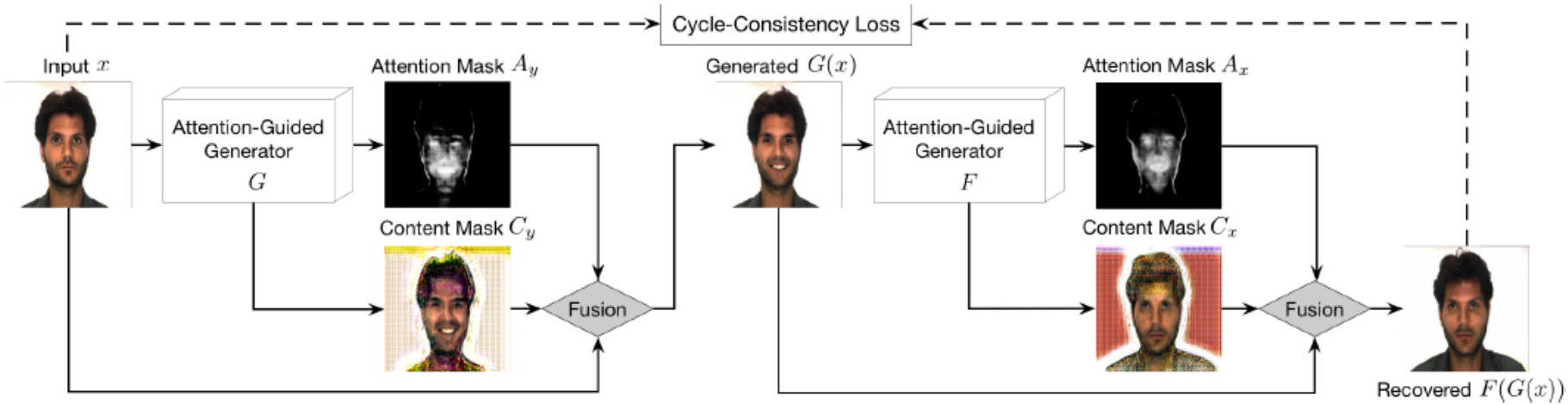


Fig. 2: Framework of the proposed attention-guided generation scheme I, which contains two attention-guided generators G and F. We show one mapping in this figure, i.e., $x \rightarrow G(x) \rightarrow F(G(x)) \approx x$. We also have the other mapping, i.e., $y \rightarrow F(y) \rightarrow G(F(y)) \approx y$. The attention-guided generators have a built-in attention module, which can perceive the most discriminative content between the source and target domains. We fuse the input image, the content mask and the attention mask to synthesize the final result.

하지만, 사람 얼굴과 같이 fore/background가 단순히 분리되는 것 말고
complex sementic translation에서는 매우 불안정했다.
그리하여 3개의 문제점을 파악해서 새 네트워크를 제안하려고 한다.

1. Attention mask와 Content mask가 출력되는 네트워크는 공통 된 네트워크라는 점
2. 기존 v1에서는 foregroun만을 위한 mask를 생성했고, 이는 backgroun를 preserve하는데 문제가 된다.
3. Contents mask와 foreground attention mask의 생성을 다방면으로 늘려 학습하지 않으면 복잡한 이미지 해결이 힘들다

Proposed model

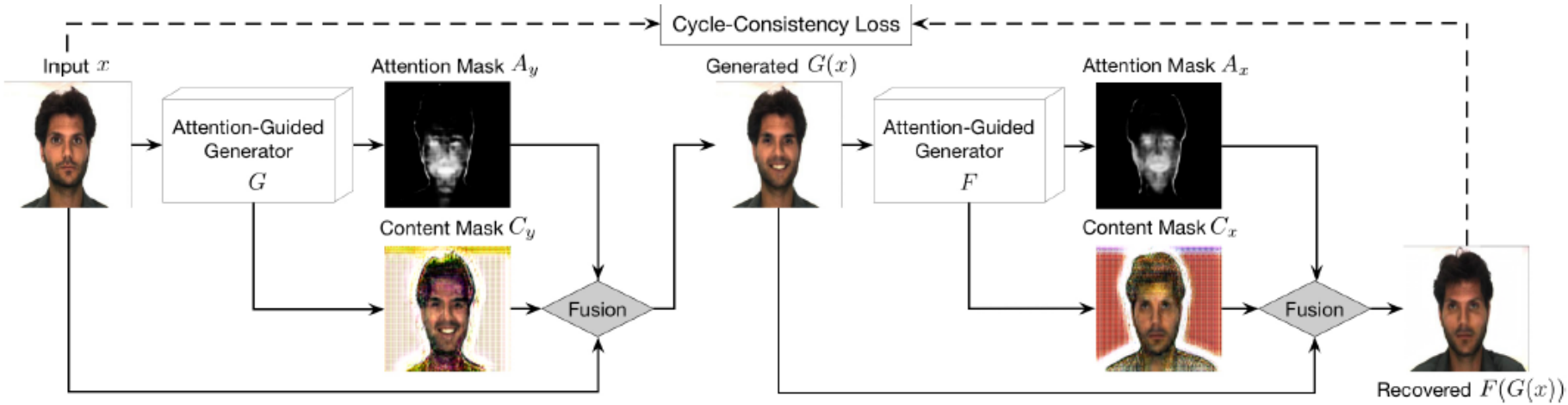


Fig. 2: Framework of the proposed attention-guided generation scheme I, which contains two attention-guided generators G and F. We show one mapping in this figure, i.e., $x \to G(x) \to F(G(x)) \approx x$. We also have the other mapping, i.e., $y \to F(y) \to G(F(y)) \approx y$. The attention-guided generators have a built-in attention module, which can perceive the most discriminative content between the source and target domains. We fuse the input image, the content mask and the attention mask to synthesize the final result.

기존 방식에 대한 간단한 설명

- 두 가지의 매핑을 진행함

$$G: x \to [A_y, C_y] \to G(x) \qquad F: y \to [A_x, C_x] \to F(y),$$

- A_x A_y는 C_x와 C_y와 결합되어 pixel당 어느정도의 강도를 줄지에 대한 결정을 내리는 수단이 됨

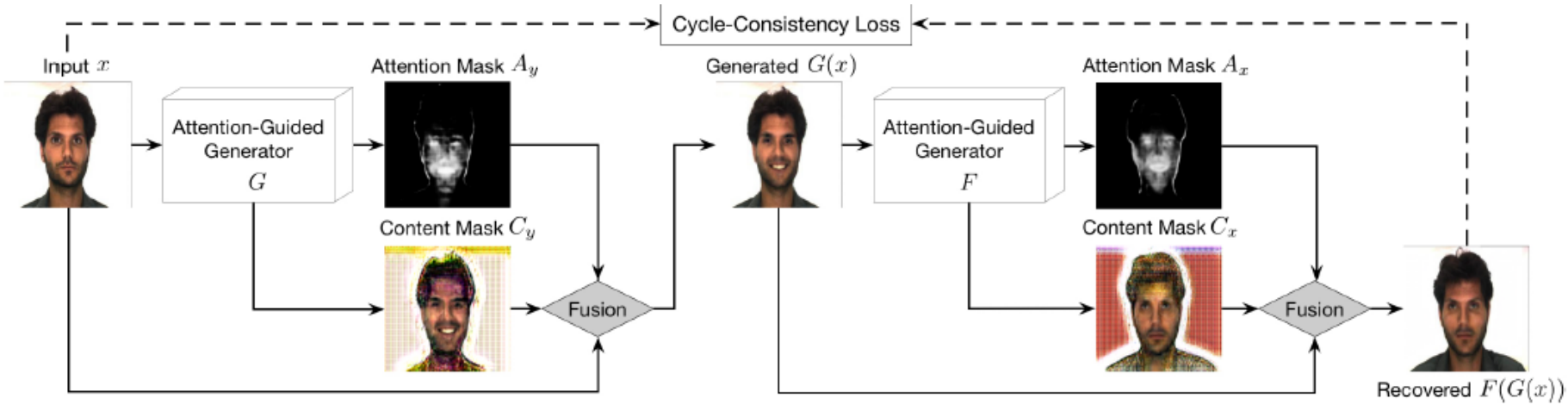


Fig. 2: Framework of the proposed attention-guided generation scheme I, which contains two attention-guided generators G and F. We show one mapping in this figure, i.e., $x \rightarrow G(x) \rightarrow F(G(x)) \approx x$. We also have the other mapping, i.e., $y \rightarrow F(y) \rightarrow G(F(y)) \approx y$. The attention-guided generators have a built-in attention module, which can perceive the most discriminative content between the source and target domains. We fuse the input image, the content mask and the attention mask to synthesize the final result.

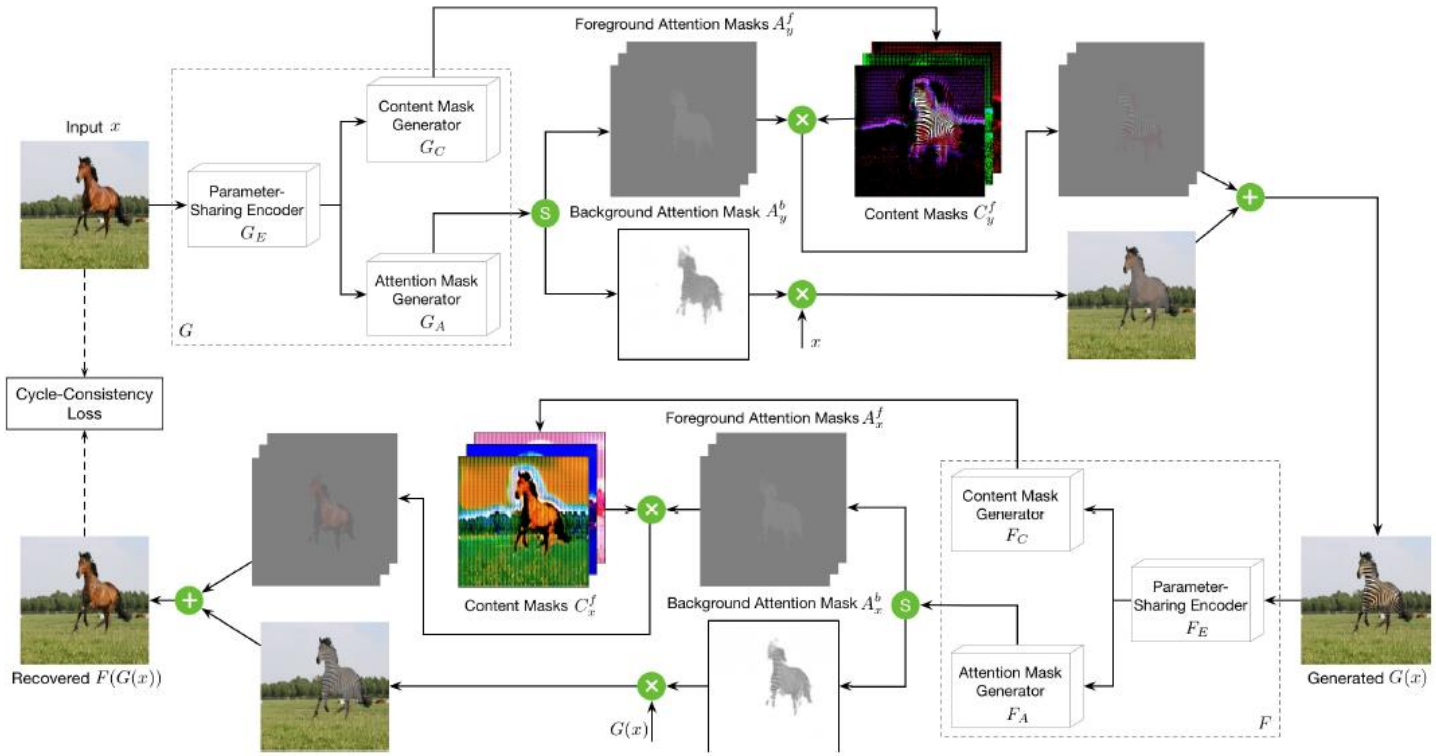
기존 방식에 대한 간단한 설명

- 이렇게 되면 집중해야 할 특징점(눈, 입 등 표정을 주는 부분)에 집중함.
- 반대로 집중하지 않아도 될 부분(머리카락, 옷 등)은 그대로 감

$$G(x) = C_y * A_y + x * (1 - A_y),$$

- Contents mask에는 집중해야 할 부분에 대한 Attention
- real data에는 집중하지 않아도 될 부분에 대한 Attention

Attnetion GAN v2



Attnetion GAN v1의 문제점 3가지

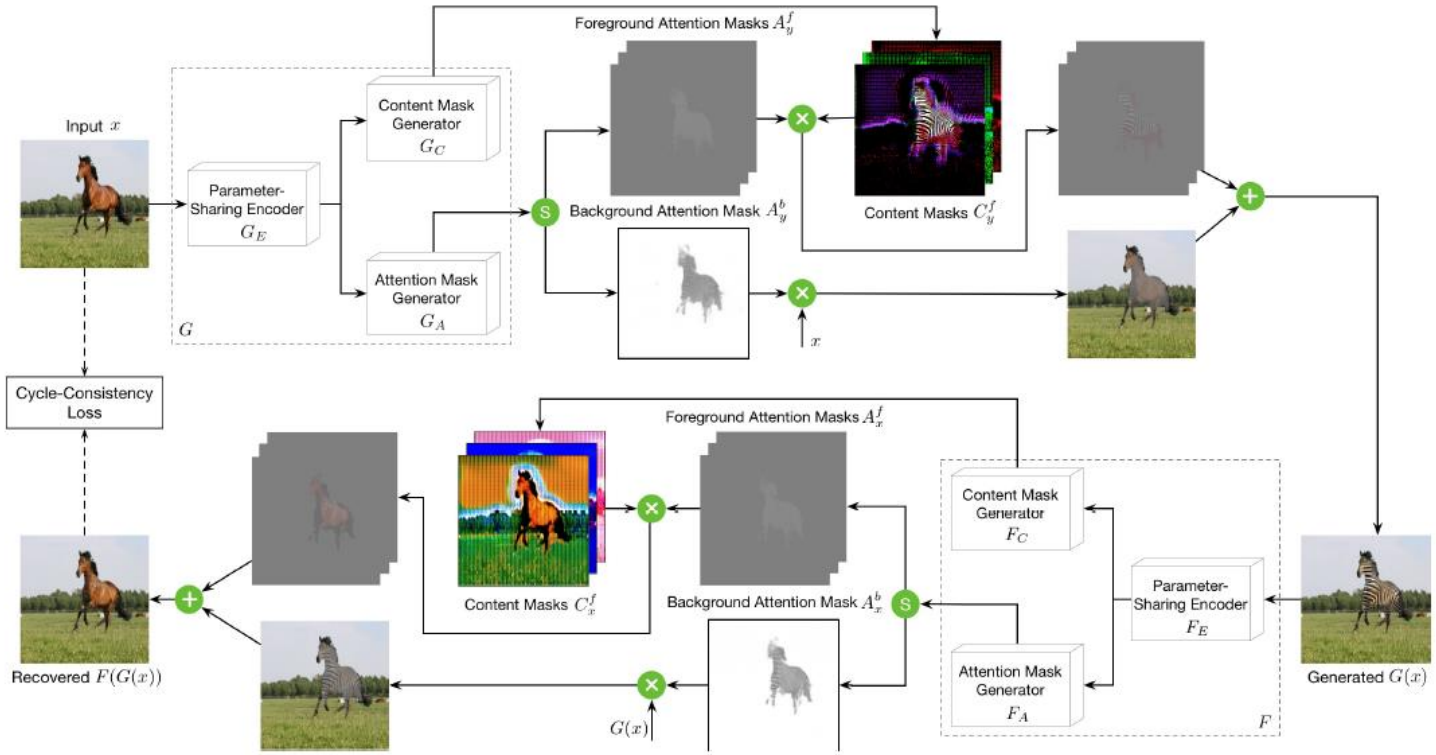
1. Attention mask와 Content mask가 출력되는 네트워크는 공통 된 네트워크라는 점

2. 기존 v1에서는 foregroun만을 위한 mask를 생성했고, 이는 backgroun를 preserve하는데 문제가 된다.

3. Contents mask와 foreground attention mask의 생성을 다방면으로 늘려 학습하지 않으면 복잡한 이미지 해결이 힘들다

Attnetion GAN v2

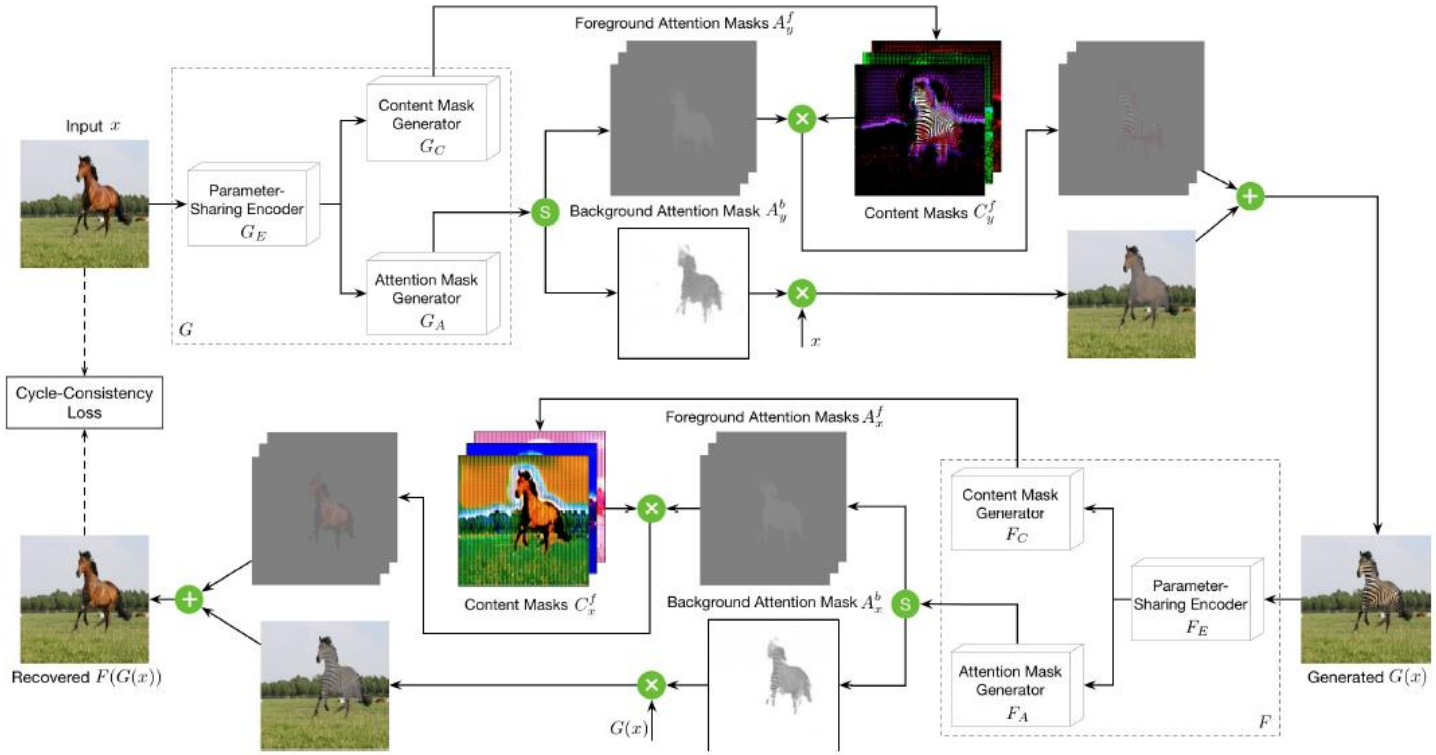
1. Attention mask와 Content mask가 출력되는 네트워크는 공통 된 네트워크라는 점



- Parameter Sharing Encoder G_E를 통해 high-level, low-level의 feature들을 각 G_C와 G_A에 공급
- 기존 1개의 generator에서 만들어냈던 것 과는 달리 각 네트워크의 목표에 집중할 수 있게 되어 more powerful

Attnetion GAN v2

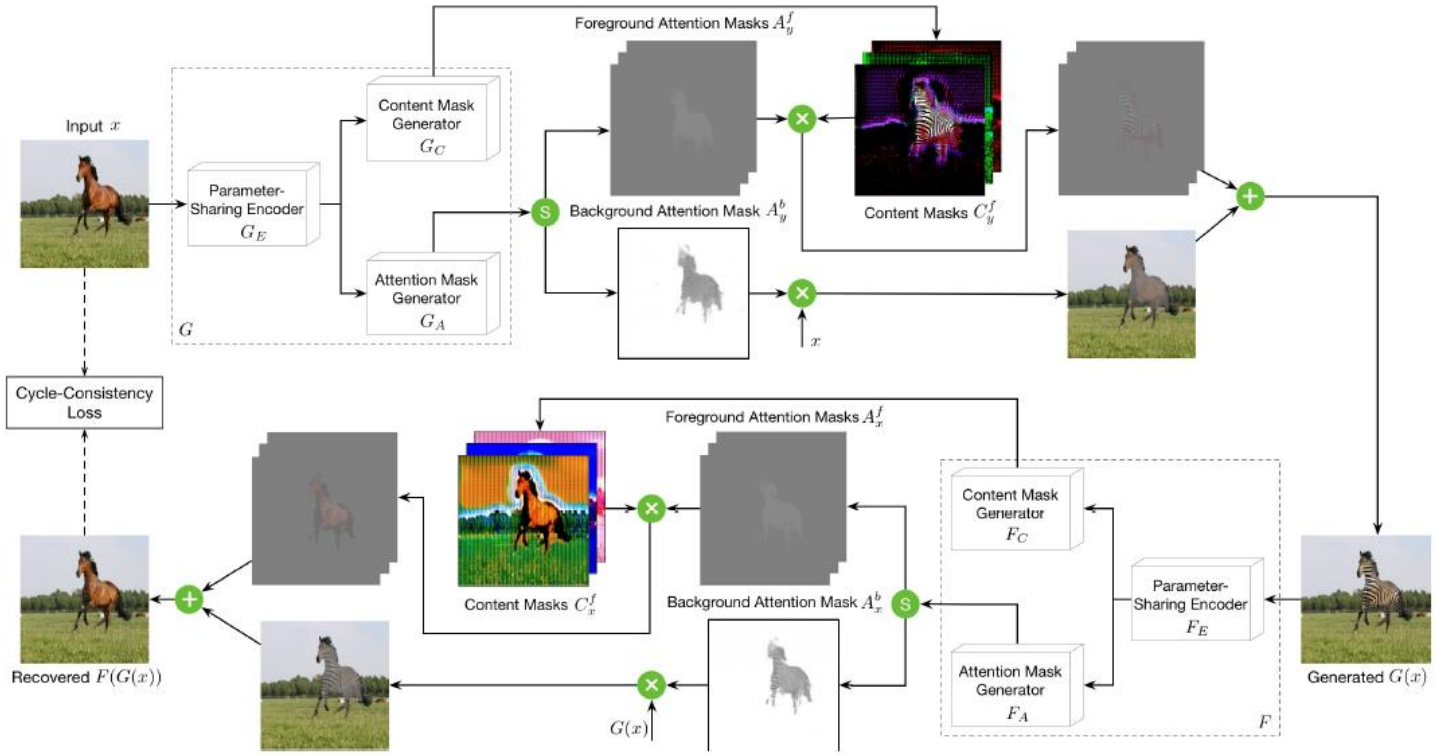
2. 기존 v1에서는 foregroun만을 위한 mask를 생성했고, 이는 backgroun를 preserve하는데 문제가 된다.



- G_A에서는 n-1개의 foreground mask를 만들었고 1개의 background mask를 만든다.
- 이는 네트워크가 foreground를 학습함과 동시에 이미지의 배경을 보존할 수 있다는 것.

Attnetion GAN v2

3. Contents mask와 foreground attention mask의 생성을 다방면으로 늘려 학습하지 않으면 복잡한 이미지 해결이 힘들다



- Content mask는 G_c에서 n-1개와 real image x를 추가하여 총 n개를 통해 중간 content mask를 생성한다.
- 이렇게 되면 3-channel generation space가 3n-channel generation space로 확대됨.
- 복잡한 이미지를 mapping하기에 적절해진다.

Attnetion GAN v2

- 결과물을 식으로 나타내자면 다음과 같음

$$G(x) = C_y * A_y + x * (1 - A_y), \quad \Longrightarrow \quad G(x) = \sum_{f=1}^{n-1}(C_y^f * A_y^f) + x * A_y^b,$$

- Attention mask를 foreground와 (1-foreground)에서 fore/back 확실히 구분 (동시 학습 시 mapping이 정확해짐)
- contents/attention mask를 여러방향으로 생성하여 복잡한 이미지를 생성하기 적합하게 바꿈

- CycleGAN based model 이다보니 Cycle-Consistency 부분도 바뀜

$$F(G(x)) = C_x * A_x + G(x) * (1 - A_x), \quad \Longrightarrow \quad F(G(x)) = \sum_{f=1}^{n-1}(C_x^f * A_x^f) + G(x) * A_x^b,$$

Attnetion GAN v2

- Optimization Objective

$$\mathcal{L} = \mathcal{L}_{GAN} + \lambda_{cycle} * \mathcal{L}_{cycle} + \lambda_{id} * \mathcal{L}_{id},$$

- Total Variation regularization

$$\mathcal{L}_{tv} = \sum_{w,h=1}^{W,H} |A_x(w+1,h,c) - A_x(w,h,c)| + |A_x(w,h+1,c) - A_x(w,h,c)|,$$

  - 쉽게 포화(모두 1로 수렴) 될 수 있는 Attention mask generator를 정규화

- pixel loss (contents loss)

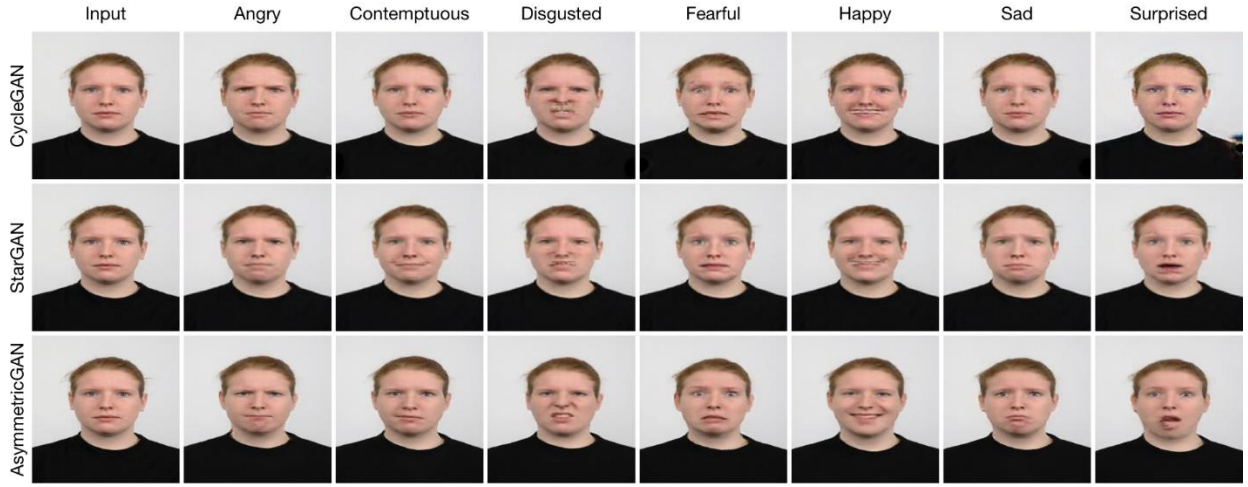
$$\mathcal{L}_{pixel}(G,F) = \mathbb{E}_{x\sim p_{\text{data}}(x)}[\|G(x) - x\|_1] + \mathbb{E}_{y\sim p_{\text{data}}(y)}[\|F(y) - y\|_1].$$

Experiments

Datasets

Selfie2Anime

CelebA



RaFD

Selfie2Anime

Experiments

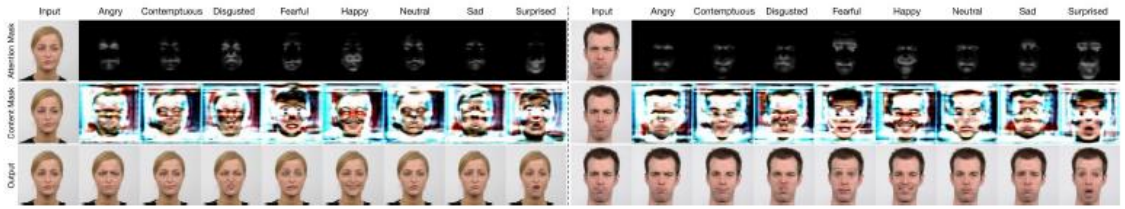


Fig. 11: Attention and content masks on RaFD.

Fig. 12: Attention and content masks on CelebA.

TABLE II: Quantitative comparison on facial expression translation task. For both AMT and PSNR, high is better.

Model	Publish	AR Face		CelebA
		AMT ↑	PSNR ↑	AMT ↑
CycleGAN [3]	ICCV 2017	10.2	14.8142	34.6
DualGAN [4]	ICCV 2017	1.3	14.7458	3.2
DiscoGAN [5]	ICML 2017	0.1	13.1547	1.2
ComboGAN [20]	CVPR 2018	1.5	14.7465	9.6
DistanceGAN [19]	NeurIPS 2017	0.3	11.4983	1.9
Dist.+Cycle [19]	NeurIPS 2017	0.1	3.8632	1.3
Self Dist. [19]	NeurIPS 2017	0.1	3.8674	1.2
StarGAN [14]	CVPR 2018	1.6	13.5757	14.8
ContrastGAN [7]	ECCV 2018	8.3	14.8495	25.1
Pix2pix [2]	CVPR 2017	2.6	14.6118	-
Enc.-Decoder [2]	CVPR 2017	0.1	12.6660	-
BicycleGAN [30]	NeurIPS 2017	1.5	14.7914	-
AttentionGAN	Ours	**12.8**	**14.9187**	**38.9**

TABLE III: AMT results of facial attribute transfer task on CelebA dataset. For this metric, higher is better.

Method	Publish	Hair Color	Gender	Aged
DIAT [37]	arXiv 2016	3.5	21.1	3.2
CycleGAN [3]	ICCV 2017	9.8	8.2	9.4
IcGAN [13]	NeurIPS 2016	1.3	6.3	5.7
StarGAN [14]	CVPR 2018	24.8	28.8	30.8
AttentionGAN	Ours	**60.6**	**35.6**	**50.9**

TABLE IV: KID $\times$ 100 $\pm$ std. $\times$ 100 of selfie to anime translation task. For this metric, lower is better.

Method	Publish	Selfie to Anime
U-GAT-IT [28]	ICLR 2020	**11.61 ± 0.57**
CycleGAN [3]	ICCV 2017	13.08 ± 0.49
UNIT [38]	NeurIPS 2017	14.71 ± 0.59
MUNIT [39]	ECCV 2018	13.85 ± 0.41
DRIT [40]	ECCV 2018	15.08 ± 0.62
AttentionGAN	Ours	12.14 ± 0.43

Experiments

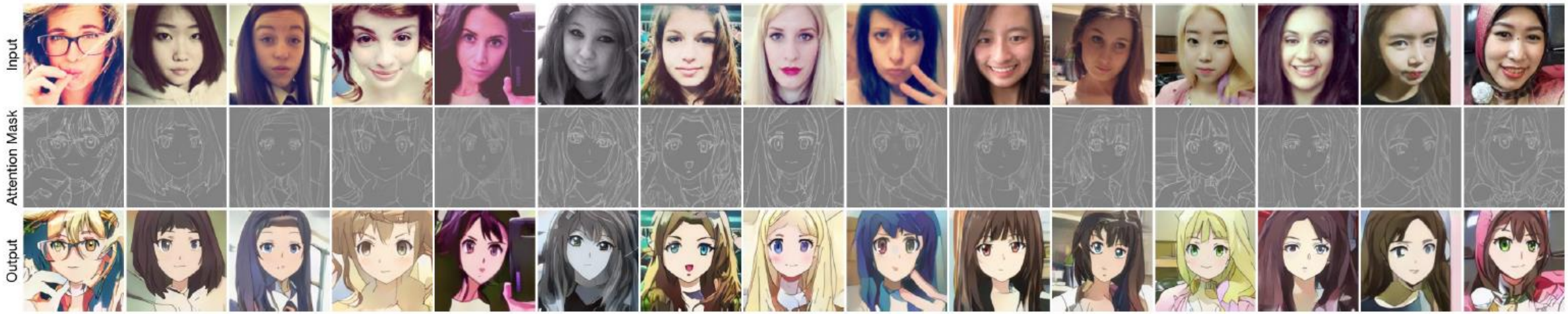


Fig. 13: Attention mask on selfie to anime translation task.

Experiments

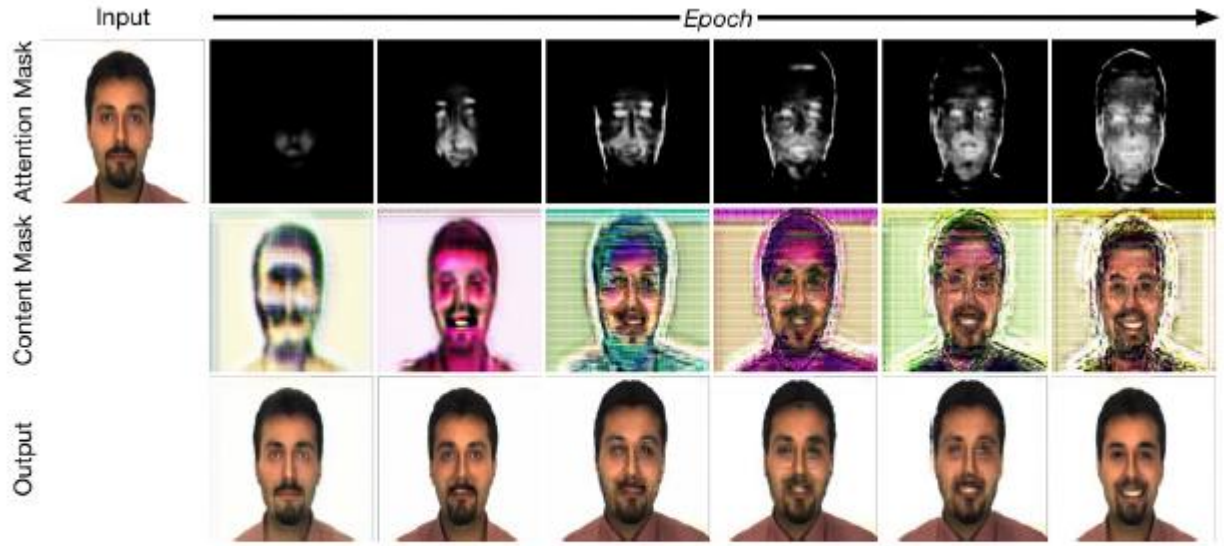


Fig. 14: Evolution of attention masks and content masks.

TABLE V: Overall model capacity on RaFD (m=8).

TABLE V: Overall model capacity on RaFD (m=8).

Method	Publish	# Models	# Parameters
Pix2pix [2]	CVPR 2017	m(m-1)	57.2M×56
Encoder-Decoder [2]	CVPR 2017	m(m-1)	41.9M×56
BicycleGAN [30]	NeurIPS 2017	m(m-1)	64.3M×56
CycleGAN [3]	ICCV 2017	m(m-1)/2	52.6M×28
DualGAN [4]	ICCV 2017	m(m-1)/2	178.7M×28
DiscoGAN [5]	ICML 2017	m(m-1)/2	16.6M×28
DistanceGAN [19]	NeurIPS 2017	m(m-1)/2	52.6M×28
Dist.+Cycle [19]	NeurIPS 2017	m(m-1)/2	52.6M×28
Self Dist. [19]	NeurIPS 2017	m(m-1)/2	52.6M×28
ComboGAN [20]	CVPR 2018	m	14.4M×8
StarGAN [14]	CVPR 2018	1	53.2M×1
ContrastGAN [7]	ECCV 2018	1	52.6M×1
AttentionGAN	Ours	1	52.6M×1