sauna59479@korea.ac.kr

Re-Destyle: 개선된 Facial Destylization 을 활용한 예시 기반 신경망 스타일 전이 연구

유 주원
고려대학교

Re-Destyle: Exemplar-Based Neural Style Transfer using Improved Facial Destylization

Joowon Yoo
Korea University

요 약

예술적 스타일 전이는 예술 작품이 지닌 특징을 다른 이미지에 적용하는 이미지 처리의 오랜 화두 중 하나로, 최근에는 StyleGAN 과 같이 미리 학습된 GAN(생성적 적대 신경망)을 통해 제한된 데이터로도 고해상도의 예술적 초상화를 생성하도록 학습하는 연구가 다양한 방면에서 성과를 내고 있다. 본 논문에서는 2 가지 경로의 StyleGAN 과 Facial Destylization 을 통해 고해상도의 예시 기반 스타일 전이를 달성한 DualStyleGAN 연구에 대해 소개하고, 기존 연구에서 사용된 Facial Destylization 방법이 지닌 한계점을 분석한 뒤, 이를 개선한 새로운 방법, Re-Destyle 을 제안한다. 새로운 Re-Destyle 방법으로 Facial Destylization 을 적용할 경우 학습 시간을 기존 연구의 방법보다 20 배 이상 개선할 수 있으며 그 결과 1000 개 이하의 적은 데이터와 1~2 시간의 추가 학습만으로도 원하는 타겟 초상화 스타일에 대해 1024×1024 수준의 고해상도의 예시 기반 초상화 스타일 전이 및 이미지 생성 모델을 학습할 수 있다.

1. 서론

실제 사람의 얼굴 이미지에 예술 초상화 스타일의 특징을 반영하는 예술적 스타일 전이는 다양한 상업 분야에서 관심을 받는 주제 중 하나이다. 최근에는 틱톡, 스노우 등 유명 앱에서 카메라로 촬영한 실제 얼굴의 사진이나 동영상을 만화나 애니메이션의 예술적 초상화 스타일로 전이하는 필터 기능이 큰 인기를 끌기도 했다. 해당 필터 기능과 같이 스타일 전이 기술에 대한 상업적 수요가 증가하면서, 다양한 초상화의 색상, 질감, 구조 등의 스타일 특징을 효과적으로 전이하면서 동시에 짧은 학습 시간 내에 모델 학습이 가능한 신경망 기반 스타일 전이에 대한 필요성이 함께 대두되고 있다.

지난 몇 년간 GAN 은 이미지 합성을 크게 발전시켰고 이는 스타일 전이 분야에 있어서도 예외는 아니었다. 최근에는 고해상도 얼굴 이미지 생성에 효과적인 StyleGAN [2, 3]을 기반으로 전이 학습하여 비교적 적은 데이터와 짧은 학습 시간만으로 효과적으로 스타일 전이를 달성하는 방법 [15]이 연구되고 있으며, 그 외 다양한 스타일 전이 혹은 Image-to-image 변환 모델 역시 활발하게 연구되고 있다 [12, 13, 14]. 그 중 DualStyleGAN [1]은 두 개의 StyleGAN 을 사용해 내부와 외부 스타일의 경로를 분리한 예시 기반 스타일 전이 모델로, 안정적인 학습을 위해 Facial Destylization 기법을 활용하여 타겟 초상화에서 스타일 요소를 제거한 현실적인 얼굴을 생성한 뒤 초상화-얼굴 한 쌍의 데이터를 적합한 초기값으로 사용하는 아이디어를 제시한다.

DualStyleGAN 이 Facial Destylization 을 활용하여 예시 기반 스타일 전이에 있어 높은 효용성을 보여줬지만, 기존의 Facial Destylization 방법은 1) 최적화 기반 잠재 공간 탐색 과정에서 시간이 오래 걸리고, 2) 적은 데이터로 달성하기 힘든 StyleGAN 튜닝이 별도로 필요하다는 한계점을 갖고 있다. 따라서 본 논문에서는 기존의 한계를 극복하기 위해 최적화 방식이 아닌 인코더만을 통해 직접 잠재 공간에 매핑하는 새로운 Facial Destylization 방법을 연구하도록 한다. 본 논문에서는 다양한 인코더 [4, 5, 6, 7]를 비교한 뒤 그 중 잠재 공간 매핑에 가장 적합한 Restyle [6]을 중첩 적용하는 것으로 짧은 시간 내에 동일한 수준의 Facial Destylization 목표를 달성하는 Re-Destyle 방법을 제안한다. 본 논문의 Re-Destyle 방법을 사용할 경우 최적화 기반 Facial Destylization 보다 속도를 20 배 이상 개선 가능할 뿐 아니라 별도의 튜닝 추가 모델에 대한 학습이 필요하지도 않으면서 초상화의 특징을 정확하게 반영하는 현실의 얼굴을 생성할 수 있다. 또한 Re-Destyle 로 진행한 결과 데이터를 통해서도 DualStyleGAN 에 적용하여 보다 짧은 학습 시간으로 동일한 수준의 스타일 전이 이미지를 생성 가능한 것도 확인할 수 있다.

본 논문의 구성은 다음과 같다. 2 절에서는 관련 연구와 함께 DualStylaGAN 이 Facial Destylization 을 통해 초상화 스타일 전이를 어떻게 달성하였는지에 대해 소개한다. 3 절은 DualStyleGAN 의 개선 방향 및 Re-Destyle 이라는 새로운 Facial Destylization 방법에 대해 설명하고, 4 절에서는 제안한 Re-Destyle 방법을 적용해 기존 방법과 비교해 어떠한 개선이 이루어졌는지 실험을 통해서 확인한다. 마지막으로 5 절에서는 본 논문의 결론을 맺는다.

2. 관련 연구

2.1. GAN Inversion

이미지 생성 분야에 있어 GAN 이 많은 발전을 이루면서 GAN 모델의 잠재 공간을 탐색하고 조작하고자 하는 연구 또한 꾸준히 이어져 왔다. 그 중 대표적인 분야가 GAN Inversion [11]으로, StyleGAN 과 같이 사전 훈련된 GAN 이 주어진이미지를 정확히 재구성할 수 있는 잠재 벡터를 찾는 것을목표로 하고 있다. 정확도 높은 GAN Inversion 을 달성하기위해 입력 이미지와 생성 이미지 사이의 오류 손실이최소화되도록 잠재 벡터를 직접 최적화하거나 [8, 9], 주어진이미지를 잠재 공간에 직접 매핑하도록 인코더를 훈련하거나[4, 5], 혹은 양쪽을 혼합하는 방식을 사용한다 [10]. 최근에는인코더를 반복 정제하여 잔차를 예측하는 것으로 잠재 공간매핑의 정확도를 올리거나 [6] 훈련 과정에서 관측되지 않았던도메인 외부의 초상화 이미지에 대해서 일반화가 가능한수준의 인코더 기반 GAN Inversion 모델도 연구되고 있다 [7].

2.2. DualStyleGAN 과 Facial Destylization

DualStyleGAN [1]은 예시 기반 스타일 전이 모델 중 하나로, 두 개의 StyleGAN 을 사용하여 하나는 내부 스타일 경로, 다른 하나는 외부 스타일 경로로 동작하여 보다 안정적인 스타일 전이를 달성하도록 한다. 해당 연구에서는 두 경로의 모델을 안정적으로 동시 학습하기 위해 최적화 기반의 Facial Destylization 방법을 제안한다. Facial Destylization 의 목표는 초상화로부터 인코더로 매핑한 z_e^+와 초상화에 대한 현실적인 얼굴에 대응하는 새로운 잠재 벡터 z_i^+를 찾는 것으로, 이 과정에서 z_i^+를 찾기 위해 FFHQ StyleGAN g [2]로부터 초상화 스타일로 튜닝된 새로운 StyleGAN g' 로 생성된 이미지가 입력 초상화 이미지와 비교하여 perceptual loss [16]와 identity loss [17]가 최소화되도록 최적화를 진행한다 [8, 15]. 이렇게 계산된 z_e^+와 z_i^+ 한 쌍의 잠재 벡터는 이후 모델의 학습 과정에서 얼굴 구조를 변형하는 방법에 대한 유효한 감독으로 사용된다.

3. 제안 방법

3.1. 최적화 기반 Facial Destylization 의 문제 확인 및 개선

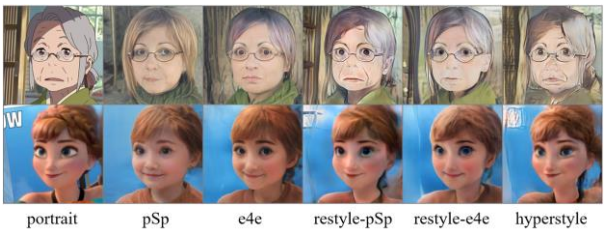
본 논문은 DualStyleGAN 을 기반으로 동일한 수준의 예시 기반 스타일 전이를 달성하되, 기존의 Facial Destylization 과정에서 발생하는 문제점을 제기하고 이를 해결하고자 한다. 기존 방식의 경우, 입력 초상화 이미지로부터 실제 얼굴에 가까운 잠재 벡터를 찾는다는 점에서 넓은 범위에서 잠재 공간 임베딩에 해당하며 인코더와 최적화를 같이 사용하는 측면에서 GAN Inversion 중 혼합 방식과 유사한데, 이는 곧 최적화 방식 혹은 혼합 방식 GAN Inversion 과 동일한 문제 및 한계를 갖게 된다. 즉, 기존의 방식은 최적화 기반 GAN Inversion 과 마찬가지로 1) 최적화 과정에서 비용이 많이 들고 수렴하는데 많은 시간이 소요되며, 2) 반전된 잠재 벡터가 의미 있는 특징을 반영하지 못해 동일 인물에 대한 초상화라도 전혀 다른 실제 얼굴을 생성하게 된다 [4]. 거기에 더해, 기존의 방법은 튜닝된 StyleGAN g'을 Destylization 과정에서 별도로 요구하게 되는데, 만약 데이터가 부족할 경우 전이 학습 과정에서 과적합 문제로 g'를 학습하는데 실패할 가능성이 있을 뿐 아니라 [18] 어느 정도의 품질로 튜닝된 StyleGAN 을 사용해야 하는지 정확한 기준을 제시하기 어렵다. 이는 적은 데이터로도 학습이 가능하다는 모델의 장점을 훼손하게 되며 결국 다량의 초상화 데이터를 확보하지 못할 경우 여전히 DualStyleGAN 학습에 실패할 가능성이 있음을 암시한다.

본 논문에서는 이러한 문제를 해결하기 위해 GAN Inversion 중 주어진 이미지를 잠재 공간에 직접 매핑하는 인코더 방식을 채택해 그 장점을 취하고자 한다. 즉, 입력된 초상화로부터 실제 얼굴에 해당하는 잠재 벡터를 직접 매핑 할 수 있도록 사전 훈련된 인코더만을 이용해 빠른 시간 내에 정확도 높은 Facial Destylization 을 달성하는 새로운 방법을 제안하는 것이 본 논문 연구의 목표이다.

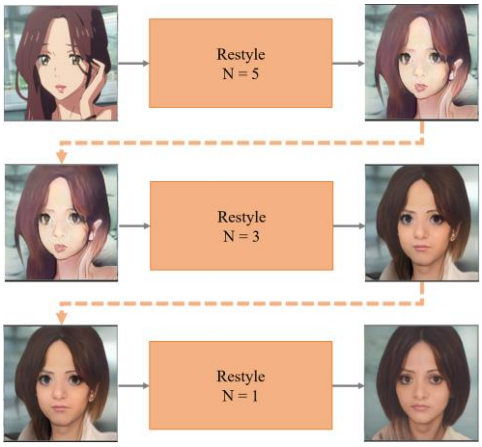
3.2. 인코더 모델 선택을 위한 초상화 해석력 비교

기존의 Facial Destylization 에서 필요한 최적화 과정을 생략하기 위해 GAN Inversion 중 잠재 공간에 직접 매핑하는 인코더 모델을 사용하기로 했다. 본래 GAN Inversion 은 입력된 이미지와 같은 이미지를 생성할 수 있는 잠재 벡터를 찾는 것을 목표로 하지만, 이미지에서 잠재 벡터를 찾는 기능은 동일한 만큼 Facial Destylization 에서의 잠재 공간 매핑에도 기존의 GAN Inversion 인코더가 효과적으로 동작할 것으로 보고 이에 적합한 인코더를 탐색하였다.

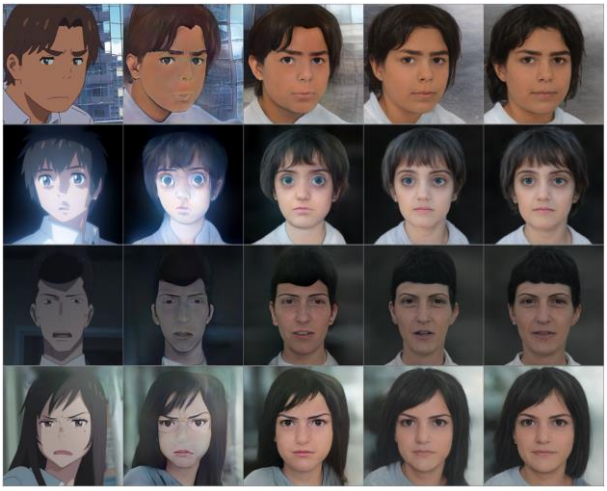
미리 학습된 인코더는 기본적으로 실제 얼굴에 대해 학습된 모델이지만, 최신 인코더는 실제 얼굴로 학습하더라도 학습 과정에서 경험하지 않았던 얼굴 초상화에 대해 얼굴과 관련된 특징을 잘 해석하는 성능을 보인다. <그림 1>은 FFHQ 데이터셋 [2]으로 학습된 6 가지의 인코더를 비교한 결과이며, 본 논문에서는 최종적으로 초상화에 대해 보다 잘 복구하면서 점진적으로 실제 얼굴의 비중을 섞을 수 있는 Restyle 방식을 선택하기로 결정했다. 그 중에서도 기존 DualStyleGAN 의 미리 학습된 pSp 기반 모델과의 호환성을 위해 Restyle-pSp 모델을 채택했다 [4, 5, 6, 7].



<그림 1> 초상화 해석력 비교를 통한 인코더 모델 검토



<그림 2> Restyle 의 N 의 점진적 감소를 통한 Re-Destyle 절차



<그림 3> Re-Destyle 에 의한 Facial Destylization 적용 결과

3.3. Re-Destyle: Restyle 의 점진적 적용을 통한 Destylization

본 논문에서는 Restyle 인코더를 기반으로 하여 Facial Destylization 을 달성하는 Re-Destyle 을 제안한다. Restyle [6]은 입력된 이미지를 StyleGAN 잠재 공간으로 GAN Inversion 시키는 인코더로, 단일 정방향 패스에서 반전하도록제약된 기존의 인코더를 반복 적용하는 것으로 보다 정확한탐색을 가능하게 하는 특징이 있다. 여기서 반복 개선 횟수N 이 클 경우 <그림 1>처럼 실제 얼굴에 대해서만 학습했어도초상화와 같은 스타일이 더해진 이미지에 대해서도 특징을 잘잡아내며, 반대로 N 을 1 에 수렴하는 수준까지 줄일 경우초기화 단계에서 각 생성기의 평균 스타일 합성 이미지와벡터가 더해지면서 현실적인 얼굴 요소가 강하게 드러나게된다.

Re-Destyle 은 Restyle 의 N 을 점진적으로 낮추어 현실적인 얼굴에 적응하도록 유도하는 아이디어를 사용한다. <그림 2>은 Re-Destyle 의 절차를 요약한 것으로, Restyle 의 반복 횟수 N 을 점진적으로 낮추어 첫 회에는 N 이 높아 초상화의 특징이 잘 반영되도록 합성하는 잠재 공간을 찾되 이후 과정에서는 N 을 줄여 초상화에 대한 특징은 줄어들고 평균 얼굴에 대한 특징이 더해질 수 있도록 유도한다. 이렇게 5-3-1 혹은 6-4-2 순으로 N 을 줄이게 되면 마치 최적화방식과 유사하게 반복 개선이 가능하면서도 미리 학습된인코더를 사용하기 때문에 빠르고 정확하게 <그림 3>과 같이Facial Destylization 결과 잠재 벡터를 탐색할 수 있다.

4. 실험 결과 및 분석

4.1. 데이터셋 및 구현 세부사항

본 논문에서는 데이터셋으로 Cartoon, Kiminonawa 2 종을 사용하였다. 두 데이터셋은 모두 1024×1024 의 고해상도 이미지로 구성되어 있으며, Cartoon 은 기존 DualStyleGAN 연구에서 사용된 317 개의 이미지, Kiminonawa 는 <너의 이름은>을 비롯한 다나카 마사요시가 캐릭터 디자인한 일본 애니메이션 5 종으로부터 추출한 633 개의 이미지이다. 학습 GPU 는 RTX 3090 1 대를 사용했으며 PyTorch 프레임워크를 사용하여 구현되었다.

	Learning Time	
Baseline	5.15h	300it / (2.56it/s) / 4 * 633
Re-Destyle	**0.256h**	1.45s/it * 633

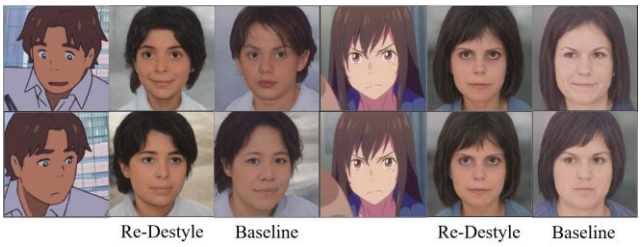
<표 1> 기존 Baseline 과 Re-Destyle 의 학습 시간 비교

4.2. Facial Destylization 시간 개선

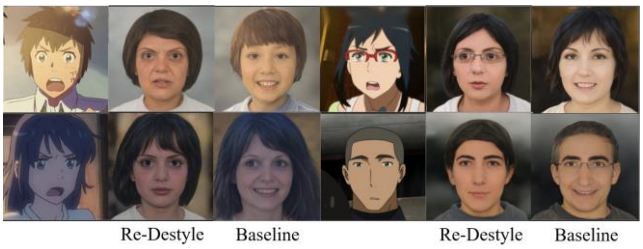
기존 최적화 방식의 Destylization 방식을 Baseline 으로 633 개의 kiminonawa 데이터셋에 적용하여 시간을 비교하였다. <표 1>은 시간 측정 결과를 비교한 것으로, 기존에 300 회에 달하는 반복 최적화를 생략하고 미리 학습된 인코더에 의해 직접 매핑이 가능하기 때문에 새로운 Re-Destyle 을 통하여 Facial Destylization 은 기존 Baseline 에 비해 20.1 배에 달하는 학습 시간 단축 효과를 확인할 수 있었다.

4.3. Facial Destylization 잠재 공간 탐색 정확도 개선

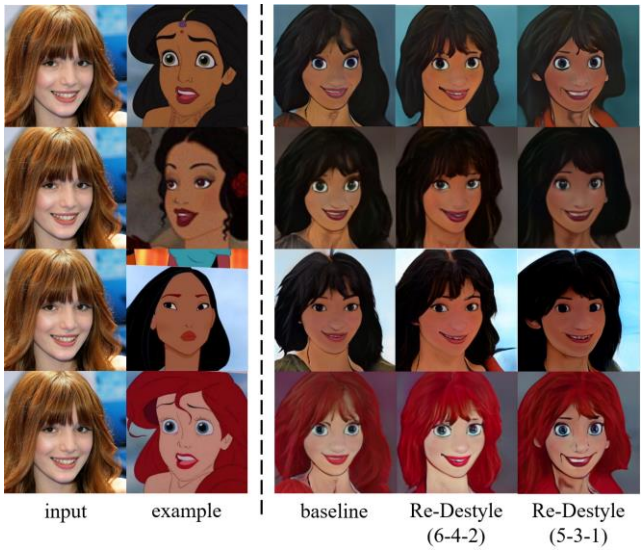
Facial Destylization 의 결과가 실제 얼굴 생성기가 지닌 잠재 공간을 정확하게 반영하고 있는지를 확인하기 위해 1) 동일한 인물을 묘사한 초상화에 대해 유사성이 강한 실제 얼굴 이미지가 생성되었는지, 2) 표정 등 초상화에 담긴 인물 정보가 생성된 실제 얼굴에 반영되는 비율이 높은 지 확인하였다. <그림 4>는 동일 인물에 대한 초상화를 Facial Destylization 했을 때의 결과 유사성을 비교한 것으로, Re-Destyle 은 동일 인물 초상화에 대해 유사성이 높은 실제 이미지를 생성하는 반면, Baseline 인 기존의 최적화 방식은 차이가 거의 없는 초상화에 대해서도 전혀 다른 실제 얼굴을 생성하는 경우가 발생하는 것을 확인할 수 있었다. <그림 5>에서는 Re-Destyle 을 사용한 결과, 표정이나 안경 같은 세부 정보가 유실되지 않고 보다 잘 반영되는 것을 확인할 수 있다. <그림 4>, <그림 5>에서 보인 결과는 본 논문의 Re-Destyle 이 학습된 인코더를 사용함으로써 단순한 손실 최적화를 통한 Facial Destylization 보다 올바른 잠재 공간 탐색이 가능함을 암시하고 있다.



<그림 4> Destyle 정확도 비교: 동일 인물 초상화 유사성



<그림 5> Destyle 정확도 비교: 표정 등 초상화 인물 특징 반영



<그림 6> Re-Destyle 을 통한 DualStyleGAN 학습 결과

4.4. DualStyleGAN 학습 감독 유효성 확인

마지막으로 Re-Destyle 로 Facial Destylization 하여 얻은 z_e^+와 z_i^+가 DualStyleGAN 의 학습에 유효한 감독으로 동작하는지 확인했다. 정확한 비교를 위해 기존 DualStyleGAN 연구에서 사용된 Cartoon 데이터셋을 그대로 사용하고 Loss 의 람다 계수나 학습 횟수 등의 하이퍼 파라미터도 동일하게 적용했다. 그 결과 <그림 6>에서 확인하듯이 Re-Destyle 로 인코더를 통해 직접 매핑한 z_e^+와 z_i^+라도 DualStyleGAN 모델 학습이 가능하며 예시 기반 스타일 전이에 대한 감독 역할로서 충분한 효과가 있다는 사실을 보여주고 있다.

5. 결론

본 논문에서는 기존 DualStyleGAN 의 최적화 기반 Facial Destylization 의 한계점을 분석하고 Restyle 인코더의 반복 개선 횟수를 점진적으로 줄이는 방식으로 최적화 과정 없이 인코더만으로 Facial Destylization 을 달성하는 Re-Destyle 방법을 제안하였다. Re-Destyle 은 기존 최적화 기반 방식에 비해 속도와 정확도 양 측면에서 개선되었고, Re-Destyle 의 결과를 DualStyleGAN 의 학습에 적용하여 예시 기반 전이 학습의 감독 역할을 수행 가능하다는 사실도 확인할 수 있었다.

참고문헌 (References)

[1] Shuai Yang, Liming Jiang, Ziwei Liu, Chen Change Loy. Pastiche Master: Exemplar-Based High-Resolution Portrait Style Transfer. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[2] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, pages 4396-4405, 2019.

[3] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, pages 8107-8116, 2020.

[4] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.

[5] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation, 2021.

[6] Alaluf, Yuval and Patashnik, Or and Cohen-Or, Daniel. ReStyle: A Residual-Based StyleGAN Encoder via Iterative Refinement. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2021.

[7] Yuval Alaluf and Omer Tov and Ron Mokady and Rinon Gal and Amit H. Bermano. HyperStyle: StyleGAN Inversion with HyperNetworks for Real Image Editing. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[8] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In Proc. Int'l Conf. Computer Vision, pages 4432–4441, 2019.

[9] Rameen Abdal, Yipeng Qin, and Peter Wonka.Image2stylegan++: How to edit the embedded images? In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8296–8305, 2020.

[10] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. arXiv preprint arXiv:2004.00049, 2020.

[11] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. Generative visual manipulation on the natural image manifold. In European conference on computer vision, pages 597–613. Springer, 2016.

[12] Junho Kim, Minjae Kim, Hyeonwoo Kang, and Kwang Hee Lee. U-GAT-IT: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. In Proc. Int'l Conf. Learning Representations, 2019.

[13] Bing Li, Yuanlue Zhu, Yitong Wang, Chia-Wen Lin, Bernard Ghanem, and Linlin Shen. AniGAN: Style-guided generative adversarial networks for unsupervised anime face generation. IEEE Transactions on Multimedia, 2021.

[14] Min Jin Chong and David Forsyth. GANs N' Roses: Stable, controllable, diverse image to image translation. arXiv preprint arXiv:2106.06561, 2021.

[15] Justin NM Pinkney and Doron Adler. Resolution dependent gan interpolation for controllable image synthesis between domains. arXiv preprint arXiv: 2010. 05334, 2020.

[16] Justin Johnson, Alexandre Alahi, and Fei Fei Li. Perceptual losses for real-time style transfer and super-resolution. In Proc. European Conf. Computer Vision, pages 694–711. Springer, 2016.

[17] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, pages 4690–4699, 2019.

[18] S. Mo, M. Cho, and J. Shin.Freeze the discriminator: a simple baseline for fine-tuning GANs.*CoRR*, abs/2002.10964, 2020.