딥페이크 (2)

양유진

한성대학교 HANSUNG UNIVERSITY

유튜브 주소: https://youtu.be/2MmdqzfOry4

CryptoCraft LAB
https://crypto.modoo.at

Contents

01 딥페이크 영상 생성 알고리즘

02 데이터 기반 딥페이크 탐지 알고리즘

CryptoCraft LAB

딥페이크 영상 생성 알고리즘(1) DeepFaceLab

<특징>

1. GAN 기반의 최신 딥페이크 영상 생성 알고리즘

2. 가장 유명함

- 현재 만들어진 딥페이크 영상의 95%가 이 프로그램

3. DeepFakes 생성기법에 해당

- 얼굴의 특징을 목표영상의 표정, 반응으로 나타내는 기법

딥페이크 영상 생성 알고리즘(1) DeepFaceLab

<1단계> 특징 추출 단계

1) Face detection: 이미지에서 얼굴 있는 부분 찾음
2) Face alignment: 얼굴 특징 따라서 *랜드마크 표시 → 표정 표현
3) Face segmentation: 분할(segmentation)작업으로 얼굴을 가리는 물체를 제거한 **mask 생성

- 다양한 구도, 표정 수작업으로 추출한 마스크로 전체 이미지 훈련

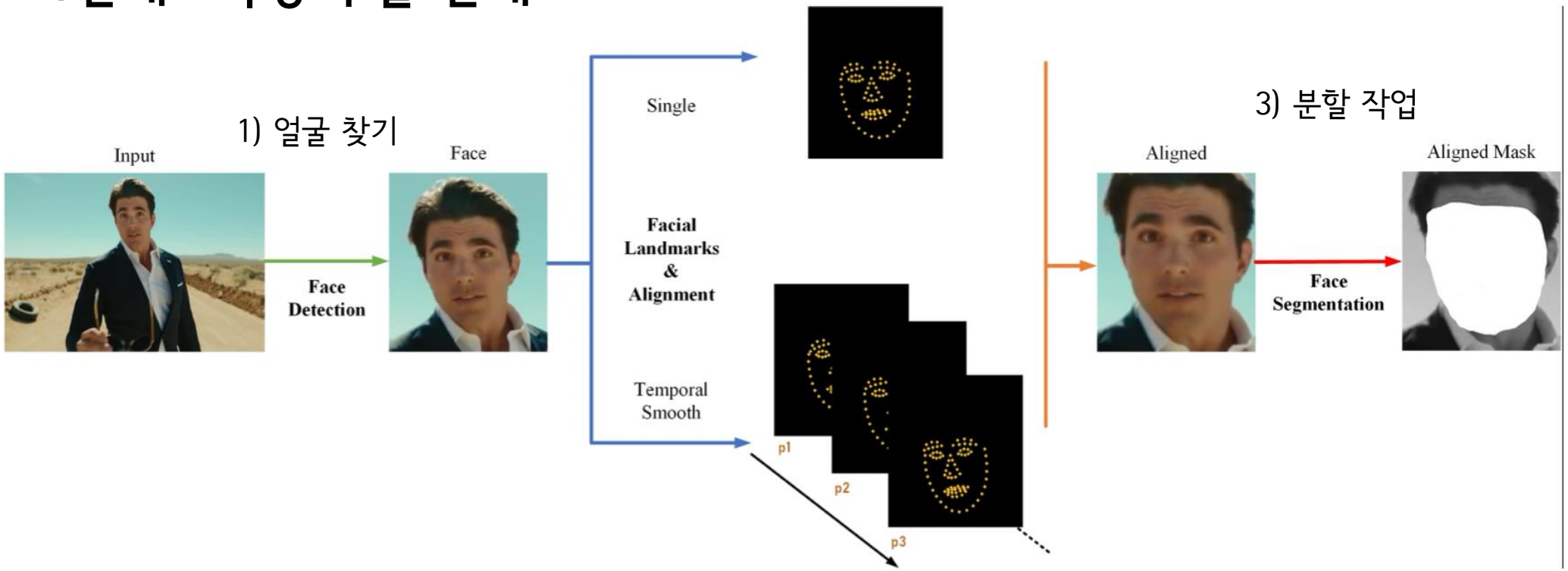
→ 머리카락/얼굴 경계 구분하는 최종 마스크 생성

*랜드마크 표시 – 목표영상의 표정이 결과물에 명확하게 나타내게 해줌.

**Mask = segmentation 된 object label

딥페이크 영상 생성 알고리즘(1) DeepFaceLab

<1단계> 특징 추출 단계



출처: http://asq.kr/FjhfUkG7ihGZRNg

딥페이크 영상 생성 알고리즘(1) DeepFaceLab

<2단계> 학습 단계

: 1단계에서 생성한 마스크로 학습 시작

1. DF구조

1) 원본 이미지 쌍(원본이미지, 원본마스크)과 목표 이미지 쌍을 encoder에 입력으로 넣음
2) encoder의 출력값들을 모두 inter layer에 넣음
3) 원본 이미지-원본 decoder에, 목표 이미지-목표 decode에 각각 넣음
4) 생성된 예측 이미지 쌍을 각각의 discriminator(판별기)에 넣어서 진위여부(real/fake) 분류함.

딥페이크 영상 생성 알고리즘(1) DeepFaceLab

<2단계> 학습 단계

2. LIAE구조

1) 원본 이미지 쌍(원본이미지, 원본마스크)과 목표 이미지 쌍을 encoder에 입력으로 넣음

2) encoder의 출력값 중 InterAB엔 원본,목표이미지를 넣고, InterB에는 목표 이미지만 넣음

3) InterAB를 거친 원본이미지 2개를 concat()하여 합치고, InterAB를 거친 목표이미지와 InterB를 거친 목표이미지를 concat()하여 합침.

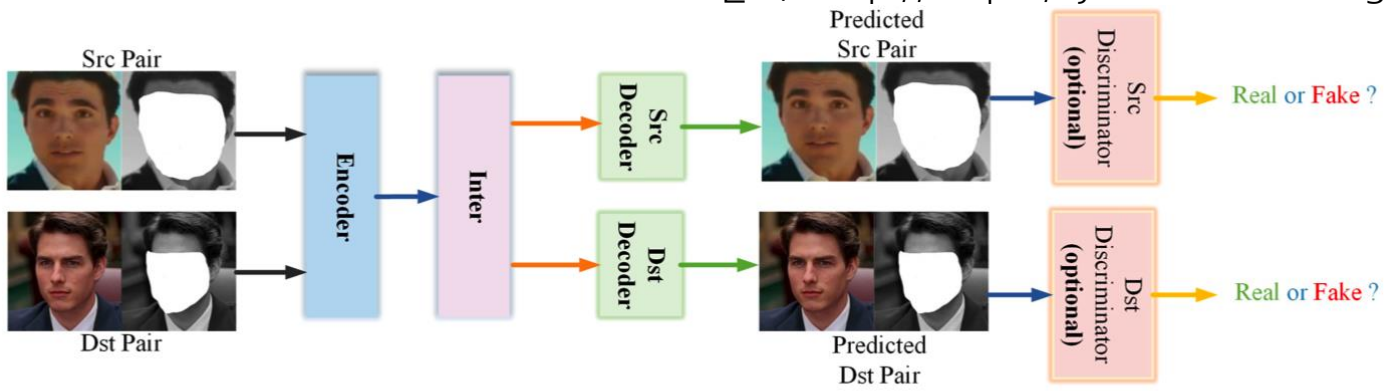
4) 3번 과정에서 나온 2가지 값을 하나의 공유된 decoder에 넣음

5) 생성된 예측 이미지 쌍을 각각의 discriminator(판별기)에 넣어서 진위여부(real/fake) 분류함.

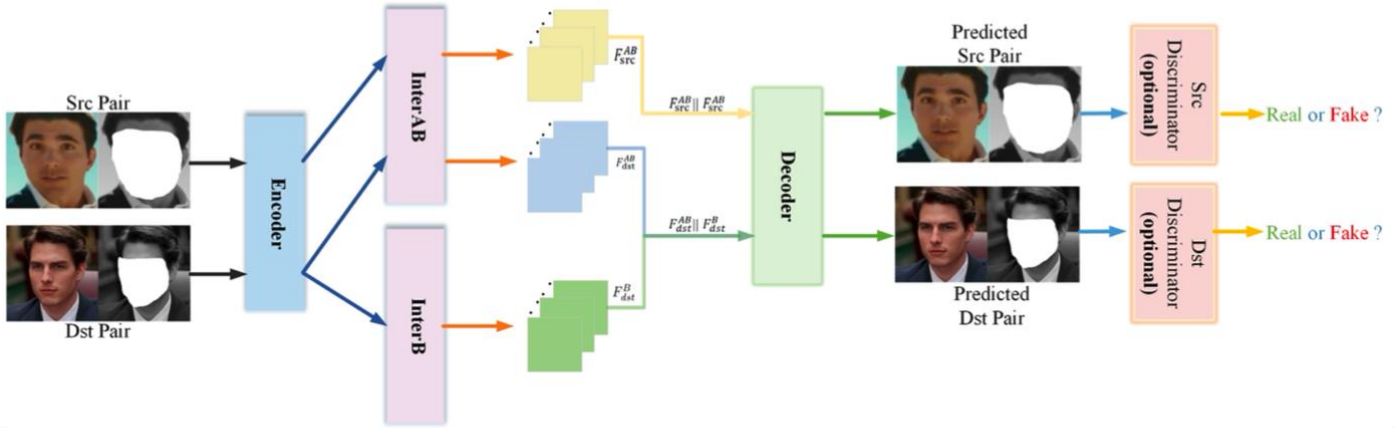
딥페이크 영상 생성 알고리즘(1) DeepFaceLab

<2단계> 학습 단계

출처: http://asq.kr/FjhfUkG7ihGZRNg



(a) **DF** structure



(b) **LIAE** structure

딥페이크 영상 생성 알고리즘(1) DeepFaceLab

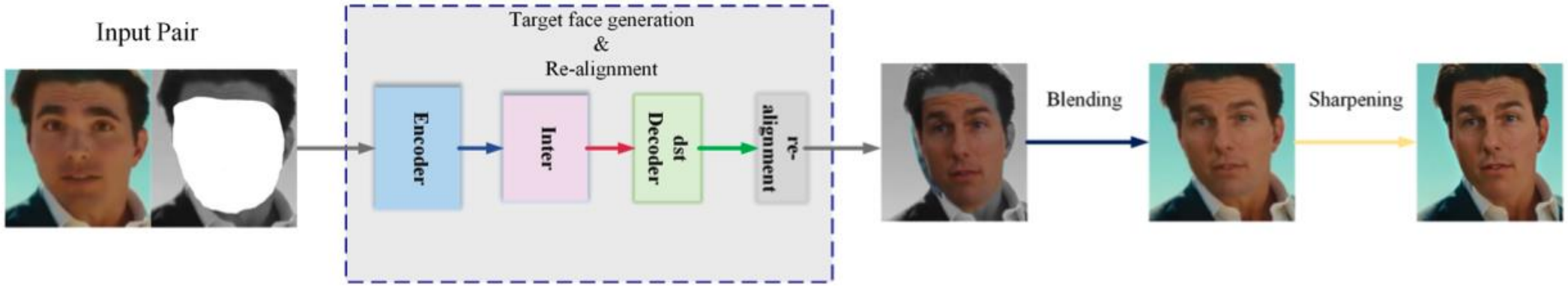
<3단계> 전환(Conversion) 단계

: 이미지를 합치고 후처리 해주는 단계_마스크를 제외한 부분은 모두 일치해야함.

1) Blending: 예측 모형 2가지를 통합하여 정확도가 더 높은 하나의 예측 모형 만들어줌
2) Sharpening: 출력 이미지의 화질을 선명하게 만들어줌. (super-resolution)

딥페이크 영상 생성 알고리즘(1) DeepFaceLab

<3단계> 전환(Conversion) 단계

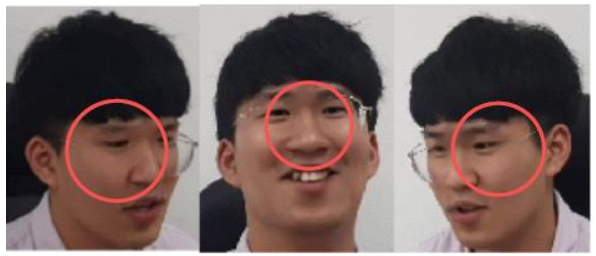


출처: http://asq.kr/FjhfUkG7ihGZRNg

딥페이크 영상 생성 알고리즘(1) DeepFaceLab

장점
- 짧은 수렴 시간 → 전체 생성 시간 짧음
- 수동 후처리 작업 → 높은 해상도의 영상 생성 가능.

단점
- 합성 영상 만드는 작업 자동화 불가
  - 합성 영상에 대응되는 모델을 만들어야 함
  - 마스크 생성 시 수동 작업 불가피
- 가림현상
  - 원본 영상에서 학습할 수 없는 액세서리가 목표 영상에 있을 경우 마스크에 반영되지 않음.



딥페이크 영상 생성 알고리즘(2) FaceShifter

<특징>

1. DeepFaceLab의 문제점 보완

- 높은 해상도의 딥페이크 이미지 생성
- 가림 현상 보완

2. 2단계 모델 구조 제시

- AEI-Net & HEAR-Net

딥페이크 영상 생성 알고리즘(2) FaceShifter

Adaptive Embedding Integration Network

<1단계> AEI-Net - 얼굴 스와핑 예상 결과 생성

1) ID인코더에 원본 이미지를 입력으로 넣음 → 특징추출

2) 표현을 담은 목표 이미지를 인코더/디코더 구조 모델에 입력으로 넣음
→* $\hat{Y}_{s.t}$생성

- 모든 레이어에 특징벡터(포즈, 윤곽,표정, 헤어스타일, 피부색, 배경, 장면, 조명 등) 넣음
→ 얼굴 특징 보존

- 다중 레이어 출력 모두 사용 → 다각도의 표현 학습 가능

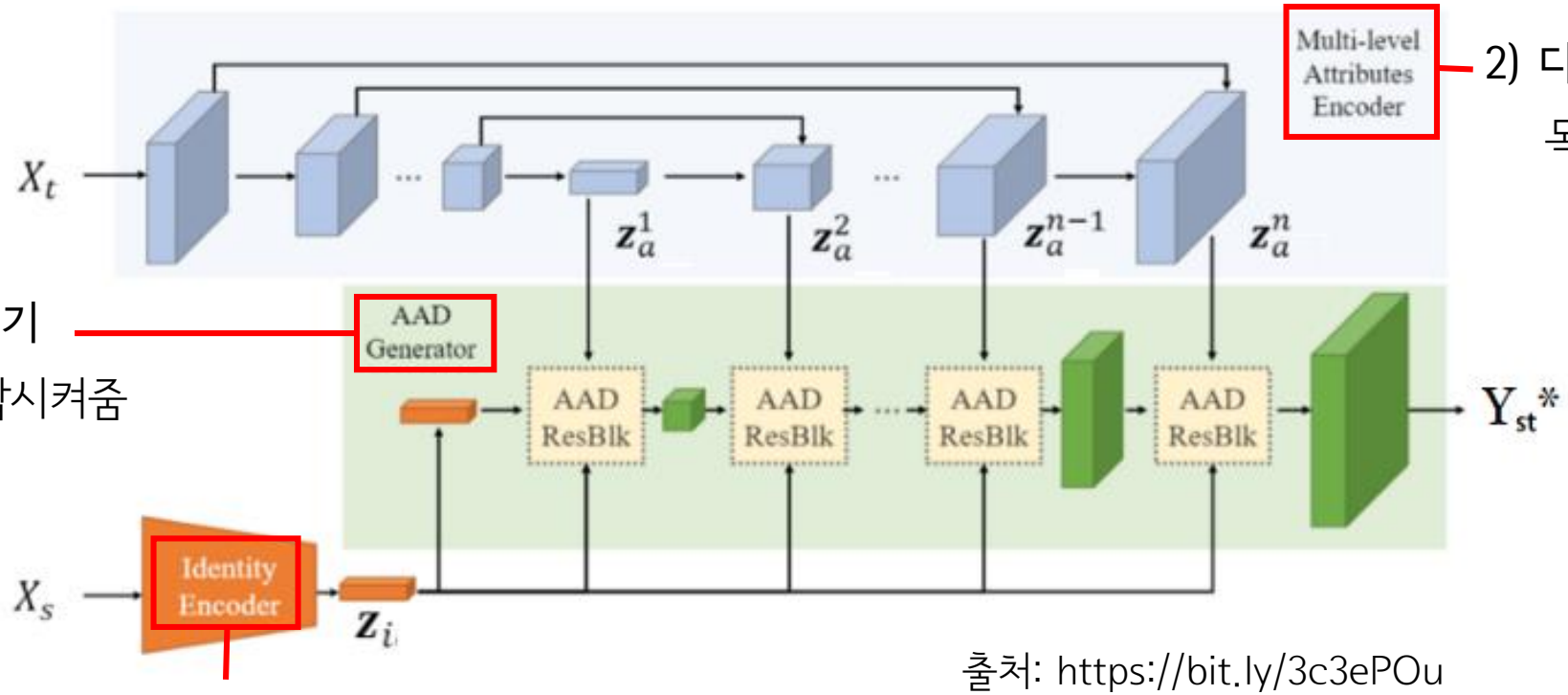
3) AAD 생성기에서 값들을 통합하여 최종 출력값 생성.

- 이전 AAD블록 출력에서 합성곱레이어&활성화함수(시그모이드) 취함
→ 얼굴 특징 잡은 마스크 생성

*$\hat{Y}_{s.t}$ – 레이어에서 각각의 표현 벡터들의 얼굴 변환이 완성된 이미지

딥페이크 영상 생성 알고리즘(2) FaceShifter

<1단계> AEI-Net



2) 다중 레벨 속성 인코더

목표 이미지 인코딩

3) AAD 생성기

1과 2의 출력 통합시켜줌

1) ID인코더

원본 이미지 특징 추출

출처: https://bit.ly/3c3ePOu

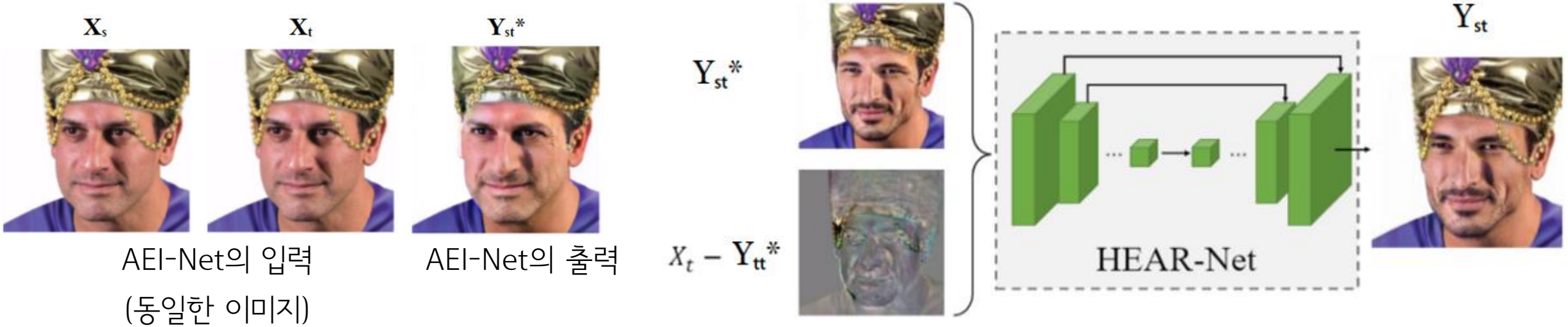
딥페이크 영상 생성 알고리즘(2) FaceShifter

Heuristic Error Acknowledging Refinement Network

<2단계> HEAR-Net : 1단계 출력 개선

AEI-Net에서 제거한 항목 (액세서리, 머리카락) 다시 복구

=학습되지 않은 물체 표현하게 해줌.



AEI-Net의 입력 (동일한 이미지)

AEI-Net의 출력

출처: https://bit.ly/3c3ePOu

딥페이크 영상 생성 알고리즘(2) FaceShifter

단점

- 공개된 코드X → 정확한 성능 파악X
- 합성과정이 너무 길어서 시간이 오래 걸림

장점

- 합성 영상 생성의 일반화된 모델
  → 영상마다 맞는 모델 만들지 않아도 됨.
- 별도의 수작업 없이 진행 가능
- 원본 영상, 목표 영상 선택 폭이 확대됨.

데이터 기반 딥페이크 탐지 알고리즘(1) ForensicTransfer

<이론적 배경>

“합성곱 신경망은 위조 탐지에 성능적으로 매우 효과적(증명O)”

한계

1) 학습데이터에 너무 의존적 → 특정한 생성 방법에 과적합
⇒ 학습할 때 보지 못한 생성 방법에 약함 (성능 감소)

2) 이미지 합성/생성 방법 다양 → 각 영상 생성 방법마다 모델 하나씩 만들어야 함

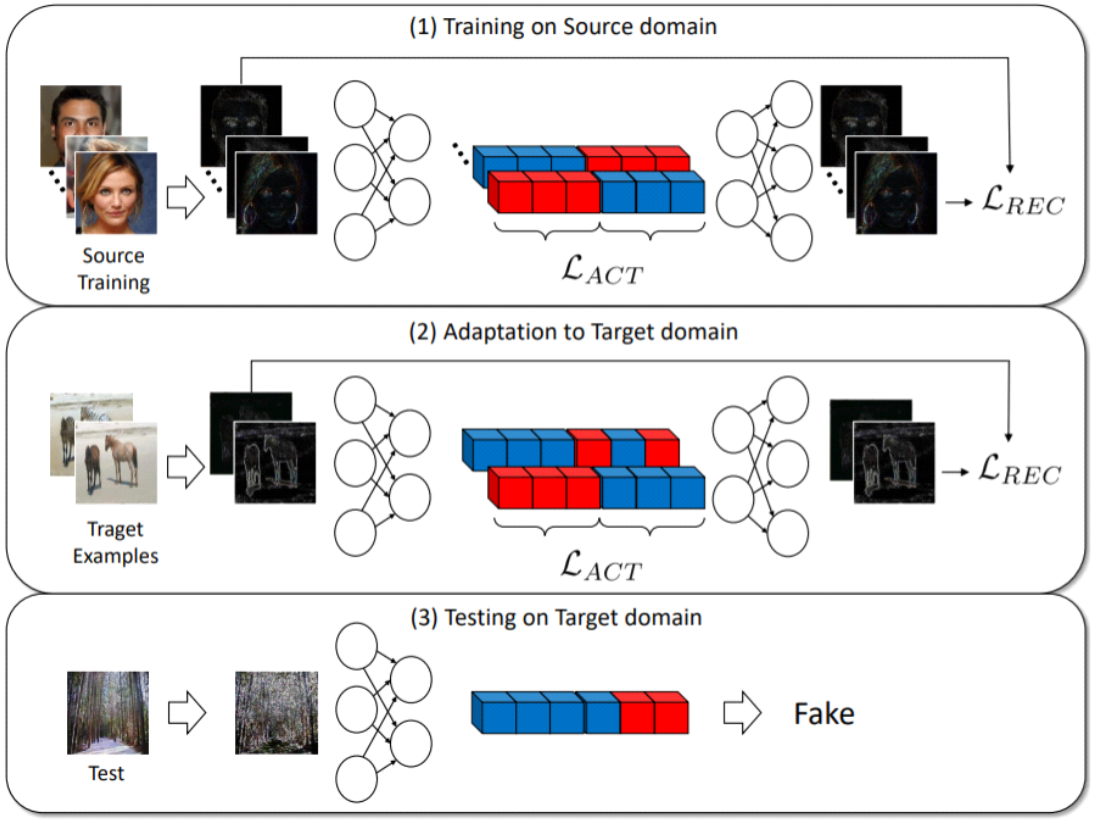
도메인 적응(Domain adaptation)

:다양한 방법으로 생성된 가짜 이미지를 하나의 네트워크로 구분할 수 있는 방법 중 하나.

데이터 기반 딥페이크 탐지 알고리즘(1) ForensicTransfer

특정한 영상 생성 방법을 학습한 모델에 새로운 생성 방법으로 만든 이미지를 소량 재학습시켜 또다른 생성 방법으로 만든 딥페이크 탐지에 사용하는 방법

데이터 기반 딥페이크 탐지 알고리즘(1) ForensicTransfer



<학습절차>

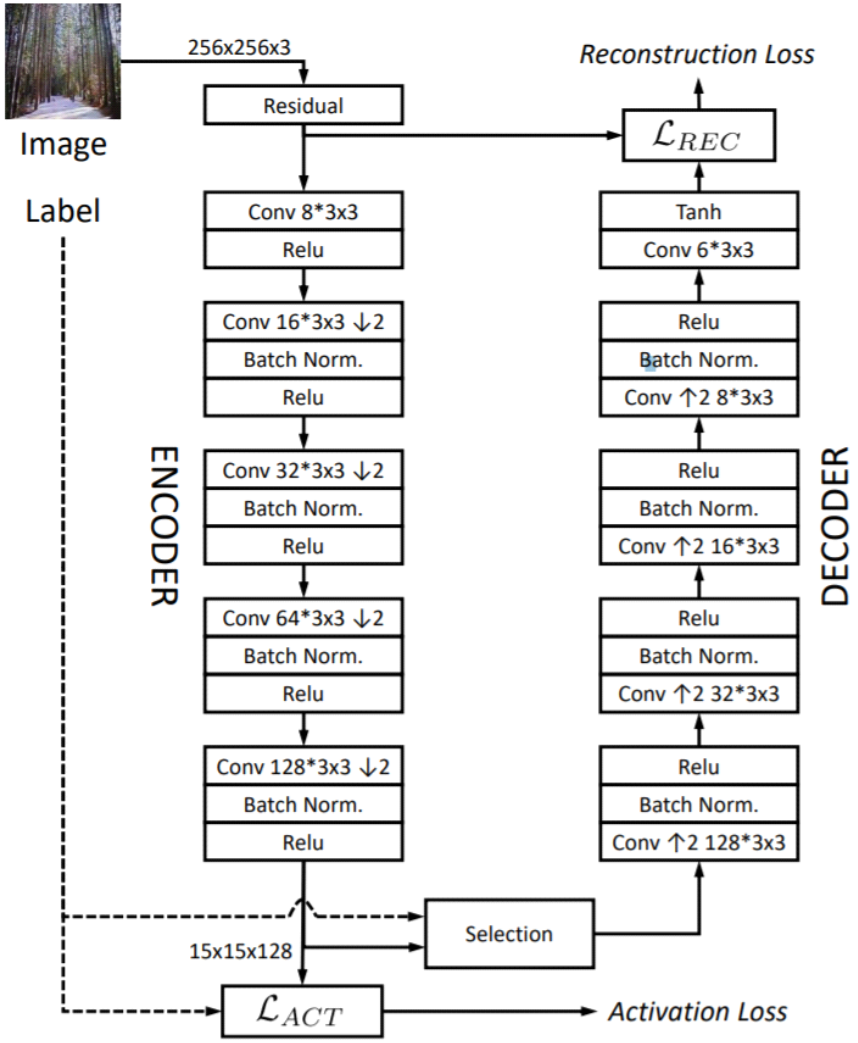
(1) 모델에 원본 도메인 학습시킴

※이미지 생성 방법이 달라야 함!

(2) 학습시킨 모델에 목표 도메인을 아주 소량 학습시킴

(3) 두 가지 생성 방법을 학습 모델 평가에 사용

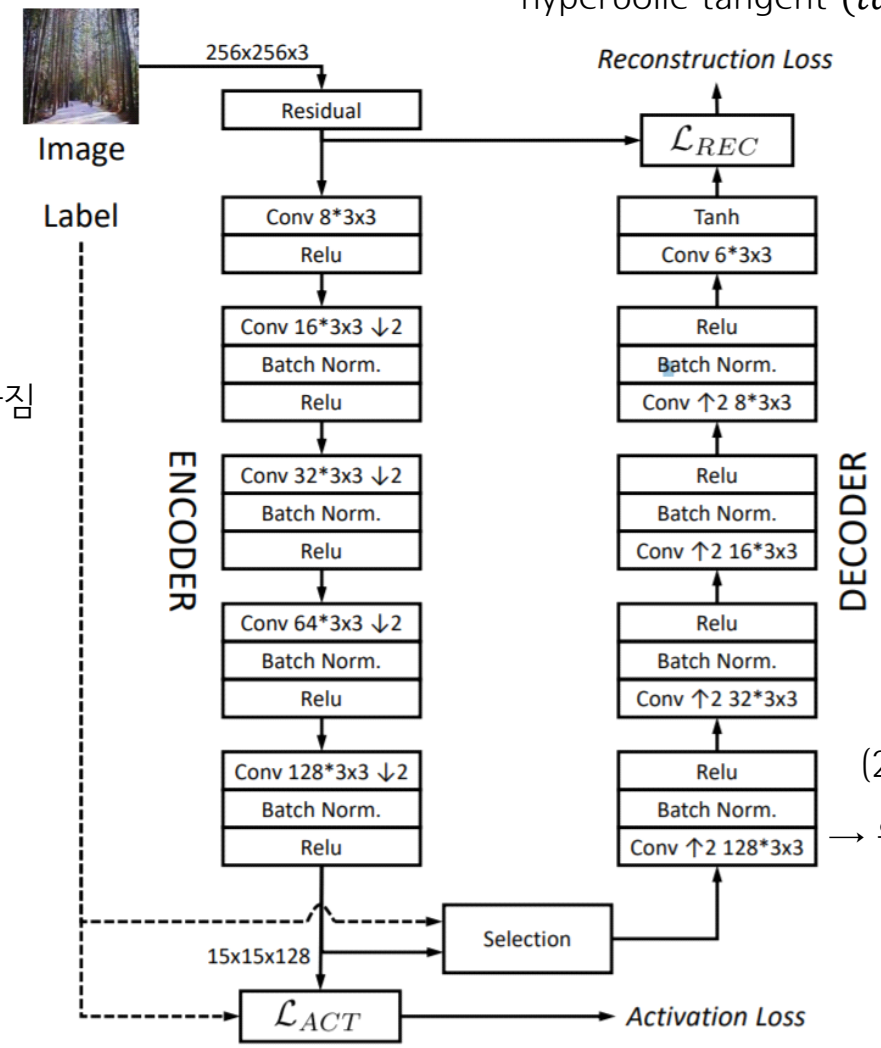
데이터 기반 딥페이크 탐지 알고리즘(1) ForensicTransfer



<신경망 구조>

- 오토인코더 사용
- 인코더 구조=디코더 구조
  =3x3 5개의 합성곱층으로 구성

데이터 기반 딥페이크 탐지 알고리즘(1) ForensicTransfer



Stride=2

→ 1/16해상도 가짐

(2x2)up-sampling

→ 원래 해상도로 복원

<신경망 구조>

1) 이미지 추출하여 *잔차(Residual) 학습
2) 인코더에 이미지 넣어 학습 진행
3) Selection block: 학습하는 데이터의 클래스와 맞지 않는 부분을 0으로 설정

Ex) If 학습 데이터=진짜 클래스

가짜 클래스와 연관된 특징맵 값 0으로 설정

4) 디코더에 latent space 넣어 이미지 재구성.

*잔차: 실제 출력 변수-예상 출력 변수간의 차

데이터 기반 딥페이크 탐지 알고리즘(1) ForensicTransfer

단점

- 목표 도메인 재학습X → 성능이 잘 나오지 않음
- 모델의 목표 도메인 데이터를 알고 있어야 함
  → 현실적으로 사용하기 힘든 기법

데이터 기반 딥페이크 탐지 알고리즘(2) T-GD

<이론적 배경>

1) 전이학습: 기존의 잘 훈련된 모델 활용 → 유사한 문제 해결

- 이미 학습된 가중치 활용 → 새로운 모델 빠르게 학습 가능

2) Self training: teacher 모델 예상 결과 활용하여 student 모델 학습
→ 도메인 적응 높임

- Student 모델 학습 시 노이즈 주입 → 과대적합 방지

데이터 기반 딥페이크 탐지 알고리즘(2) T-GD

규제, 데이터 증강기법(data argumentation), self-training, 학습 전략 활용

→ 전이 능력 향상시킴

⇒ 새로운 기법의 이미지가 들어와도 금방 모델을 학습하여 딥페이크 영상을 탐지할 수 있게 하는 기법

데이터 기반 딥페이크 탐지 알고리즘(2) T-GD

<학습 절차>

(1) CNN기반의 딥페이크 탐지 모델 학습

- 학습한 모델을 pre-trained model로 사용

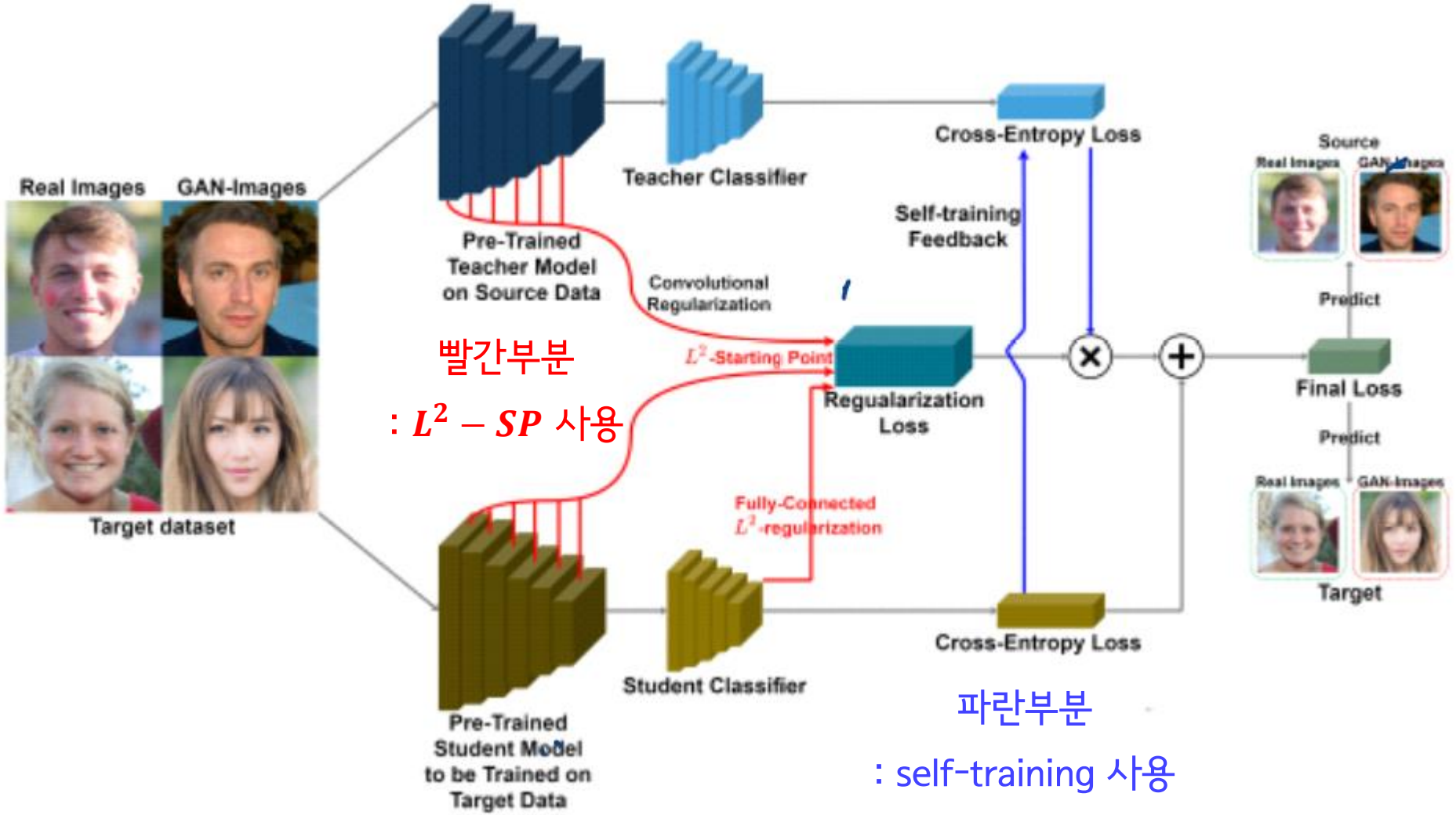
(2) $L^2 - SP$ 적용 → 전이학습 진행

- Pre-trained model 가중치 손상 방지O

(3) 전이학습 프레임워크를 self training 프레임워크로 변환

- pre-trained model:목표모델 → teacher모델:student모델
- 전이학습 수행 시 이미지에 노이즈 섞어 학습 → 과대적합 방지

데이터 기반 딥페이크 탐지 알고리즘(2) T-GD



데이터 기반 딥페이크 탐지 알고리즘(2) T-GD

장점
- 전이 능력이 높음
- 원본 데이터셋에 대한 정보손실X

감사합니다