



2018 인공지능 챌린지 정리 자료

팀 해치 팀장 김형준

프로젝트 개요

프로젝트 명 : 딥 러닝을 활용한 합성 사진 탐지 (2018 인공지능 R&D 챌린지 출품작)

진행 기간 : 2018. 4 ~ 2018. 7

소스 코드 : https://github.com/2alive3s/Fake_image

핵심 기능 : 이미지 데이터 셋에서 임무 별로 주어진 합성 이미지를 탐지

- 임무 1 : 인공지능이 생성한 이미지 탐지
- 임무 2 : 합성된 이미지 탐지

최종 결과 : 전체 47 개 팀 중 12 위

- 임무 1 AUROC : 0.8953
- 임무 2 AUROC : 0.5489

임무 1 문제 정의

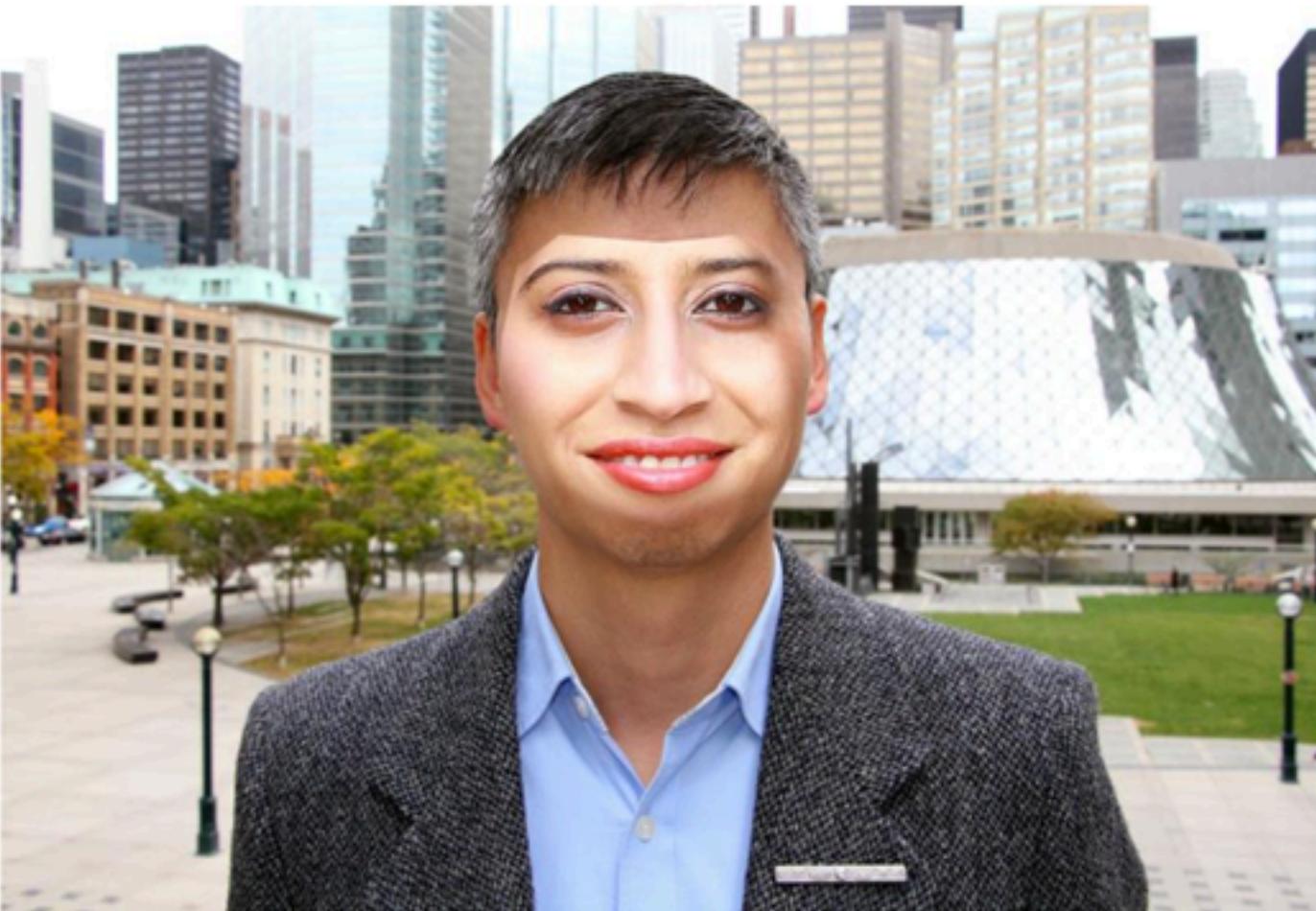
- 임무 1 정의 : 진짜 얼굴 이미지와 GAN 기술을 통해 합성된 얼굴 이미지 판별
- 출제 이미지 크기 : $64 \times 64 \sim 1024 \times 1024$
- 이미지 합성에 사용된 기법 : nVidia Progressive GAN 알고리즘을 통한 생성



임무 1 합성 이미지 예시

임무 2 문제 정의

- 임무 2 정의 : 진짜 얼굴 이미지와 얼굴 일부 합성, 얼굴 교체 등의 방법으로 합성된 이미지 판별
- 출제 이미지 포함 인물 수 : 1~4명
- 출제 이미지 크기 : 제한 없음
- 이미지 합성에 사용된 기법 : 사람이 직접 수작업



임무 2 합성 이미지 예시

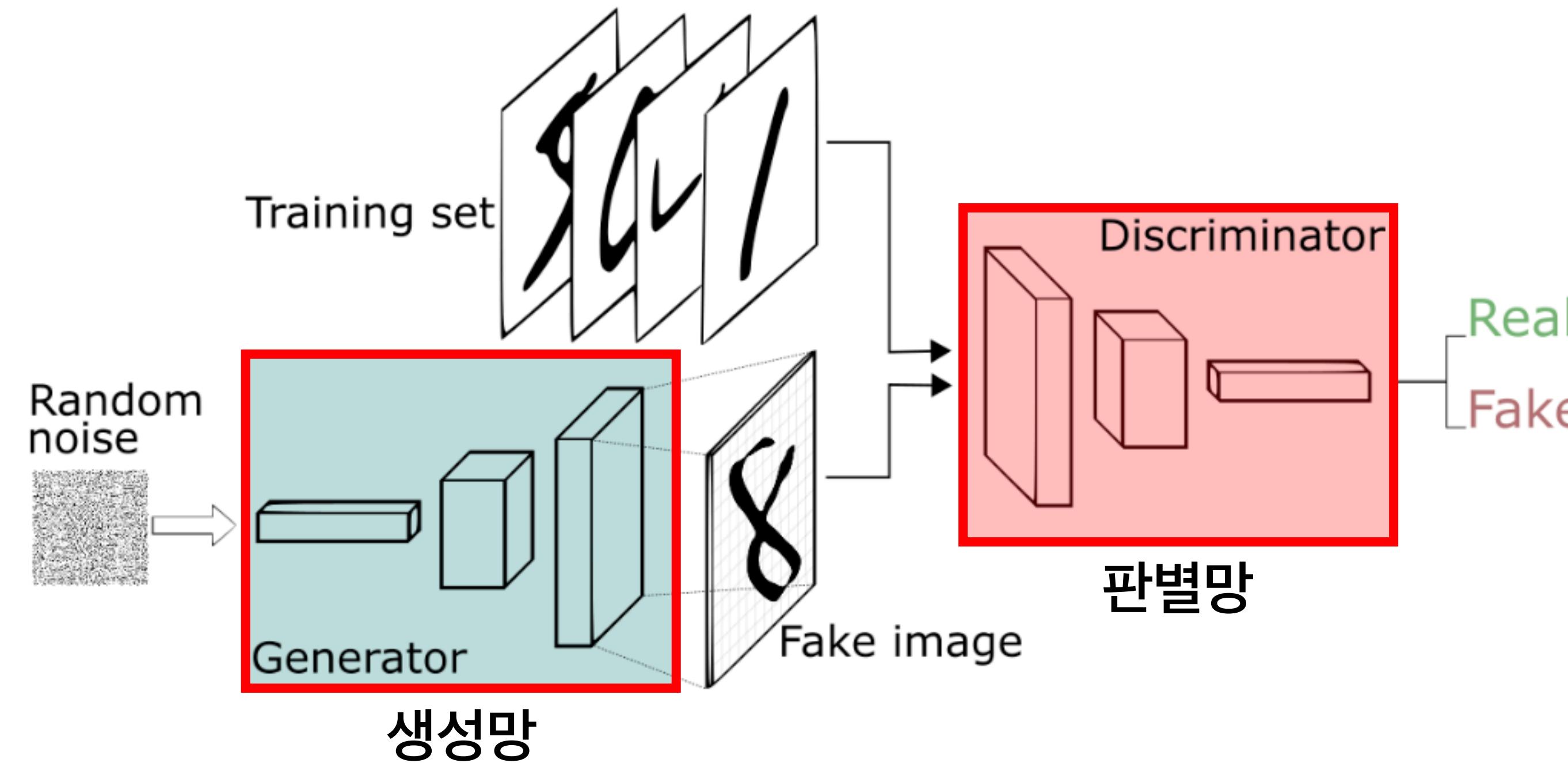
임무 별 풀이 방식

- 임무 1 : 이미 학습된 GAN 모델을 가져와 합성 이미지를 생성
이렇게 생성한 합성 이미지와 실제 이미지를 구분하는 CNN 신경망 학습
- 임무 2 : 얼굴 전체 혹은 일부분을 뒤봐꿔주는 알고리즘을 개발하여 합성 이미지 생성
이렇게 생성한 합성 이미지와 실제 이미지를 구분하는 CNN 신경망 학습

임무 1 세부 분석

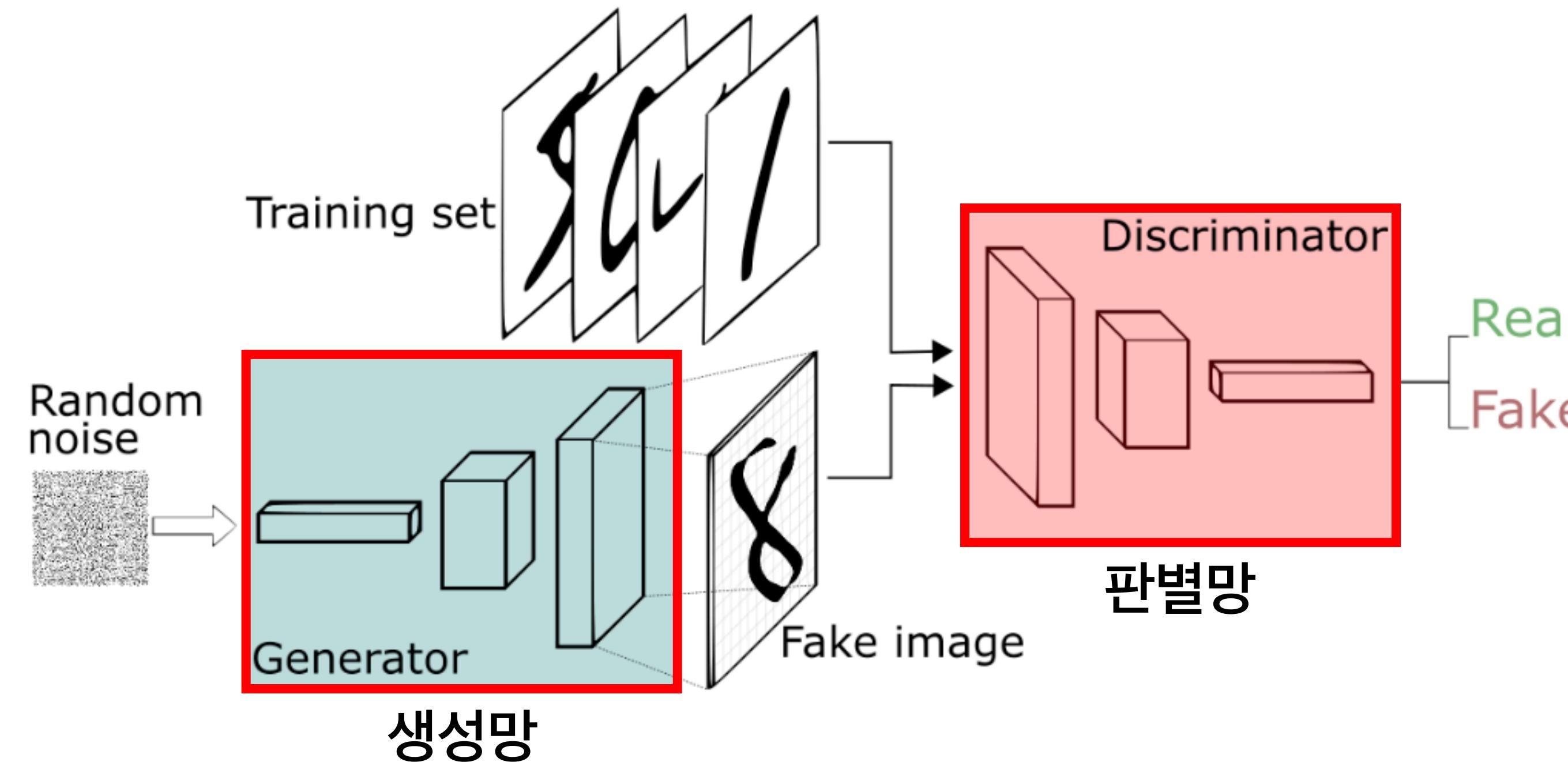
GAN : Generative Adversarial Network의 줄임말

- 합성 이미지를 만드는 생성망과 합성과 진짜를 구별하는 판별망으로 구성



임무 1 세부 분석

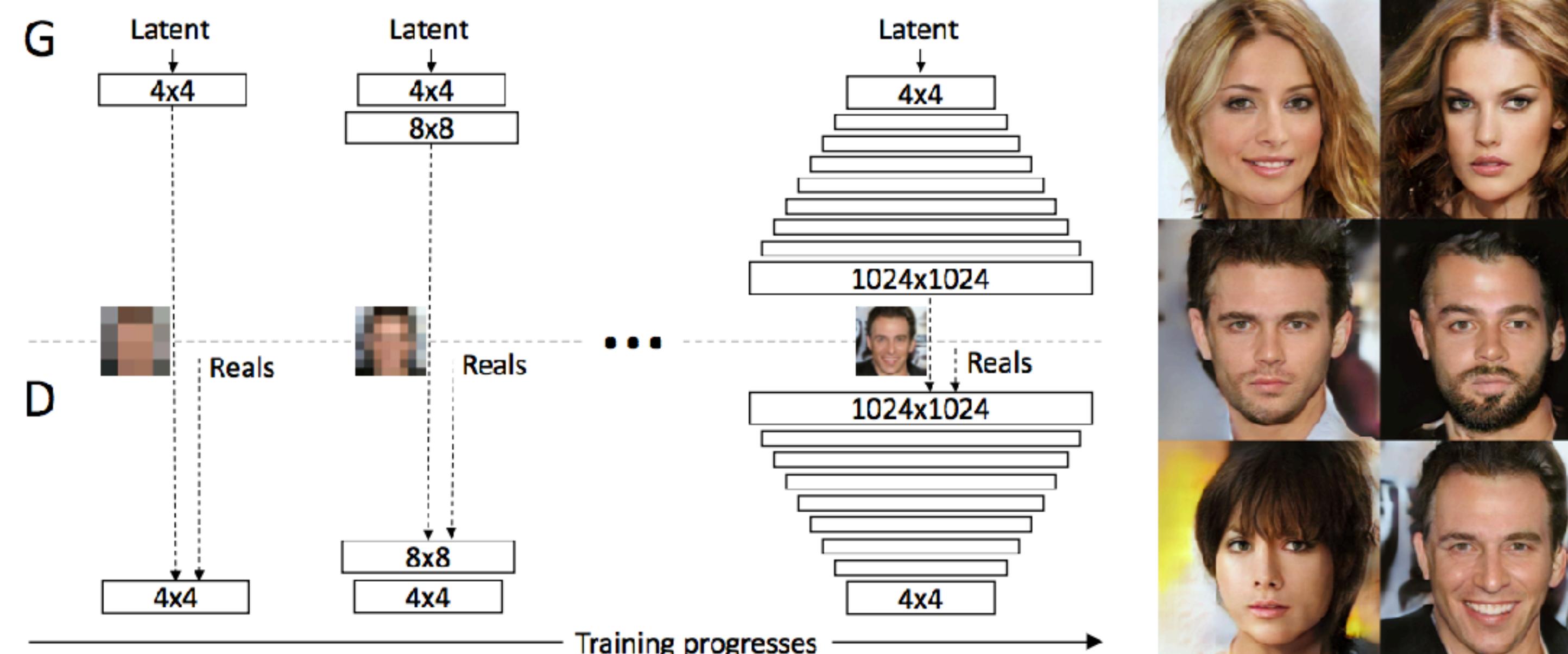
- 생성망은 판별망을 속일 수 있도록 점점더 정교한 합성 이미지를 만들어내고,
판별망은 생성망이 만든 이미지와 진짜 이미지를 점점더 정확히 구별할 수 있도록 학습 진행
- 그 결과 생성망은 진짜인지 가짜인지 구별이 힘들만큼 정교한 합성 이미지를 생성 가능



임무 1 세부 분석

Progressive GAN : nVidia가 2018년 2월에 발표한 정교한 얼굴 합성이 가능한 GAN

- 생성망과 판별망 모두 4×4 의 낮은 해상도 이미지로부터 학습 시작
- 같은 이미지를 8×8 , \dots 1024×1024 와 같이 해상도 높여가며 학습시키는 것이 특징



임무 1 데이터셋 구성

주최 측 출제 방식

- 주최 측은 $64 \times 64 \sim 1024 \times 1024$ 크기의 합성 생성망을 사용하는 것으로 추정
- 이렇게 생성한 합성 이미지와 celebA 데이터셋을 섞어서 임무 1 문제를 출제
(celebA 데이터셋이란 다양한 특징의 얼굴 이미지 202599개를 포함한 데이터셋)

이에 대응하기 위한 데이터셋 구성

- 학습된 Progressive GAN 모델을 사용하여 256×256 크기의 합성 이미지 202599 장 생성
- celebA 데이터셋을 확보하여 진짜 얼굴 이미지 202599장 확보
- 이를 모델 학습 시에 각각 64×64 , 128×128 , 256×256 크기로 조절시켜서 모델 학습 진행

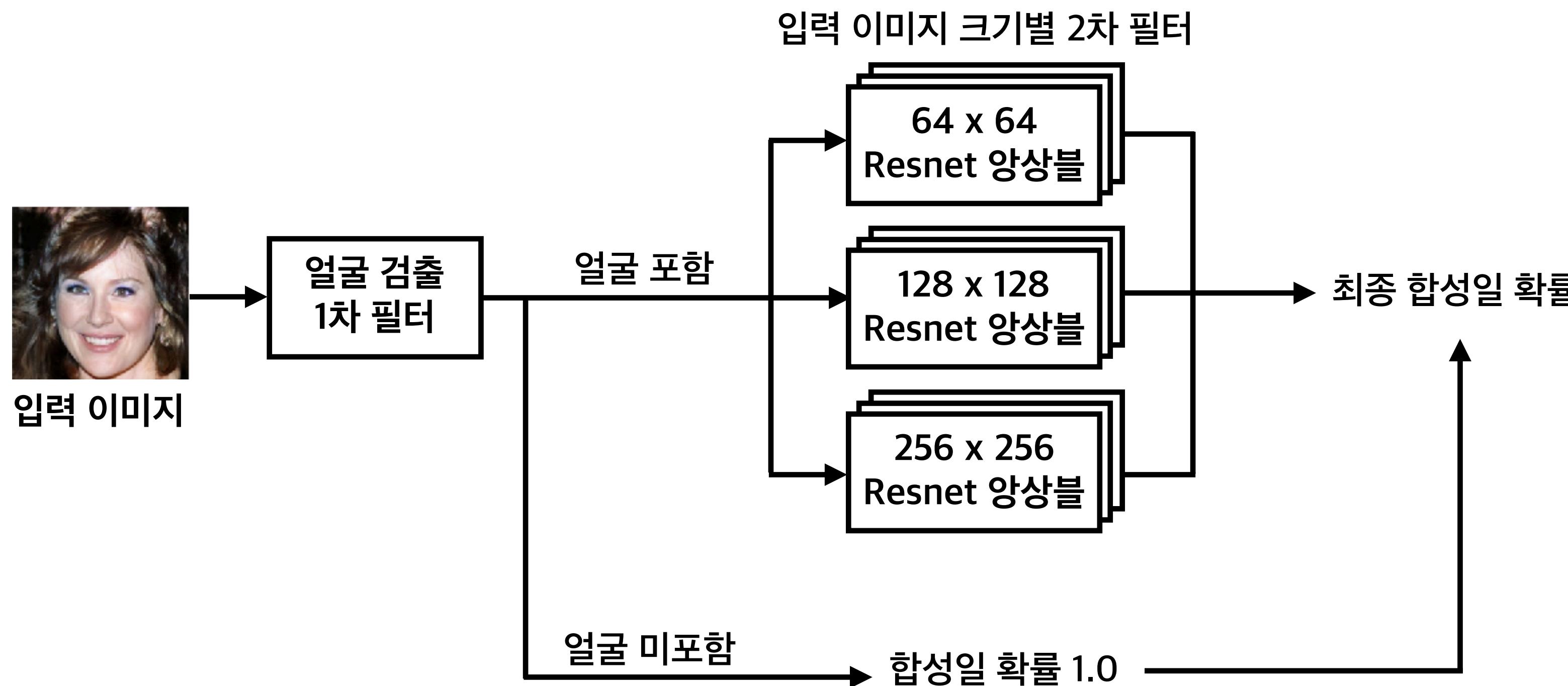
임무 1 접근 방식

기본 접근 방식 : Progressive GAN을 통해 생성된 이미지들을 가짜로 하고,
celebA 데이터 셋 이미지들을 진짜로 하여 이 둘을 구분하는 CNN을 학습시킨다.

고려한 특이 사항 : 출제되는 이미지의 크기는 64x64 부터 1024x1024까지 다양하다.
이미지의 크기를 키우거나 축소할 경우 특징이 손실될 우려가 있다.
따라서 입력 이미지의 크기 별로 합성 여부를 탐지하는 신경망 모델을 구축한다.

임무 1 모델 구성

- 1차 필터 : 얼굴이 포함되어 있는지 여부를 확인하는 신경망
- 2차 필터 : 입력 이미지 크기 별로 합성 이미지와 진짜 이미지를 판별하는 신경망 양상을
- 두 가지 필터를 통과하여 입력 이미지가 합성일 확률을 산출
- 2차 필터에서는 양상을 구성하는 각 모델별 예측 확률의 평균 값을 최종 확률로 산출

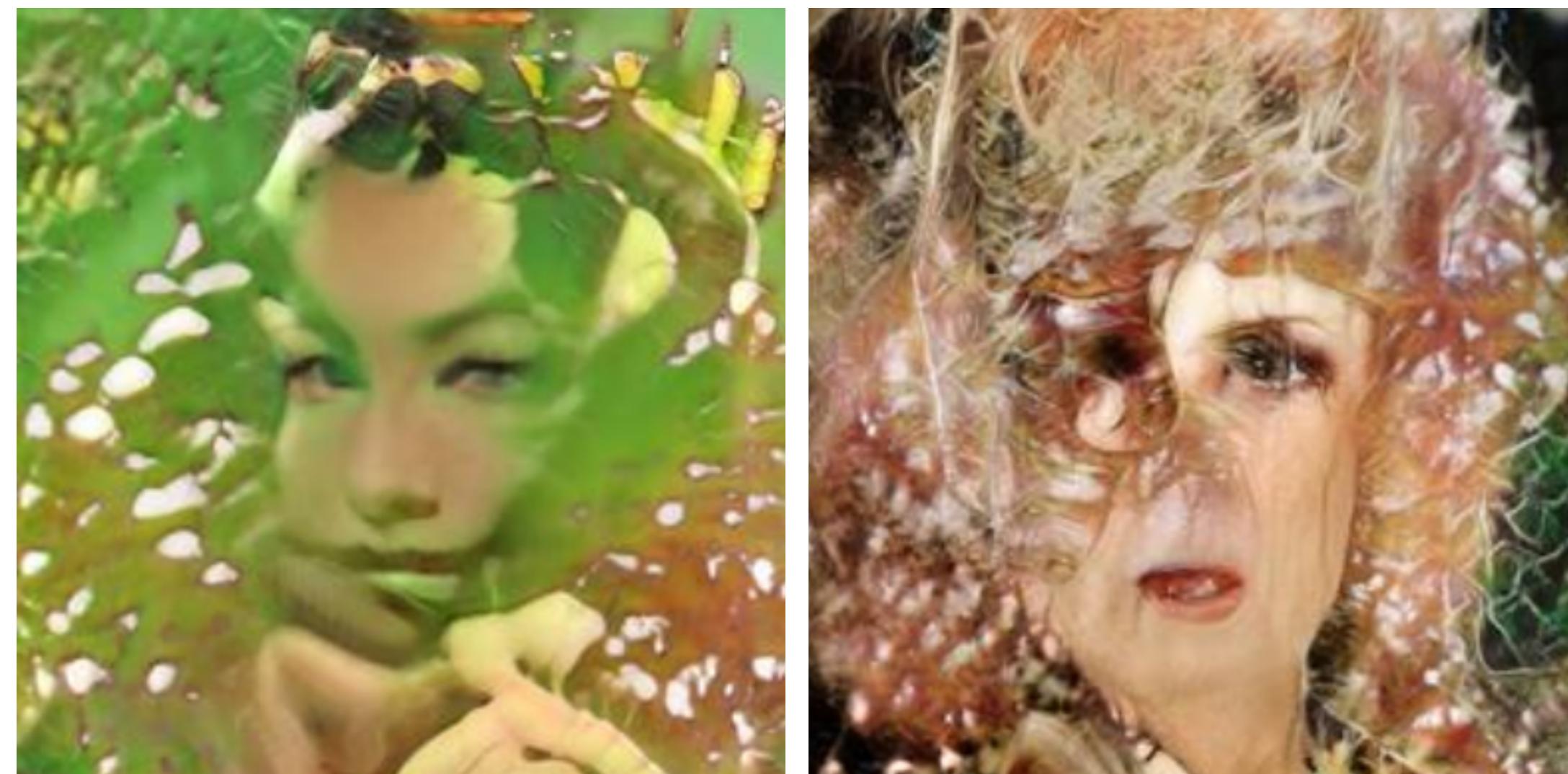


임무 1 1차 필터

입력 이미지에서 얼굴 포함 여부를 확인하는 1차 필터를 통과하는 이유는

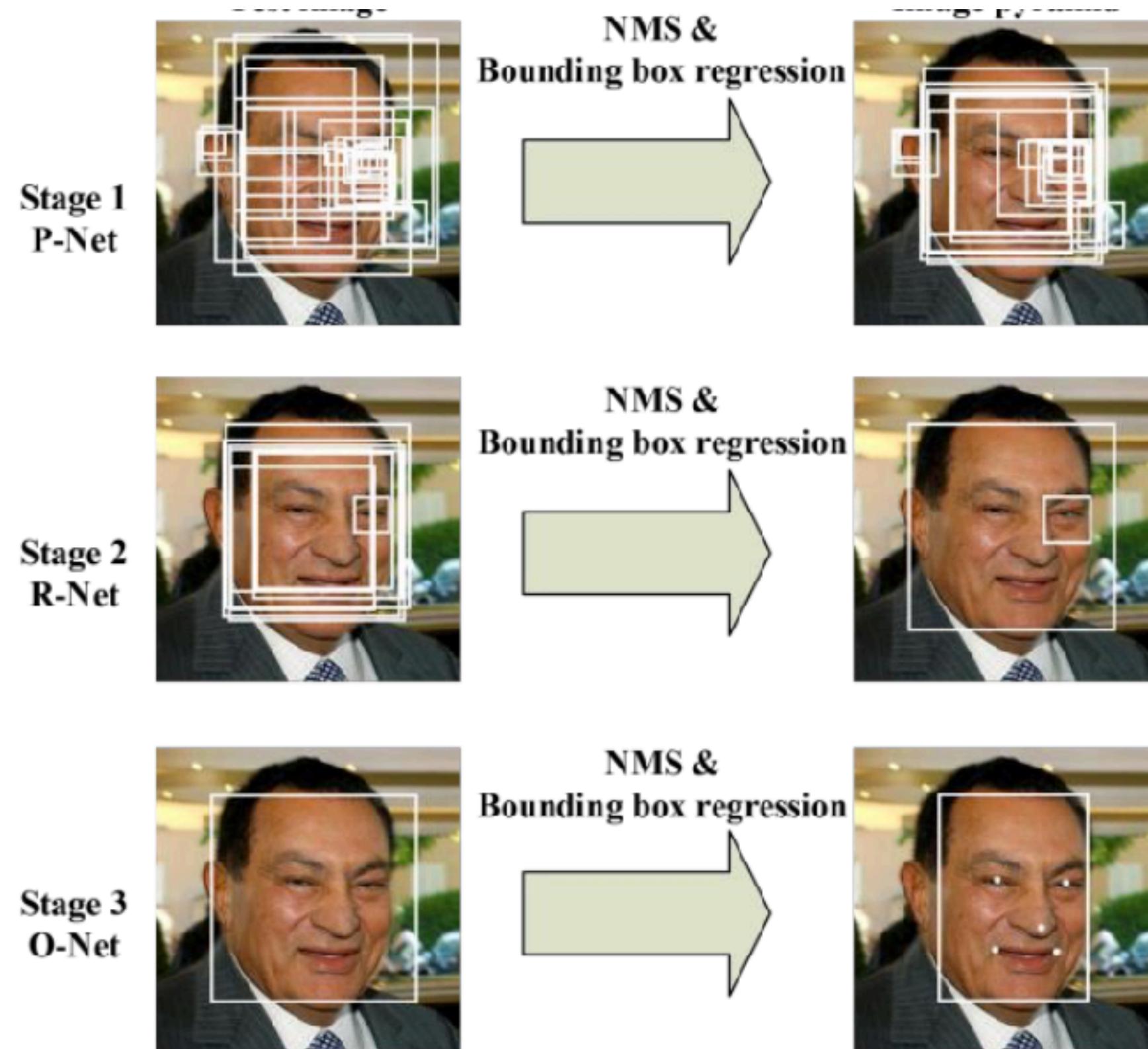
GAN을 통해 생성된 이미지들 가운데 사람 얼굴로 보기 어려운 이미지들을 탐지해내기 위함

또한 이를 통과한 정교한 합성 이미지들로만 2차 필터를 학습시켜 정확도를 높이기 위함



얼굴을 생성해내지 못한 이미지 예시

임무 1 1차 필터



MTCNN 구조

1차 필터에는 MTCNN 모델을 사용

이는 P-Net, R-Net, O-Net 세 가지 CNN을 사용하여 얼굴 검출

- P - Net : 얼굴로 추정되는 후보군 영역 검출
- R - Net : 후보군들을 간추림
- O - Net : 최종 얼굴 영역 및 얼굴 주요 지점 좌표 반환

임무 1 2차 필터

- 2차 필터는 각 크기 별로 합성과 진짜를 구별하도록 직접 CNN 학습 진행
- 모델 학습 세부 사항은 아래와 같다

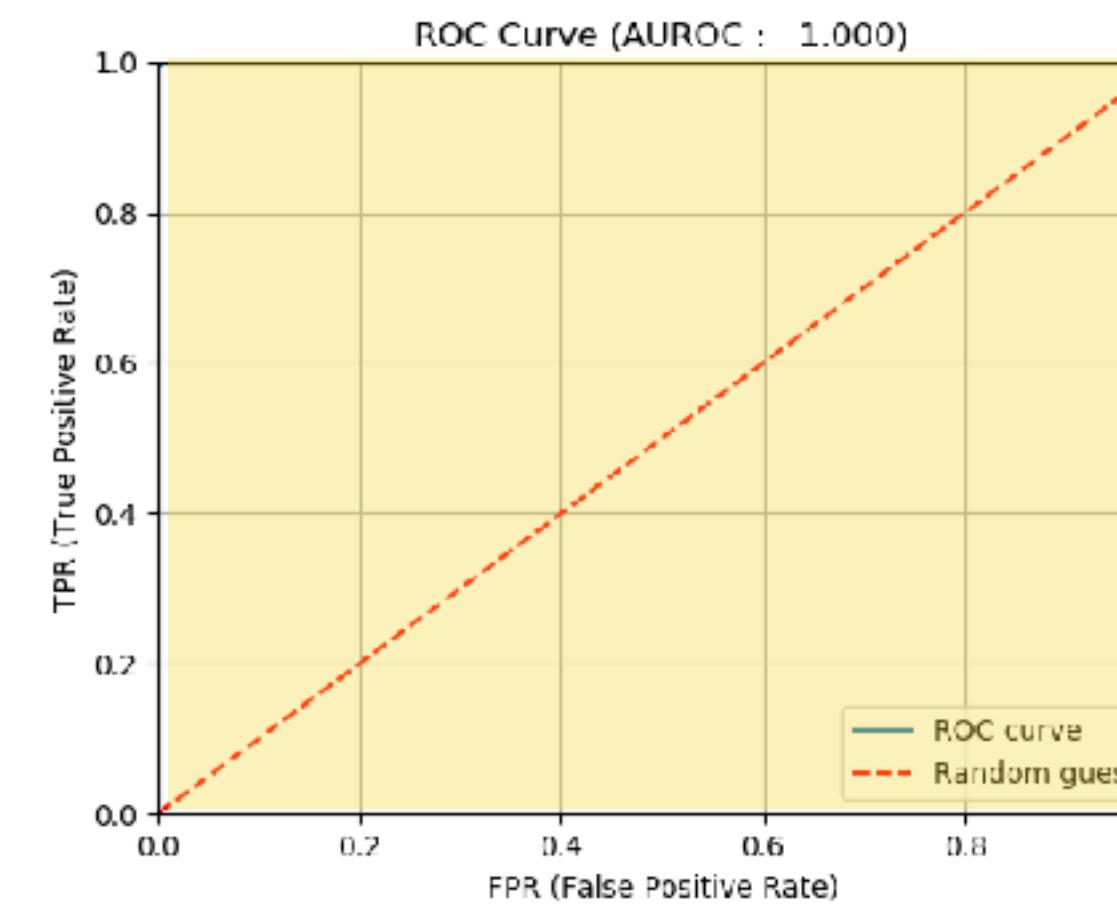
CNN 아키텍쳐	Resnet 50
학습 클래스 수	2 (진짜 / 가짜)
학습 데이터 수	약 32만 장
테스트 데이터 수	약 8 만 장
학습 epoch 수	1
양상을 구성 모델 수	3

임무 1 2차 필터

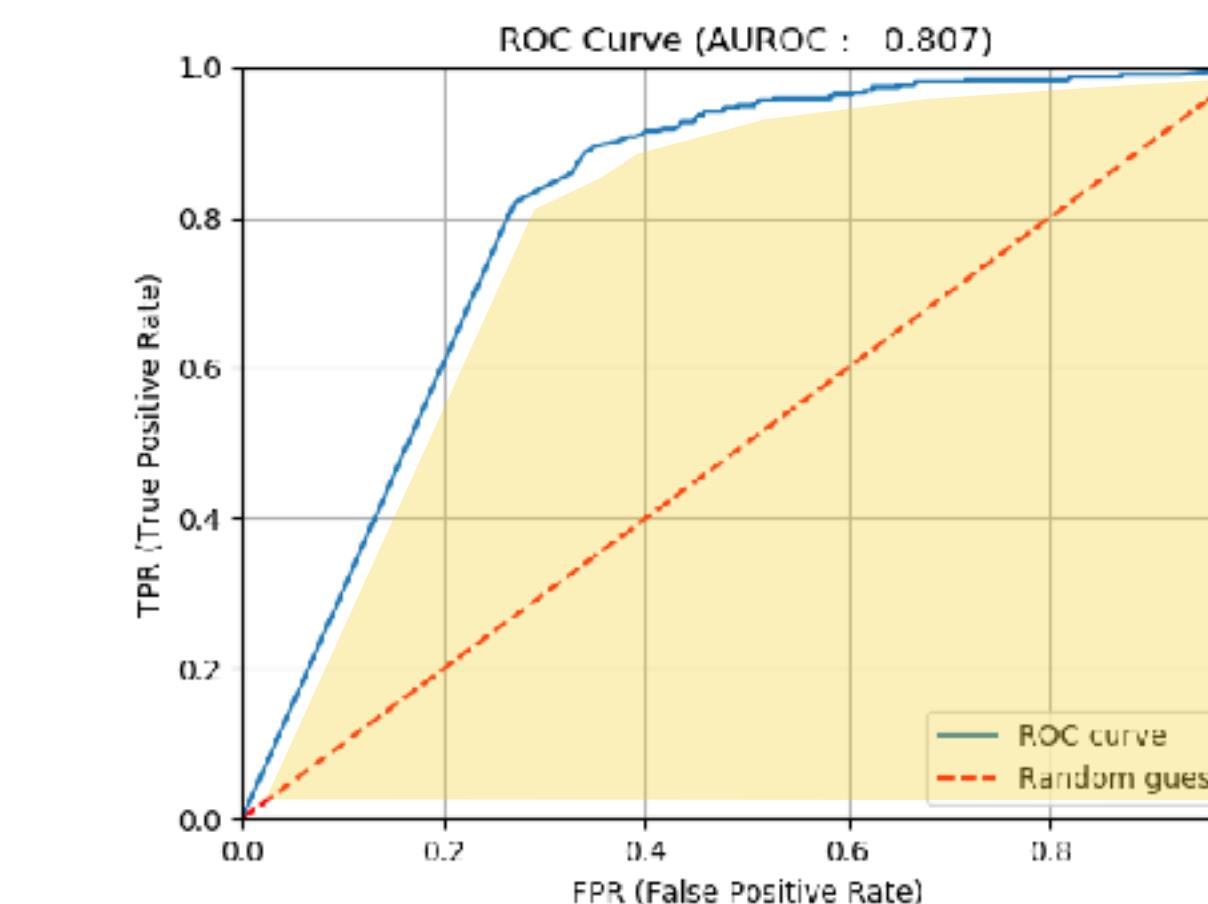
데이터 셋을 1 epoch만 학습시키는 이유는 과적합 현상을 우려해서이다.

학습을 마친 뒤 자체 테스트에서는 AUROC 1.0에 수렴하지만

주최측 테스트에서는 0.807로 낮아서 추가 학습은 과적합 현상을 심화시킬 것으로 판단



자체 테스트 데이터 셋 AUROC : 1.0



주최측 테스트 데이터 셋 AUROC : 0.807

임무 1 최종 결과 및 한계점

임무 1 최종 AUROC : 0.8953

자체 평가 : 높은 정확도를 보이며 실제 합성 이미지를 탐지하는 모습을 보인다.

하지만 학습 시보다는 낮은 정확도를 보여 과적합의 가능성이 있다고 판단한다.

추후 진행 방향 : GradCam 기술을 적용하여 합성 탐지시 신경망이 주목하는 특징을 확인해보려 한다.

또한 테스트 데이터 셋을 다양하게 만들어 과적합 여부를 검증하려 한다.

임무 2 분석

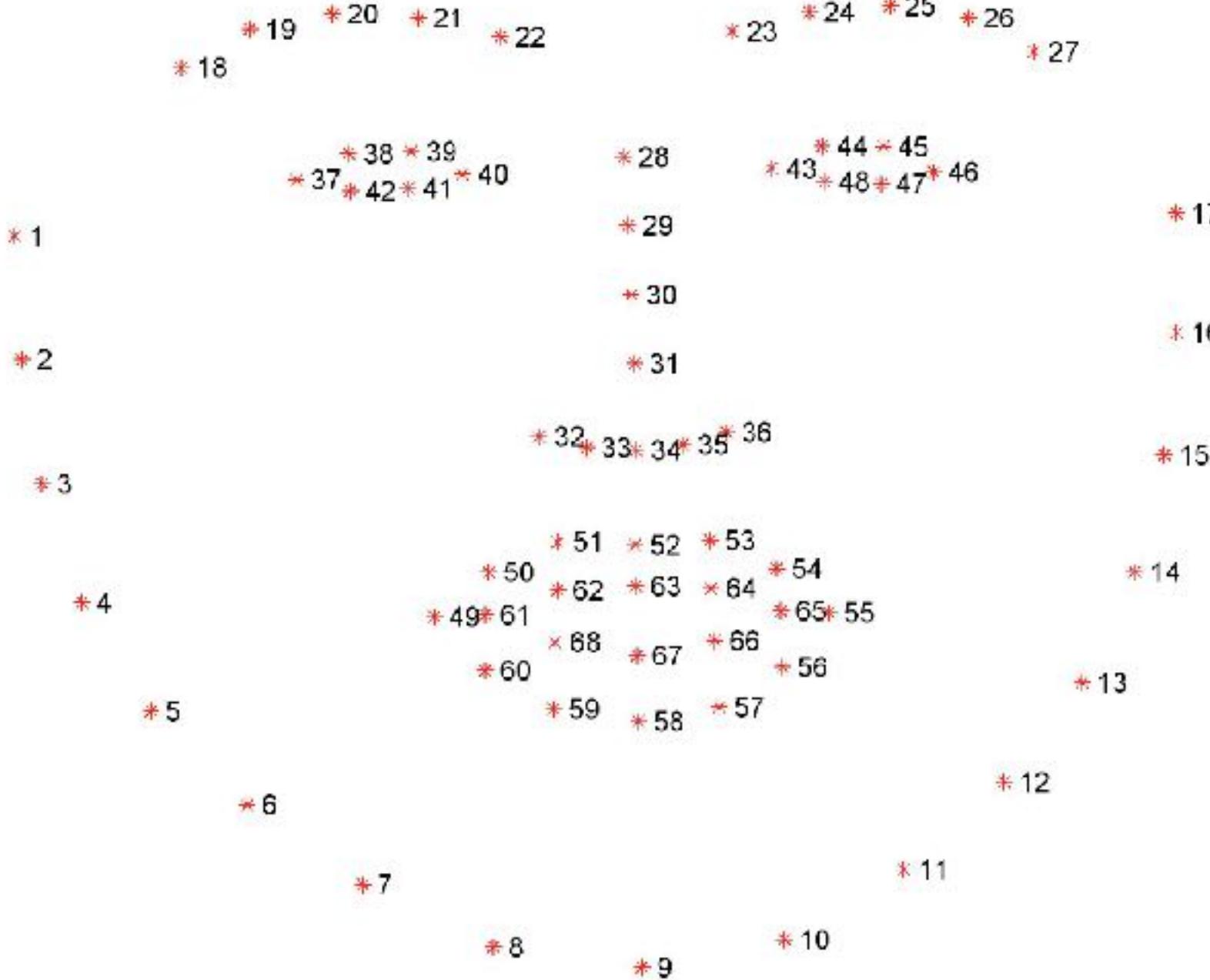
- 임무 2 출제 유형 분석
- 유형별 이미지 예시는 다음의 링크를 통해서 확인 가능 [[link](#)]

출제 유형	합성 범위
얼굴 일부분 바꾸기	눈, 코, 입 등 주요 부위를 하나 이상 선택
얼굴 절반 바꾸기	좌, 우, 상, 하 각 방향의 얼굴 절반
얼굴 전체 바꾸기	눈썹부터 턱 까지 얼굴 안쪽 영역
머리 전체 바꾸기	정수리부터 목 까지 머리 영역

임무 2 데이터 셋 구성

- 1~4명이 포함된 이미지를 수집하고, 얼굴 합성 알고리즘을 통해 합성 데이터 셋을 구성
- 한명이 포함된 이미지는 celebA 데이터 셋을 활용하였다 (202599 장),
- 여러 명이 포함된 이미지는 크롤러를 개발해 아래의 사이트에서 수집하였다. (9715장)
(이미지 수집 출처 : 구글, 네이버, 다음, 플리커, 인스타그램, 텀블러)
- 여러명 포함 이미지 수가 적은 이유는 주최측 출제 이미지와 최대한 유사한
데이터 셋을 구성하기 위해 필터링 기준을 높였기 때문

임무 2 합성 데이터 셋 생성 방식



번호 별 좌표 위치

얼굴 부위 별 합성 이미지를 생성하기 위해서 파이썬 dlib 사용

얼굴의 주요 부위에 해당하는 68개의 좌표를 추출

- 턱 : 1 ~ 17
- 눈썹 : 18 ~ 22 (왼쪽), 23 ~ 27 (오른쪽)
- 코 : 28 ~ 36
- 눈 : 37 ~ 42 (왼쪽), 43 ~ 48 (오른쪽)
- 입 : 49 ~ 68

임무 2 합성 데이터 셋 생성 방식

- 먼저 원본 이미지와 그 위에 합성될 얼굴 이미지를 준비한다



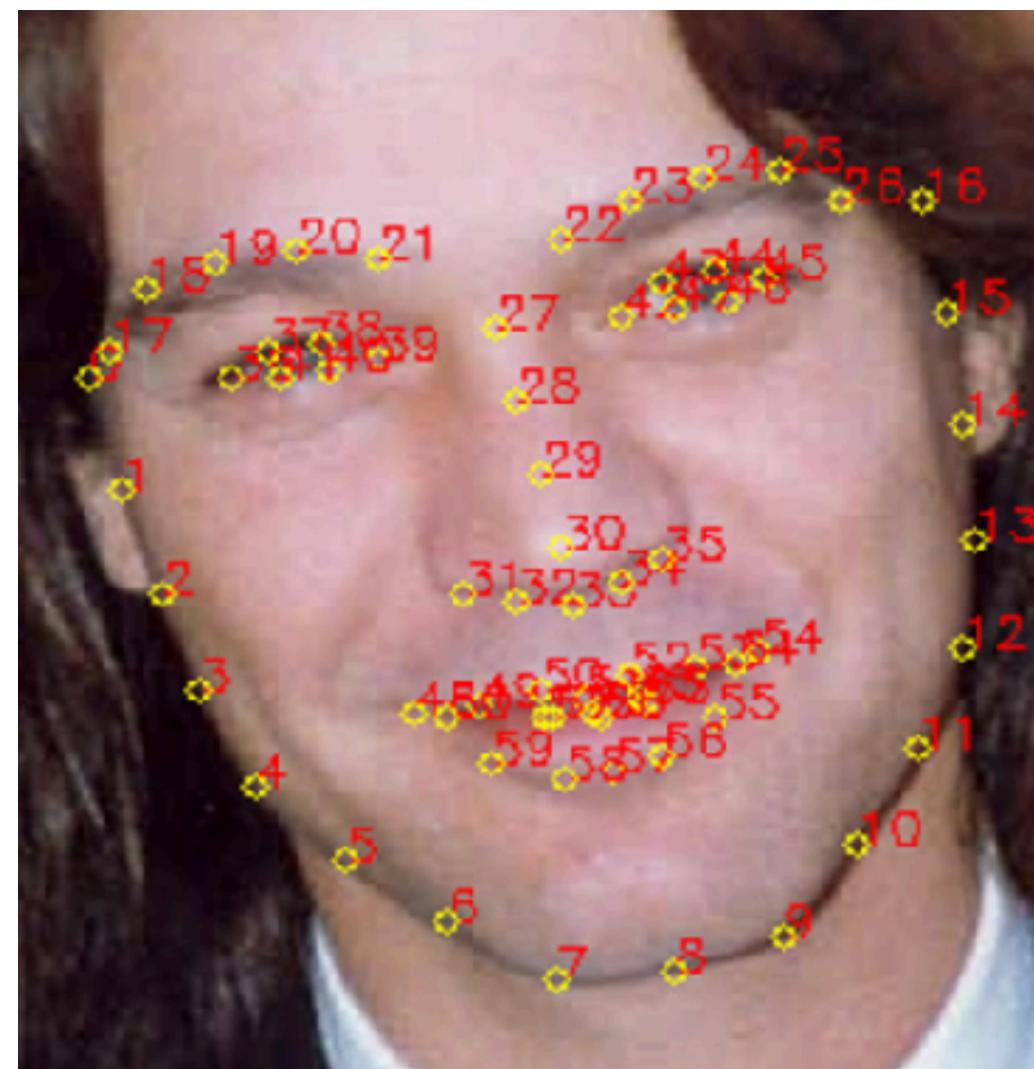
원본 이미지



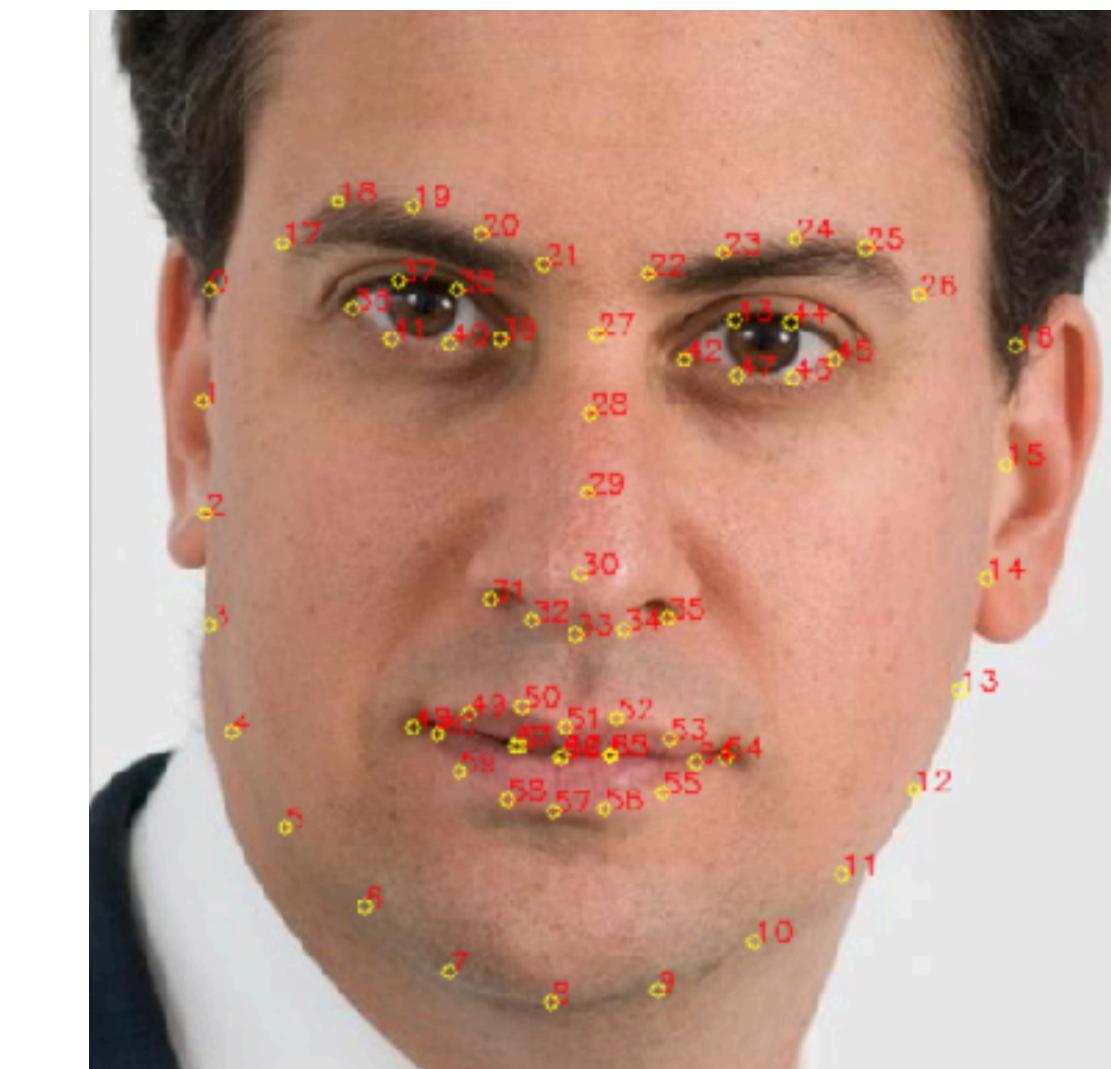
합성할 얼굴 이미지

임무 2 합성 데이터셋 생성 방식

- dlib를 활용하여 각 이미지 별로 얼굴에 해당하는 영역만 크롭한 다음,
- 68개의 얼굴 주요 지점 좌표를 추출한다



원본 얼굴 이미지



합성할 얼굴 이미지

임무 2 합성 데이터 셋 생성 방식

- 이제 합성할 이미지의 얼굴 위치가 원본 이미지 얼굴 위치와 일치하게끔
- 이미지를 회전시켜 준다



원본 이미지



회전된 합성할 이미지

임무 2 합성 데이터 셋 생성 방식

- 회전된 이미지에서 합성하고자 하는 얼굴 부위를 잘라낸다 (ex. 눈, 코, 입)
- 그 다음 원본 이미지 위에 덮어씌워 합성 이미지를 생성한다



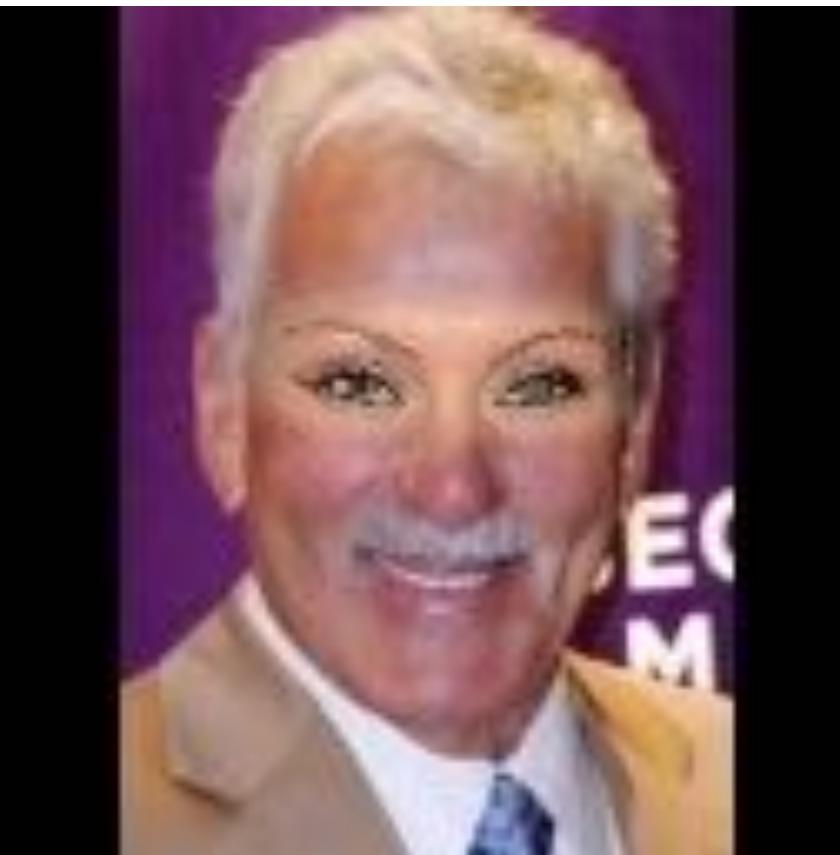
원본 이미지



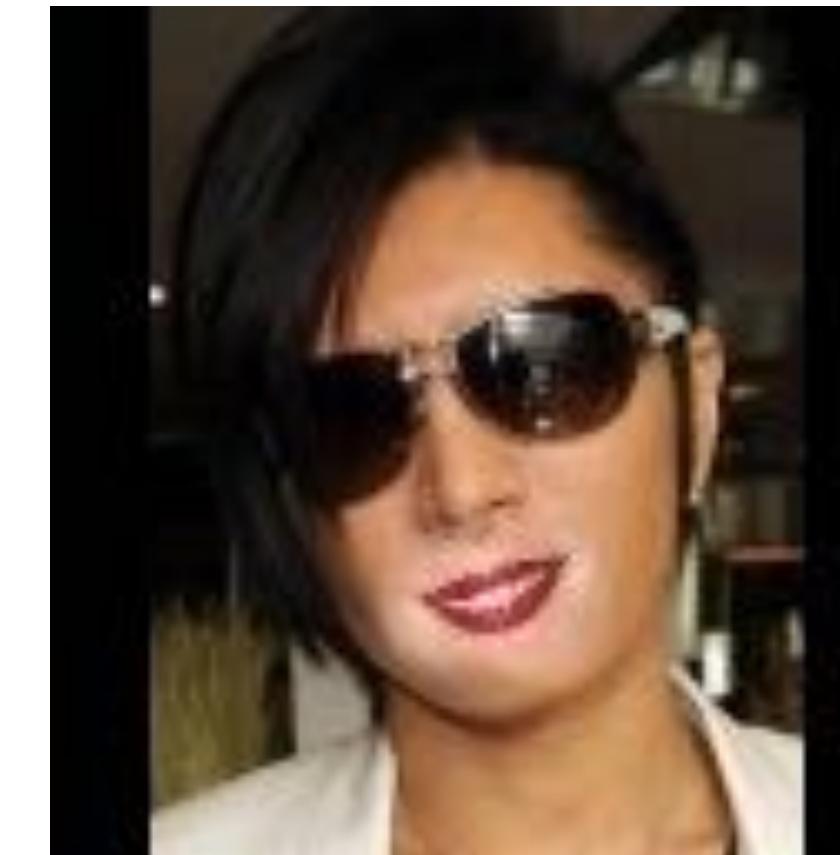
눈, 코, 입 합성 이미지

임무 2 합성 데이터 셋 생성 방식

- 잘라내는 영역을 눈, 코, 입, 얼굴 전체, 얼굴 절반 등으로 설정할 수 있다
- 가장자리 부분을 번짐과 잘림으로 구분하여 합성 데이터 셋을 생성하였다.
- 학습에 사용된 데이터 셋 샘플은 다음 링크에서 확인 가능하다 [[link](#)]



눈 번짐 합성



입 번짐 합성



얼굴 전체 잘림 합성

임무 2 접근 방식

기본 접근 방식 : 입력 이미지에서 얼굴 영역을 먼저 크롭한다.

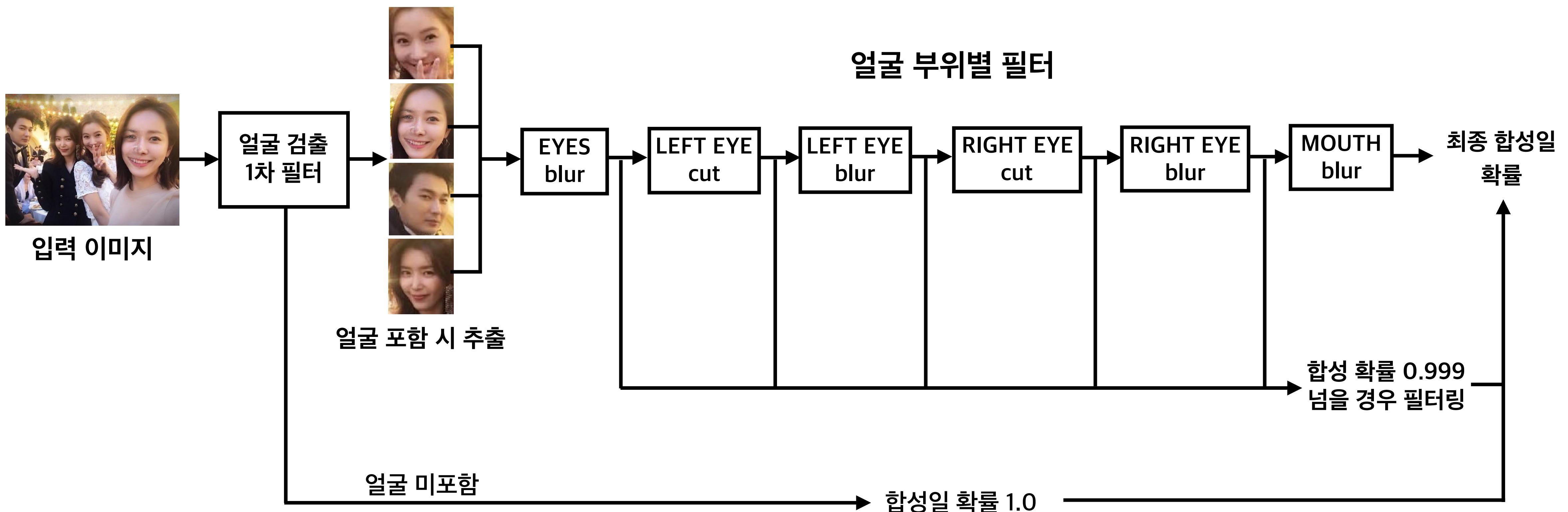
유형별 합성 이미지와 원본 이미지를 구분하는 CNN을 통과시켜 합성인지 판별한다.

고려한 특이 사항 : 눈, 코, 입, 얼굴 절반, 얼굴 전체 등 다양한 유형의 합성 이미지가 출제된다.

따라서 각 부위별로 나누어 CNN을 학습시킨 후, 이를 순차적으로 통과하도록 한다.

임무 2 모델 구성

- 1차 필터 : 얼굴이 포함된 영역을 크롭하는 신경망
- 얼굴 부위별 필터 : 합성된 부위 별로 가장자리 번짐과 잘림 처리를 구분하여 학습시킨 신경망
- 해당 필터들을 순차적으로 통과하면서 입력 이미지가 합성일 확률을 산출



임무 2 1차 필터

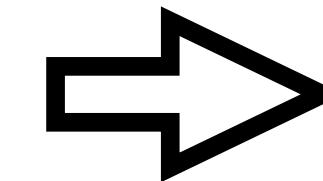
입력 이미지에서 얼굴 영역을 크롭하는 이유는

신경망 학습 시에 얼굴 이외의 배경이 많이 섞이게 될 경우 정확도 하락을 우려해서이다.

따라서 얼굴 영역을 먼저 크롭하며, 임무 1과 마찬가지로 MTCNN을 활용한다.



입력 이미지



크롭된 얼굴 이미지
(크기 : 128x128)

임무 2 얼굴 부위별 필터

- 얼굴 부위별 필터는 부위 별로 합성된 이미지와 진짜를 구분하는 CNN 학습 진행
- 1~5 에포크 학습 시킨 뒤, 정확도가 가장 높은 모델을 선택

CNN 아키텍쳐	Resnet 50
학습 클래스 수	2 (진짜 / 가짜)
학습 데이터 수	약 33만 장
테스트 데이터 수	약 9 만 장
학습 epoch 수	1 ~ 10

임무 2 얼굴 부위별 필터

- 임무 2 2차 필터 경우 데이터를 10 에포크 학습 진행
- 학습 에포크 별 AUROC 정확도 측정 후, 가장 높은 정확도를 보인 에포크 선택
- 그 결과 대부분의 필터가 1~4 에포크 사이에 가장 높은 정확도를 보였다

탐지 영역	EYES blur	LEFT_EYE cut	LEFT_EYE blur	RIGHT_EYE cut	RIGHT_EYE blur	MOUTH blur
선택된 에포크	4	1	1	2	2	4

임무 2 최종 결과 및 한계점

임무 2 최종 AUROC : 0.5489

자체 평가 : 어느 정도 합성을 탐지하긴 하지만, 정확도 측면에서 많이 아쉬운 결과이다.

이는 학습한 이미지와 출제된 이미지 사이의 성격 차이 때문에 발생한 현상으로 추측된다.

추후 진행 방향 : GradCam 기술을 적용하여 합성 탐지시 신경망이 주목하는 특징을 확인해보려 한다.

이를 통해 필터들이 의도한 대로 학습된 영역에 주목하여 합성을 검출하는지 검증할 예정이다

기간 내 진행 못한 부분

진행하지 못한 사안

- 임무 1, 임무 2 GradCam 적용, 신경망이 주목하는 특성 확인
- 임무 2 딥 페이크 활용 합성 이미지 탐지 모델 학습
- 임무 2 머리 전체 합성 이미지 탐지 모델 학습

진행하지 못한 이유

- 임무 2 필터들이 기본적인 눈, 코, 입 합성도 제대로 검출해내지 못하는 상황
- GradCam, 딥 페이크 필터, 머리 합성 필터 구현은 시간이 많이 소요됨
- 한정된 시간 내에 기본적인 필터 구현이 더 중요하다고 판단하여 위 사안들을 진행하지 못함

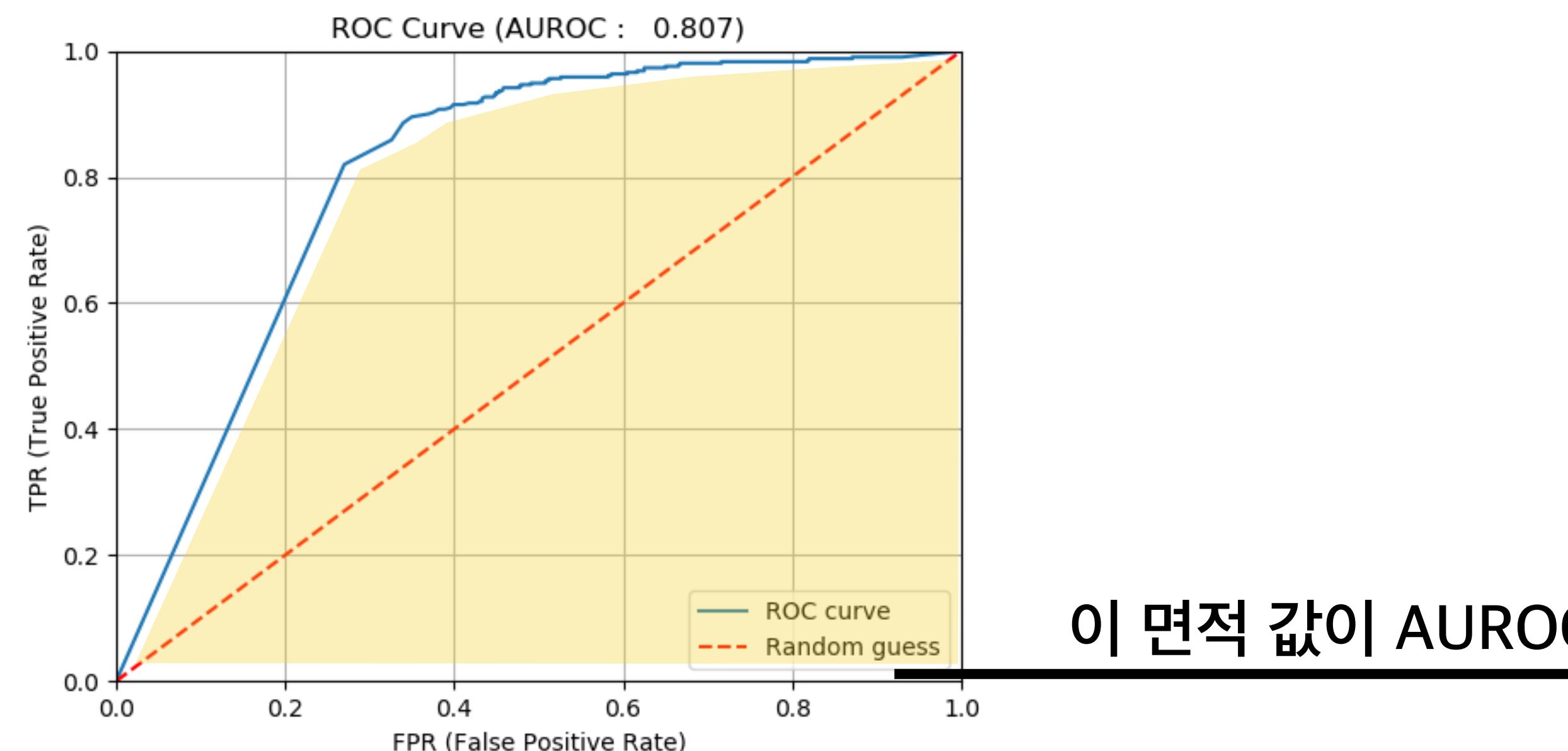
부록 1. AUROC

AUROC : Area under ROC curve의 줄임말로 ROC 곡선의 아래 면적 값을 의미

ROC curve : 합성 사진이라고 잘못 예측한 비율(FPR) 을 0에서 1까지 변화시켰을 때,

합성 사진으로 제대로 예측한 비율(TPR)이 어떻게 변화하였는지를 표현한 곡선

즉, 합성은 합성으로 원본은 원본으로 얼만큼 정확하게 구별하는지를 측정



부록 1. AUROC

TPR : 전체 합성 사진 가운데 합성 사진으로 예측한 비율

FPR : 전체 원본 사진 가운데 합성사진으로 예측한 비율

$$TPR = \frac{TP}{TP+FN}$$

$$FPR = \frac{FP}{FP + TN}$$

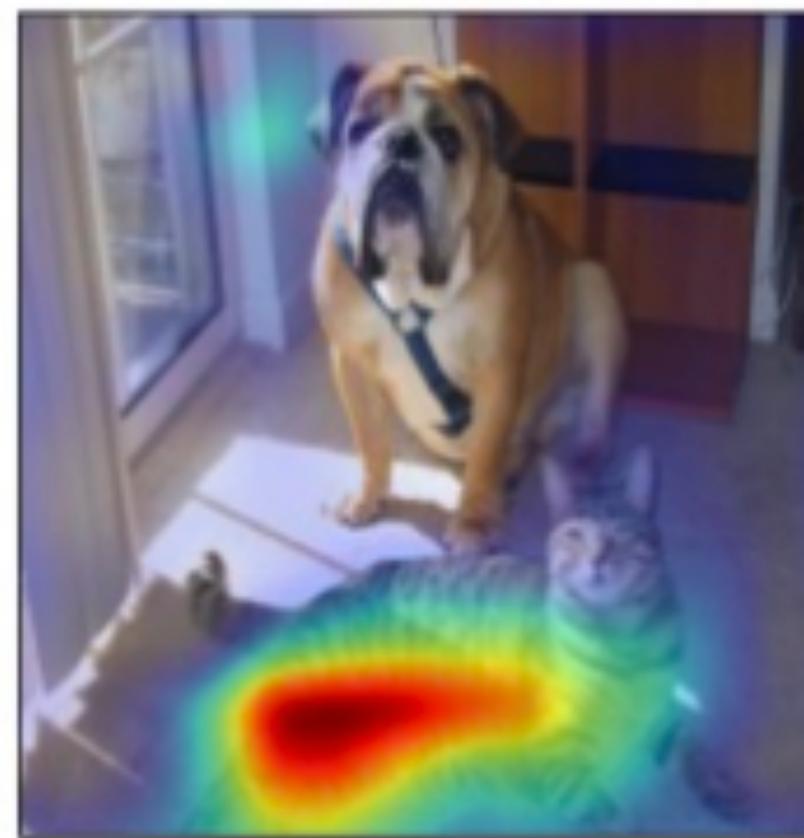
		실제 라벨	
		TRUE (합성 사진)	FALSE (원본 사진)
예측값	Positive (합성 사진)	TP	FP
	Negative (원본 사진)	FN	TN

부록 2. GradCam

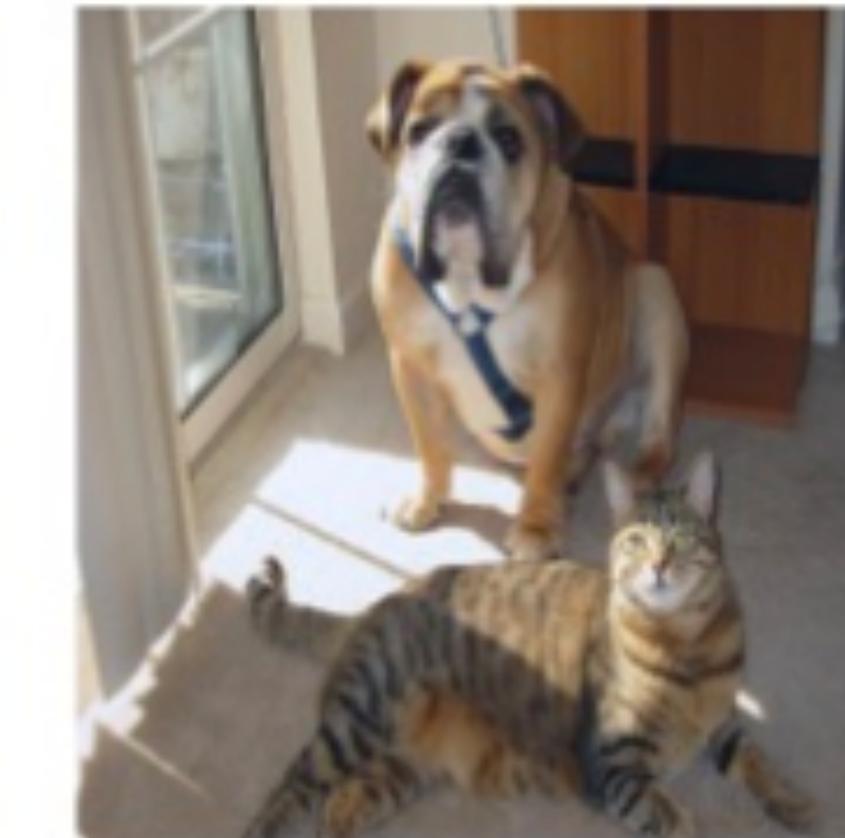
GradCam : 학습된 CNN이 입력된 이미지에서 예측값을 반환할 때

어떠한 부분에 집중하였는지를 시각화 해주는 신경망 기술

Grad-CAM for "Cat"



Grad-CAM for "Dog"



GradCam 적용 사례

부록 2. GradCam

GradCam을 통해서 임무 1과 임무 2 필터들이 합성 여부를 판별할 때
어떤 특징에 주목하여 판단을 하는지를 검증할 예정



GradCam 적용 예상 모습