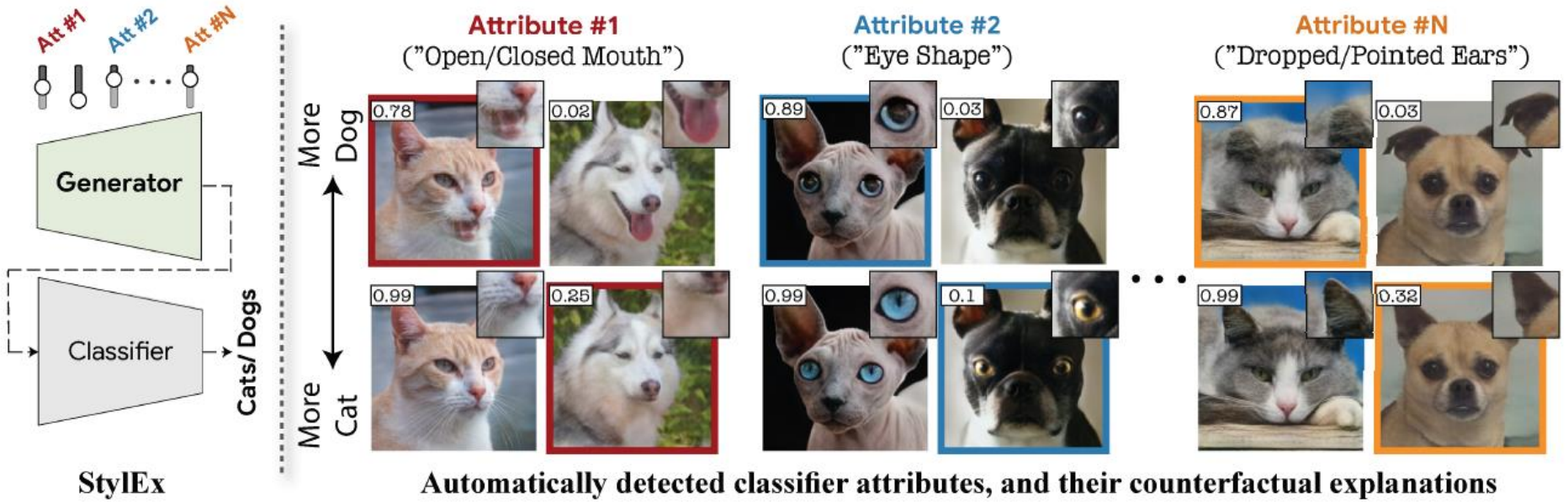
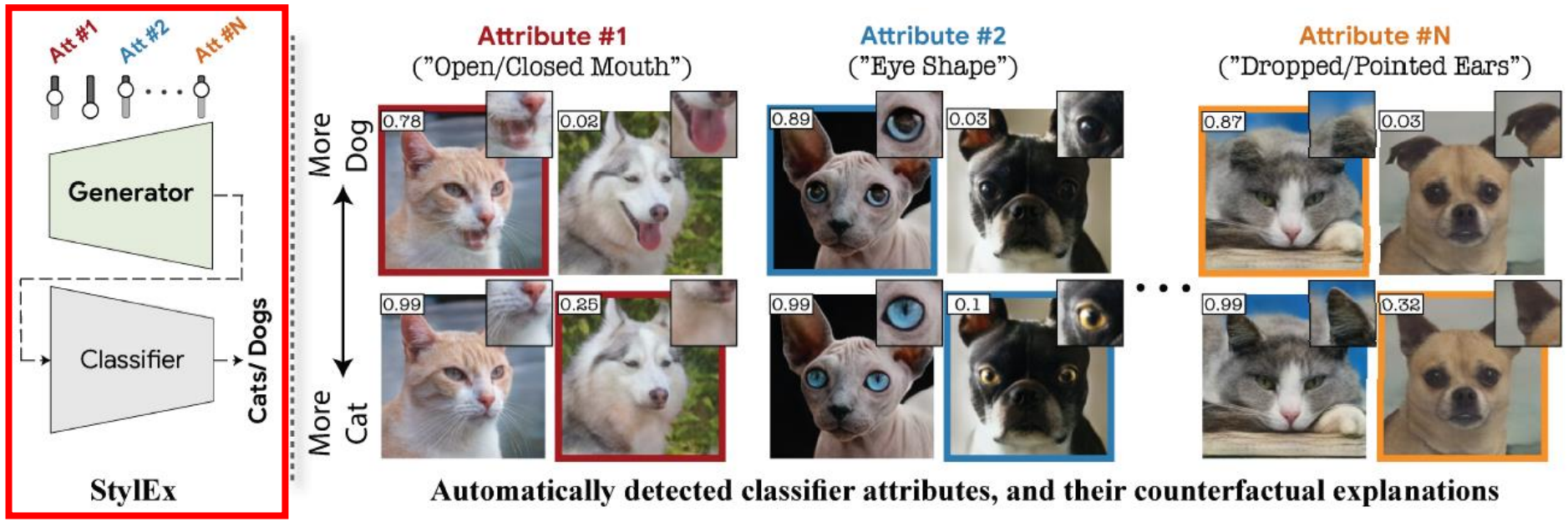
Explaining in Style: Training a GAN to explain a classifier in StyleSpace

Oran Lang[*1] Yossi Gandelsman[*1] Michal Yarom[*1] Yoav Wald[*1,2]
Gal Elidan[1] Avinatan Hassidim[1] William T. Freeman[1,3] Phillip Isola[1,3]
Amir Globerson[1,4] Michal Irani[1,5] Inbar Mosseri[1]

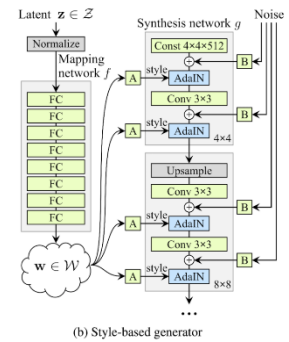
[1] Google Research [2] Hebrew University [3] MIT [4] Tel Aviv University [5] Weizmann Institute of Science

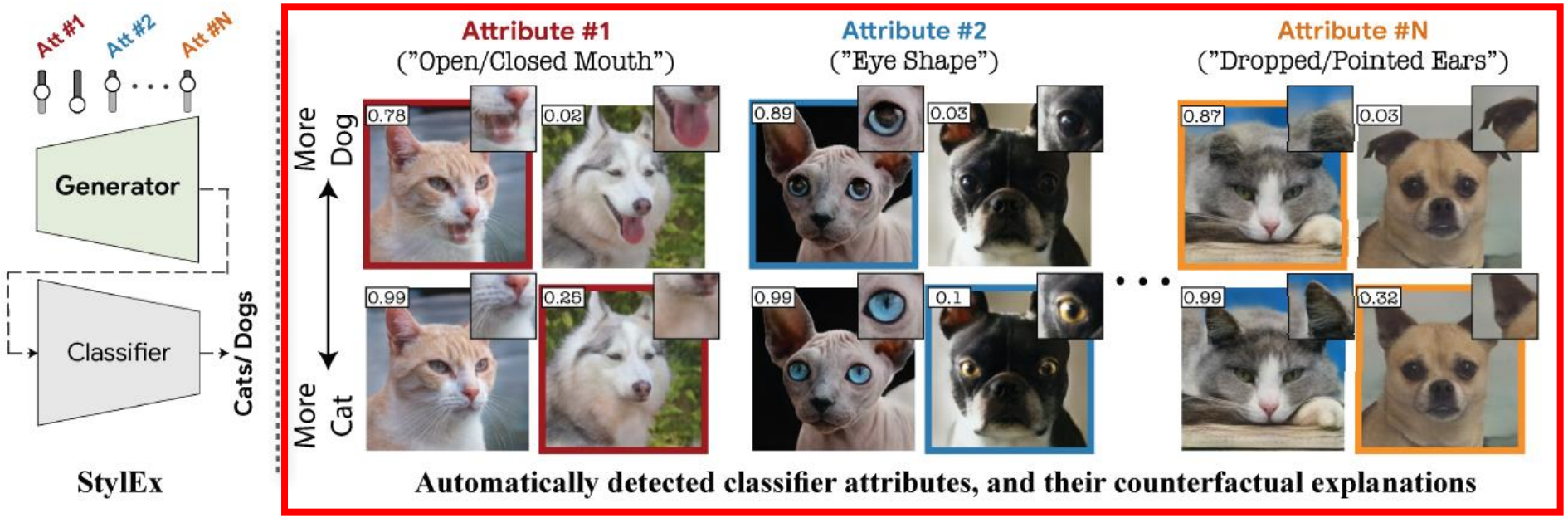


- Figure 1: Classifier-specific interpretable attributes emerge in the StylEx StyleSpace
- StylEx 는 prediction 에 영향을 주는 여러가지 attribute 를 찾고 시각화해서 classifier 의 결정을 설명한다.



- (Left) StylEx 는 classifier 를 설명하기 위해 StyleGAN 을 training 시킨다. (e.g., a “cat vs. dog” classifier)
- 따라서 classifier-specific attributes 를 포착하기 위해 GAN 의 StyleSpace 에서 latent attributes 를 유도한다.



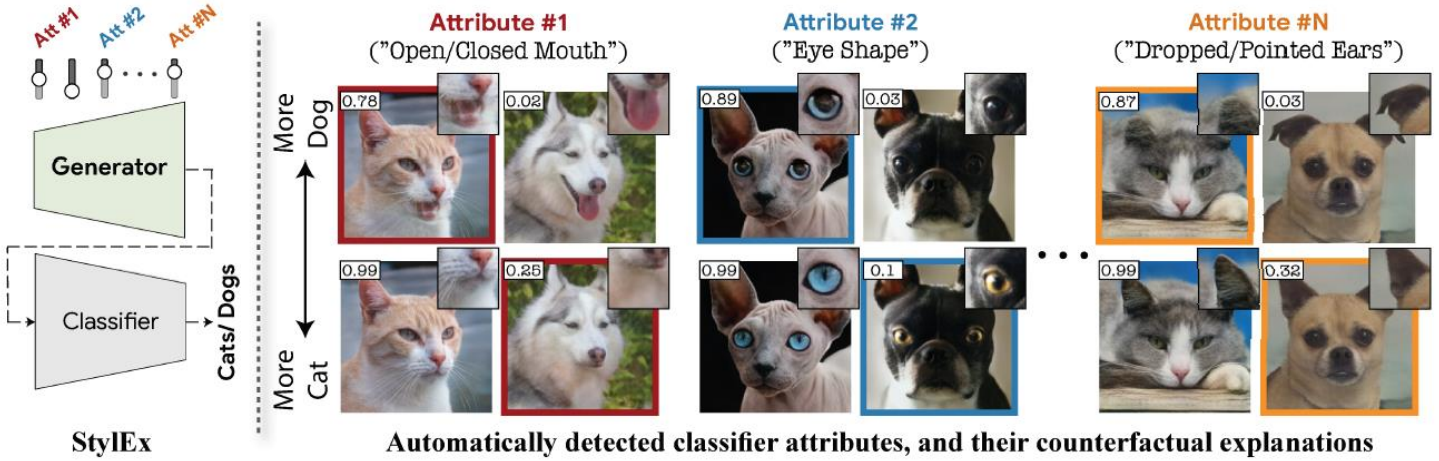


- (Right) 자동으로 StyleSpace 좌표에서 시각적으로 가장 두드러지고 classifier 의 결정을 가장 잘 설명하는 attributes 를 찾는다.
- 발견된 각 attribute 에 대해, StylEx 는 counterfactual example 을 생성해서 attribute 를 조정하는 것이 어떻게 classifier 결과의 확률에 영향을 미치는지 설명 한다.
- 생성된 counterfactual examples 는 색이 있는 상자로 나타냈고, 각 속성을 조정하는 것이 classifier 확률에 얼마나 영향을 미치는지 이미지의 top-left 에 나타냈다.
- 본 논문이 찾은 top attributes 는 실제로 cat vs. dogs 의 인식에 영향을 미치는 일관된 semantic 한 특징들 (e.g. open or closed mouth, eye shape, and pointed or dropped ears) 에 대응된다.

1. Introduction

- 다양한 형태의 설명 중에서, counterfactual explanations(조건법적 설명: 어떤 문장의 첫절이 사실과 정반대인 것을 서술할 경우의 표현법, 예를 들면 [만약 내가 알고 있었다면](if I had known) 따위.-네이버 어학사전)가 주목을 받고있다 [19, 11, 33]

- Counterfactual explanation : **"입력 x 가 $\tilde{x}$ 였다면 classifier 의 결과는 y 대신에 $\tilde{y}$ 였을 것이다"**
  - 예를 들어, 고양이와 개를 분류하는 classifier 가 있다고 해보자.
  - 고양이로 분류된 이미지를 위한 counterfactual explanation : **"만약 동공이 더 크게 만들어졌다면, classifier 가 '고양이'로 판단할 확률은 10% 감소할 것이다"**
  - Counterfactual explanation 의 장점은 입력에 어떤 부분이 분류에 중요한지, 대안적 결과를 얻기 위해 입력의 어떤 부분이 변화될 수 있는지 정확히 파악하면서 예시 별로 설명을 제공한다는 것이다.



Automatically detected classifier attributes, and their counterfactual explanations

- Counterfactual explanations 의 효과는 x 와 $\tilde{x}$ 의 차이가 얼마나 직관적인지에 크게 좌우된다.
- 유용한 counterfactual explanation 을 하기 위해 해석 가능한 features 또는 attributes 를 찾아야 한다.
  - cats vs dogs classifier 의 경우, 해석가능한 attributes 는 “pupil size” 또는 “open mouth”가 될 수 있다.
- 그것들을 시각화하기 위해 이미지에서 이러한 attributes 의 control 을 할 수 있어야 한다.

- 시각적인 attributes 를 찾고 시각화 하기 위해 생성 모델(StyleGAN2)을 사용했다.
- 그러나, 일반적인 StyleGAN2 training 이 classifier 를 포함하고 있지 않기 때문에 classifier 와 관련된 attributes 를 찾지 않는다.
- StylEx 가 이러한 문제를 해결하고 classifier explainability 를 만들게 하기 위해 다음의 방법을 제안한다.
    - (1) Classifier-specific StyleSpace 를 얻기 위해 StyleGAN training 절차에 classifier 를 통합시킨다.
    - (2) Classifier 예측에 영향을 미치는 간결한 attributes 집합을 위해 이 StyleSpace 를 mining 한다.

- Classifier 를 추가했을 때의 장점
  - GAN 의 훈련 과정에서 classifier 를 추가하면 이미지의 디테일에 집중할 수 있게 된다.
  - Generator 는 디테일을 인식하지 않고 이미지를 생성하기 때문에, generator 만으로는 이미지의 디테일을 잘 살리지 못한다.

- 본 논문은 다양한 도메인에서 StylEx 를 설명하고, 각 도메인의 분류에서 가장 두드러진 semantic attributes 를 추출하는 것을 보여준다.
- 본 논문의 contributions 은 다음과 같다.
    1. StyleGAN2 의 classifier-based training 을 위한 **StylEx** 를 제안하고, **classifier-specific attributes** 를 포착하기 위해 StyleSpace 를 유도한다.
    2. StyleSpace 좌표에서 classifier 와 관련된 attributes 를 발견하고, 이를 **counterfactual explanations** 에 사용하는 방법을 제안한다.
    3. StylEx 는 다양한 classifiers 와 real-world complex domains 를 설명하기 위해 사용할 수 있다. StylEx 가 **사용자가 이해할 수 있는 설명을 제공**한다는 것을 보여준다.

2. Related Work

- Visual explanation 방법들은 주로 **heatmaps** 에 의존한다.
- Heatmaps 은 decision 또는 classifier 의 hidden unit 의 activation 에 가장 두드러진 이미지의 regions 를 강조한다. Heatmaps 는 이미지에서 어떤 objects 가 classification 에 기여를 하는지 이해하는 데에 유용하다.
- 그러나, heatmaps 는 '크기'나 '색상'과 같은 spatially localized 하지 않는 attributes 를 잘 시각화 하거나 설명하지 못한다. 게다가 heatmaps 는 이미지의 어떤 영역이 classification 에 영향을 주기 위해 바뀌어야 하는지 보여 줄 수는 있지만, 그 영역이 어떻게 바뀌어야 하는지는 보여주지 못한다.

- 시각적인 counterfactual explanations 을 만들기 위해 생성 모델을 사용했고, 실제로 최근 연구에서 이 방법이 괜찮은 결과를 낸다는 것을 확인했다.
- 다음은 StylEx 와 비슷한 목적을 가지지만, 다른 방법을 사용한 연구들이다.

논문 번호	설명
[31, 7, 32, 1]	Generative counterfactual explanations 이 생성되지만, Fig 2와 같이 시각적인 변화는 모든 관련 속성을 한 번에 변경한다.
[29]	Classification 결과에 영향을 미치지 않는 특징들을 개입시킨다.
[10, 5]	원하는 attributes 에 대한 전체 또는 부분적인 supervision 이 가능한 속성을 사용하여 counterfactuals 를 생성한다.
[11]	생성 모델에 기반하지 않은 counterfactual explanation 방법을 제안한다.
[21]	StylEx 방법과 가장 비슷한 방법이다. 그러나 작은 이미지에서만 동작을 하고, attributes 에서 한 가지 이상의 요소가 변화되었을 때 잘 설명하지 못한다.
[38, 6]	Superpixels 에 기반한 attributes 나 classifier 의 중간 레이어에서의 activation 을 추출하고, 생성 모델은 사용하지 않는다. 그래서 counterfactuals 역할을 하는 이미지는 만들어내지 않고, 관련 있는 이미지 패치들 또는 superpixels 를 보여주면서 attributes 를 설명한다.

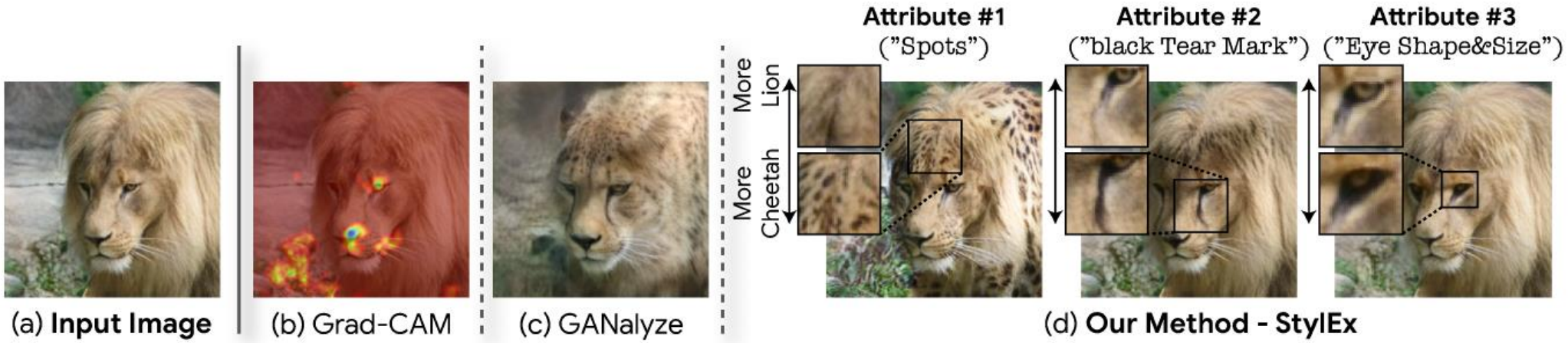


Figure 2: **Comparison to other visual explanation methods for a Lion vs. Cheetah classifier**. (b) Grad-CAM [25] and other heat-map based methods are limited in their ability to visualize attributes that are not spatially localized (e.g., eye size). (c) GANalyze [7] produces a possible counterfactual explanation, but its visualization changes all relevant attributes at once. (d) Our StylEx method provides meaningful interpertable multi-attribute explanation, by generating counterfactuals that change one attribute at a time.

- Lion vs. Cheetah classifier 에서 다른 visual explanation methods 와의 비교
  - (b) Grad-CAM [25] 과 다른 heat-map 기반의 방법들은 공간적으로 localized 되지 않은 attributes (e.g., eye size)를 시각화 하는 능력에 한계가 있다.
  - (c) GANalyze [7] 은 가능한 counterfactual explanation 을 제공하지만, 그것의 시각화는 모든 관련 있는 attributes 를 한번에 바꾼다.
  - (d) StylEx 는 한번에 한 attribute 만 바꾸는 counterfactuals 를 생성해서 의미 있는 해석가능한 multi-attribute explanation 을 제공한다.

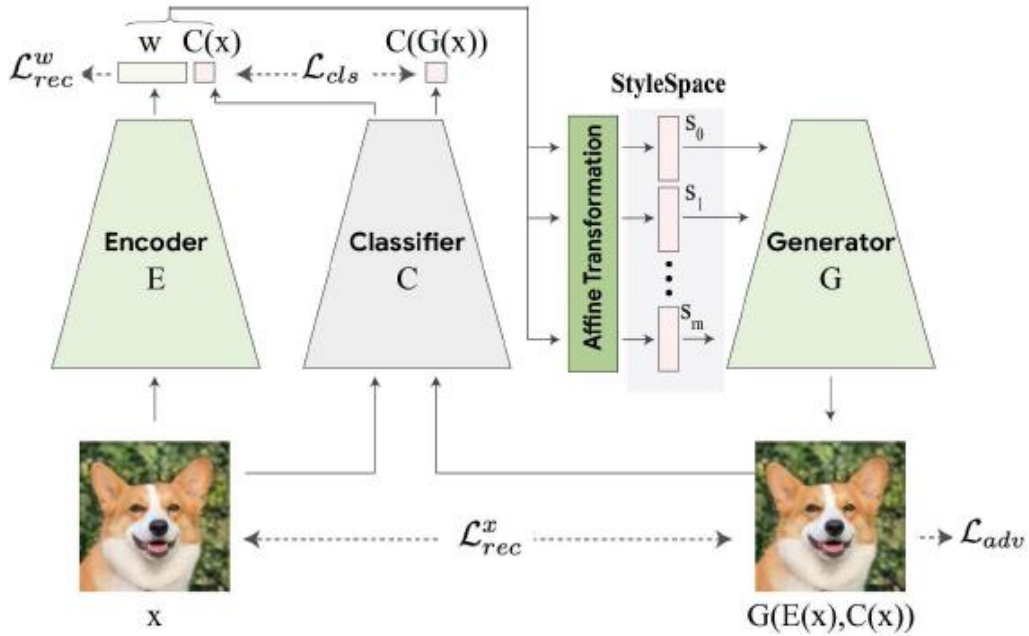
3. Method

3.1. StylEx Architecture

- 본 논문의 목표는 이미지에서 특정 attributes 를 바꿀 때 그 이미지가 어디로 분류되는지 설명하고, attributes 를 바꾼 것이 classifier 의 결과에 영향을 주는지 보여주는 것이다.
- 이것은 다음의 구성요소들을 통해 수행된다:
  - (a) Embedding w 를 output image 로 매핑하는 **conditional generative model**
  - (b) Input image 를 embedding w 로 매핑하는 **encoder**
  - (c) 이미지의 시각적 attributes 를 변경하기 위해 generation process 에 **"개입** " 하는 메커니즘

- 생성 모델로는 **StyleGAN2** 를 사용한다.
- [35] 에서 StyleGAN2 가 내재적으로 disentangled StyleSpace 를 가지는 경향이 있고, StyleSpace 는 각각의 attributes 를 추출할 수 있는 데에 사용될 수 있다는 것을 발견했다.
- 그러나, StyleGAN2 training 은 classifier 를 포함하지 않는다.
- 이 문제를 해결하고 StyleSpace 가 classifier 와 관련된 attributes 를 포함하도록 하기 위해, GAN 이 명시적으로 classifier 를 설명하도록 훈련했다.

- 마지막으로, 이미지로부터 latent vector w 를 예측하도록 훈련된 encoder 를 더한다.
- Encoder 를 필수적으로 달아야 하는 이유는 다음과 같다.
    1. Encoder 는 입력 이미지에서 classifier 결과를 설명하도록 해준다.
    2. Encoder 는 Classifier-Loss 를 사용하여 생성 모델이 classifier 와 관련된 attributes 를 포착하도록 한다.

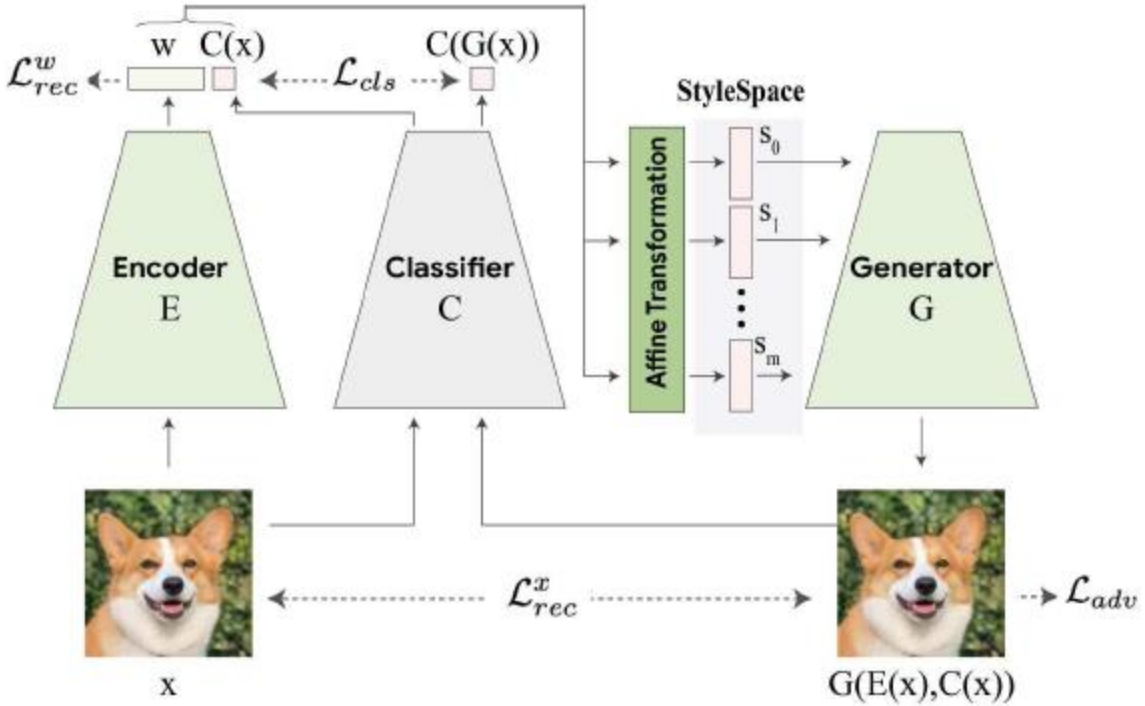


- Figure 3: StylEx architecture
  - 본 논문은 generator G, discriminator D, 그리고 encoder E 를 함께 training 한다.
  - Training 단계 동안에, 입력 이미지는 인코더를 통해서 latent vector w 로 변환된다 → w 는 이미지 x 의 classifier C 의 결과 $C(x)$ 를 concate 한다 → 결과는 affine transformation 을 통해 style vectors $s_0, \dots, s_n$ 으로 변형되고, 그것은 original image 와 비슷한 이미지를 생성해 내는 데에 사용된다.
  - Reconstruction loss 는 상응하는 encoder 의 output 사이에서 뿐만 아니라 generated image 와 original image 사이에서도 적용된다. GAN loss 는 생성된 이미지에 적용되고, KL loss 는 생성된 이미지에서 classifier C 의 결과와 input condition 사이에서 적용된다.

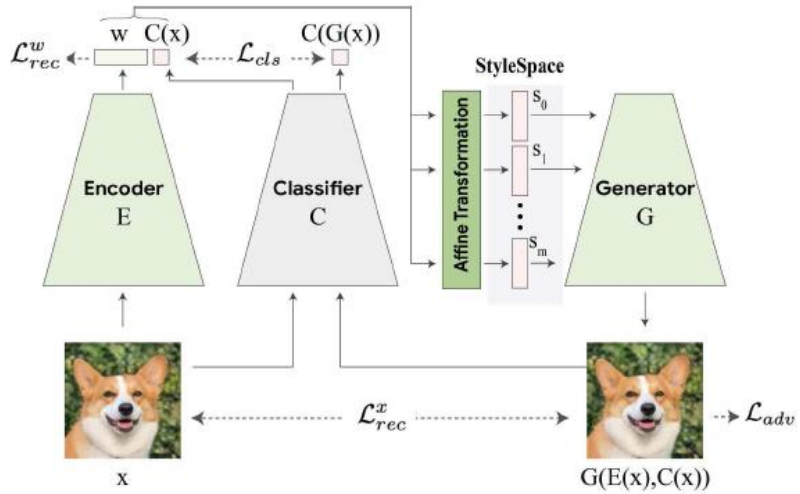
3.2. Training StylEx

- 기본적인 GAN training 방법은 generator G 와 adversarial discriminator D 를 동시에 훈련시키는 것이다.
- 본 논문은 추가적으로 reconstruction loss 를 사용하면서 generator G 와 함께 encoder E 를 훈련했다.
- 또한, training 절차에 classifier 를 도입했다.

- Conditional Training
  - 이미지를 classifier 에 넣어서 나온 결과를 generator 에 제공한다.
  - 이 조건을 추가하면 StyleSpace 가 classifier 의 결정에 영향을 미치는 attributes 를 더 많이 포함하도록 한다.



- Classifier Loss
  - 이미지셋에서 훈련된 GAN 은 반드시 특정 classifier 와 관련된 정보를 반드시 잡아내지 않는다.
  - 그래서, GAN training 동안에 Classifier-Loss 를 추가했다.
  - Classifier loss 는 생성된 이미지의 classifier 결과와 original 입력 이미지의 classifier 결과 사이에서의 KL-divergence 이다. 이 loss 는 generator 가 classification 을 위한 의미 있는 디테일을 무시하지 않도록 한다.

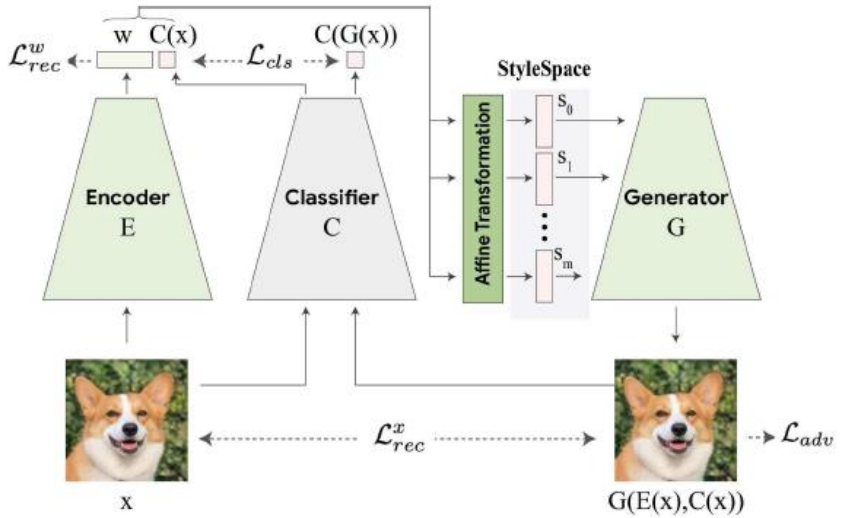


$Loss = L_{adv} + L_{reg} + L_{rec} + L_{cls}$	
L_{adv}	Logistic adversarial loss
L_{reg}	Path regularization [16]
L_{rec}	Encoding reconstruction loss • $L_{rec} = L_{rec}^{x} + L_{LPIPS} + L_{rec}^{w}$ • L_{rec}^{x} 와 L_{LPIPS} 는 input image x 와 conditioned reconstructed image $x' = G[E(x), C(x)]$ 사이에서 계산된다. • 더 구체적으로는, $L_{rec}^{x} = \|x' - x\|_1$ 이고 L_{LPIPS} 는 x 와 x' 사이의 LPIPS distance [39]이다. • L_{rec}^{w} 은 $L_{rec}^{w} = \|E(x') - E(x)\|_1$ 이다. [2]의 style reconstruction loss 에서 채택했다.
L_{cls}	Classifier loss • $L_{cls} = D_{KL}[C(x')\|\|C(x)]$

3.3 Extracting Classifier-Specific Attributes

- StyleSpace 에서 특정 좌표를 찾고 이를 변경하면, 생성된 이미지가 classifier 결과를 변경한다. 이것은 주어진 이미지에 대해서 counterfactual explanations 를 생성하게 한다.

- Algorithm 1 은 classifier-specific attributes 를 찾기 위한 AttFind 절차를 묘사한다.
  - Style vector 의 차원을 K 로, 이미지 x 에 대한 classifier 결과를 $C(x)$ 로 나타낸다.
  - AttFind 는 훈련된 모델과 Classifier 에 의해 예측된 레이블이 y (e.g., y="cat" or y="dog") 와는 다른 N 개의 이미지 집합을 입력으로 받는다.
  - 각 클래스 y 에 대해 M style 좌표(i.e., $S_y \subset [1, \dots, K]$ and $|S_y| = M$)를 바꾸는 것은 클래스 y 의 평균 확률을 증가시킨다.
  - 추가적으로 y 의 확률을 높이기 위해 이 좌표들이 어떤 방향에서 바뀌어야 할 필요가 있는지를 나타내는 "방향"들의 집합인 $D_y \in \{\pm 1\}^M$ 을 찾는다.



- AttFind 의 과정은 다음과 같다.
  1. 각 이터레이션에서 모든 K 차원의 style 좌표를 고려하고 y 의 확률에서 그들의 효과를 계산한다.
  2. 그런 다음 그것은 가장 영향을 크게 미치는 것의 좌표를 구하고, 이 좌표를 변경할 때 클래스 y 에 속할 확률에 큰 영향을 미치는 모든 이미지들을 제거한다. (즉, 이 좌표는 그 이미지들을 "설명" 하기에 충분하다; 다른 좌표로 진행 할 필요가 없다)

Algorithm 1: AttFind

Result: Set S_y of top M style coordinates & set D_y of their directions.

Data: Classifier C. A set X of images whose predicted label by C is not y. Generative model G. Threshold t.

Initialization : S_y, D_y = empty.

while $|S_y| < M$ *or* $|X| > 0$ **do**

  for x *in* X **do**

    for *style coordinate* $s \notin S_y$ **do**

      Set $\tilde{x}$ to be the image x after changing coordinate s in directions $d \in \{\pm 1\}$;

      Set $\Delta[x, s, d] = C_y(\tilde{x}) - C_y(x)$;

    end

  end

  Set $\bar{\Delta}[s, d]$ = Mean $(\Delta[x, s, d])$ over all $x \in X$;

  for *style coordinate* $s \notin S_y$ **do**

    if [4] $\bar{\Delta}[s, 1] > 0$ & $\bar{\Delta}[s, -1] > 0$ **then**

      set to Zero: $\bar{\Delta}[s, 1] = 0$ & $\bar{\Delta}[s, -1] = 0$;

  end

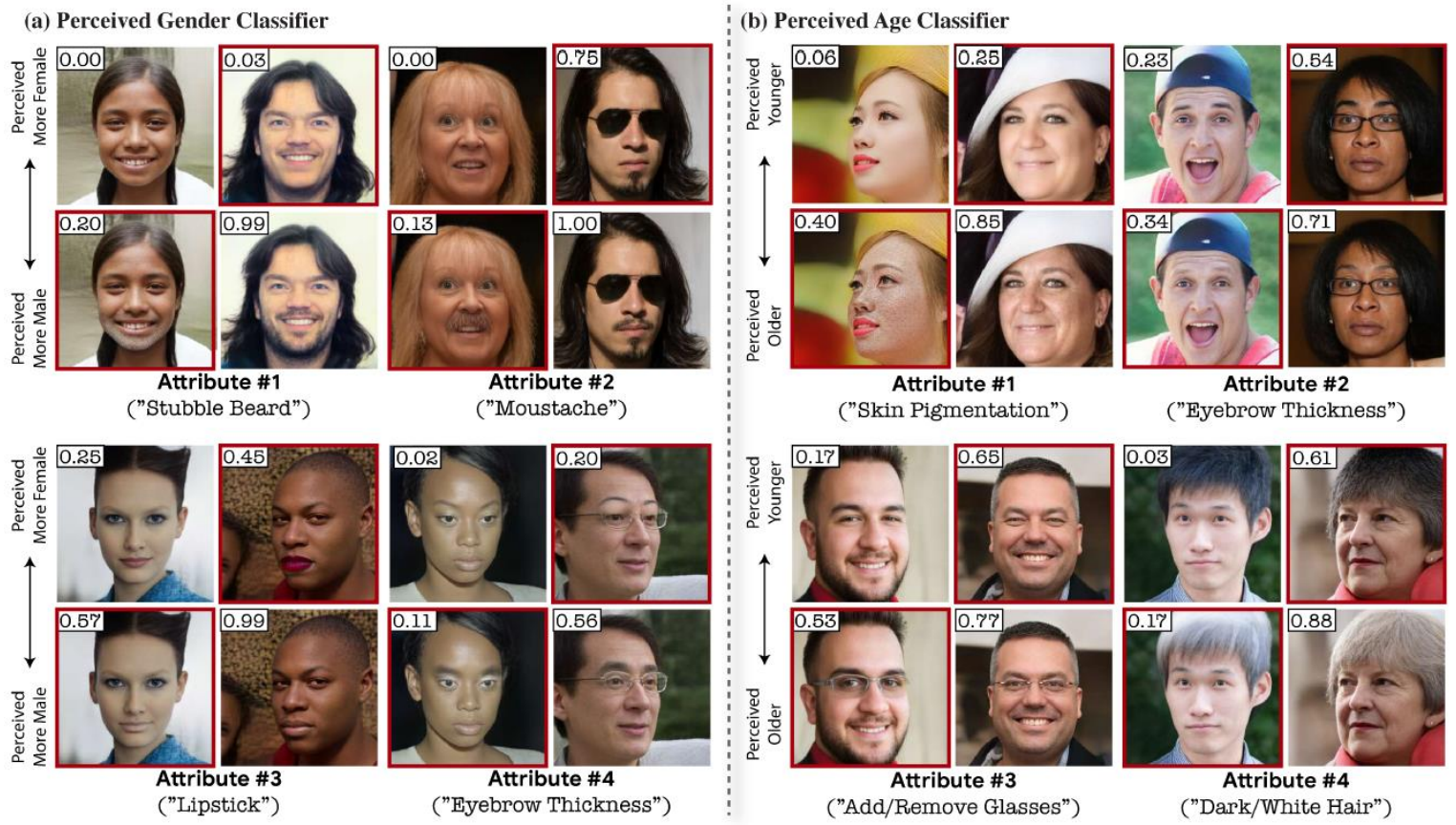
  Set s_{max}, d_{max} to be $\arg\max_{s,d} \bar{\Delta}[s, d]$;

  Add s_{max} to S_y, and d_{max} to D_y;

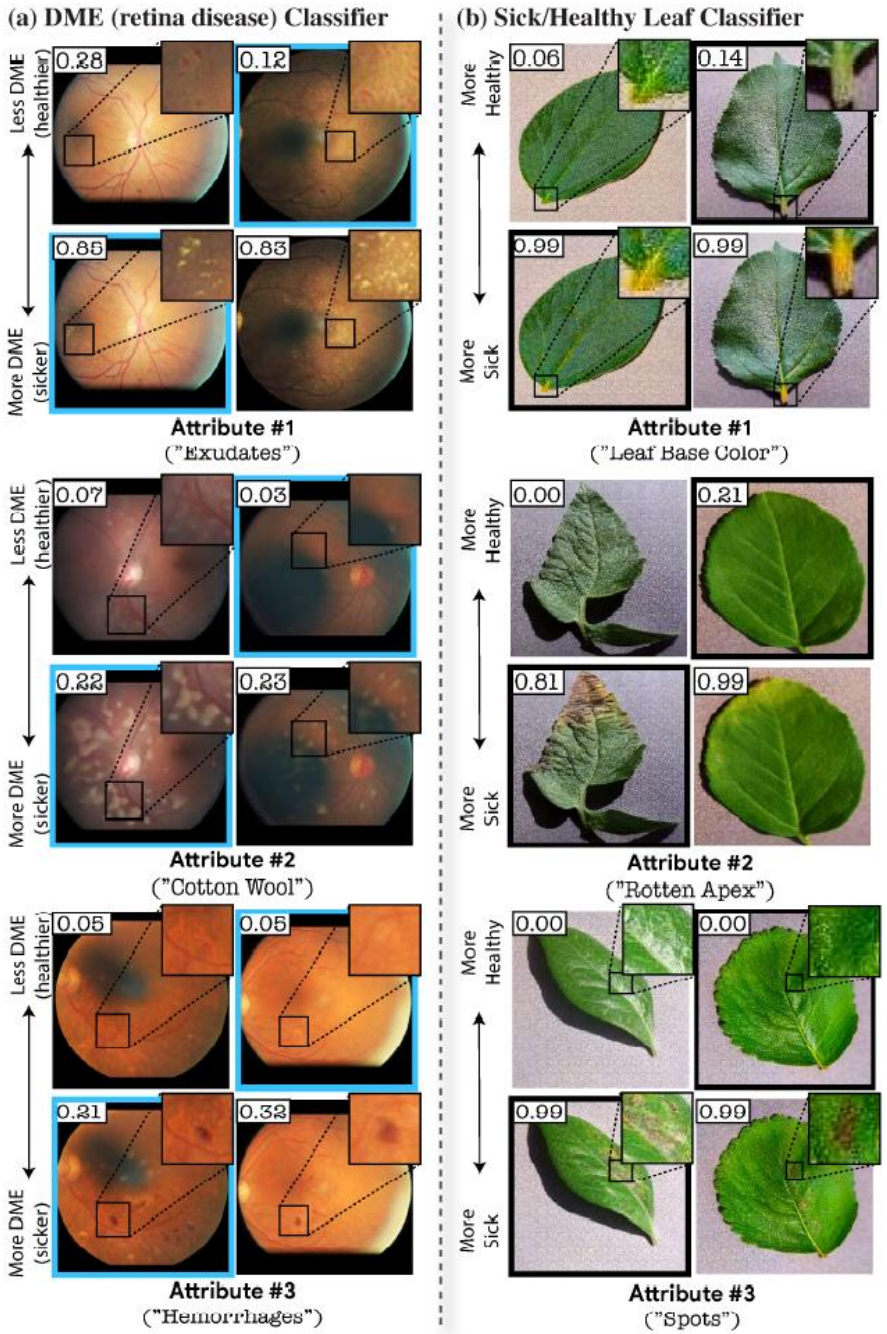
  Let $X_{explained}$ be all $x \in X$ s.t. $\Delta[x, s_{max}, d_{max}] > t$;

  Set $X = X \setminus X_{explained}$;

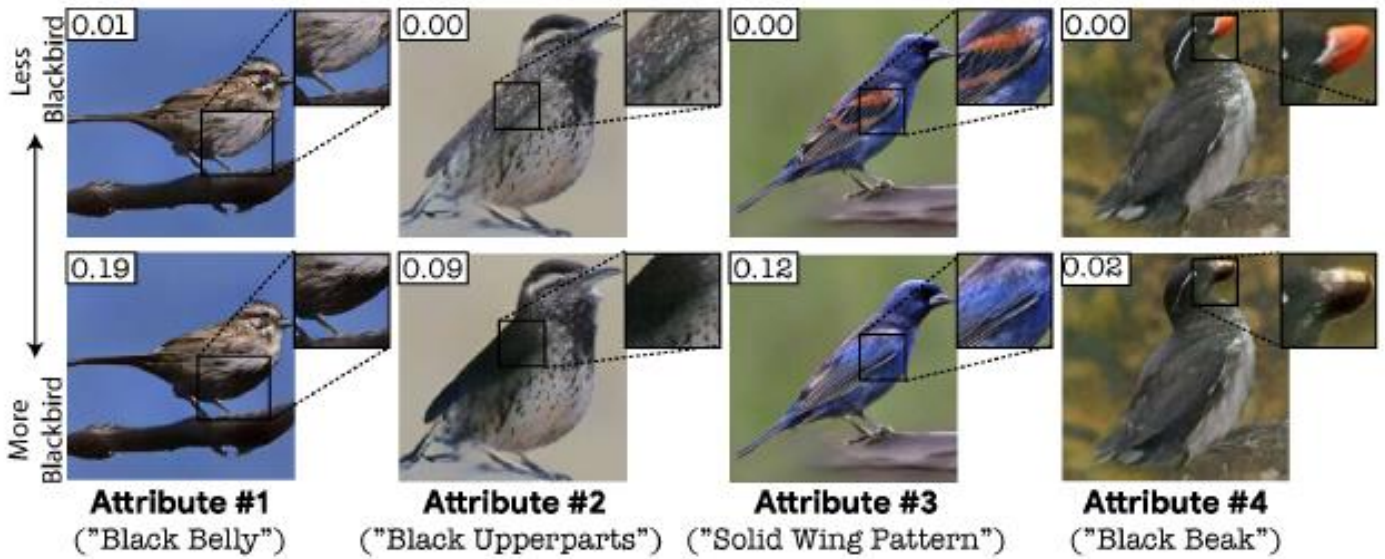
end



- Figure 4: **Top-4 automatically detected attributes for perceived-gender and perceived-age classifiers**
  - 상응하는 modifications 는 여러 이미지들 사이에서 시각적으로 일관되고, 다양한 semantic attributes 를 나타내며, 원하는 방향에 대한 classifier 의 예측에 영향을 미친다 (각 이미지의 왼쪽 상단에 나타냈다)
  - 생성된 counterfactual examples 는 프레임으로 표시했다.



- Figure 5: **Top-3 automatically detected attributes for (a) DME classifier of retina images and (b) classifier of Sick/Healthy leaf images**
  - 각 classifier 점수들은 왼쪽 상단에 표시된다.
  - 생성된 counterfactual examples 는 프레임으로 표시된다.
  - 두 개의 classifiers 모두에서 가장 많이 발견된 attributes 는 알려진 질병 지표와 잘 일치하는 것으로 나타났다.



- Figure 6: **Explaining multi-class classifiers: top-4 automatically detected attributes for the class brewer blackbird in a 200 way classifier trained on CUB-2011 [34]**
  - Brewer blackbird 클래스에 대한 classifier 점수는 왼쪽 상단에 나타냈다.
  - 가장 많이 발견된 attributes 는 CUB 분류체계 에서의 attributes 와 일치한다.

3.4. Generating Image-Specific Explanations

- StylEx 는 특정 이미지에서 classifier 의 결정을 설명하기 위한 자연스러운 메커니즘을 제공한다.
- Image specific attributes 의 집합을 찾기 위한 다양한 전략이 있다.
    1. 가장 간단한 전략은 StylEx attributes 를 반복하고, 이 이미지에 대한 classifier 결과에서 변화의 효과를 계산하고, 계산한 것 중에서 상위 k 개를 반환하는 것이다. 그런 다음 결과로 나오는 k 개의 수정된 이미지들을 시각화 할 수 있다. 이 전략을 **Independent** selection 라고 부른다.
    2. K 개의 요소를 가진 StylEx attributes 집합을 찾고 그 attributes 를 모두 함께 수정하면, classifier 의 변화는 극대화 된다. Greedy search 를 사용하여 attribute 를 찾고, classifier 가 classification 결과를 뒤집으면 중지한다. 이것을 **Subset** selection 이라고 부른다.

- Figures 7과 8은 image-specific explanations 의 예를 보여준다.
- Figure 7 : “Perceived Young” or “Perceived Old”
- Figure 8 : “Healthy” or “Sick”

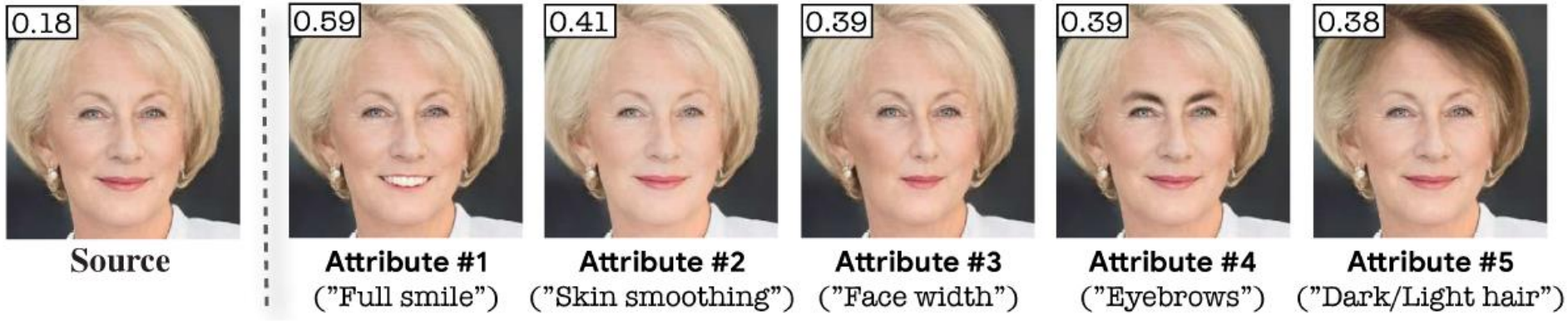


Figure 7: **Image-specific explanations:** Top-5 automatically detected attributes for explaining a perceived-age classifier for a specific image using the **Independent** selection strategy. Attributes are sorted by their effect on the classification of the specific image, resulting in different attributes from those presented in Fig. 4 which have the largest average effect over the entire dataset. The classifier probabilities of young are shown in the top-left corner.

- Figure 7: **Image-specific explanations**
  - **Independent** selection strategy 를 사용하여 특정 이미지에 대해서 perceived-age classifier 를 설명하기 위한 상위 5개의 자동으로 감지된 attributes 이다.
  - Attributes 는 특정 이미지의 classification 에서 그들의 효과에 의해 정렬되었고, Fig. 4 에 제시된 전체 데이터 셋에서 가장 큰 average effect 를 가지는 것들과는 다른 attributes 를 초래한다.
  - Young 의 classifier 확률은 top-left corner 에서 볼 수 있다.

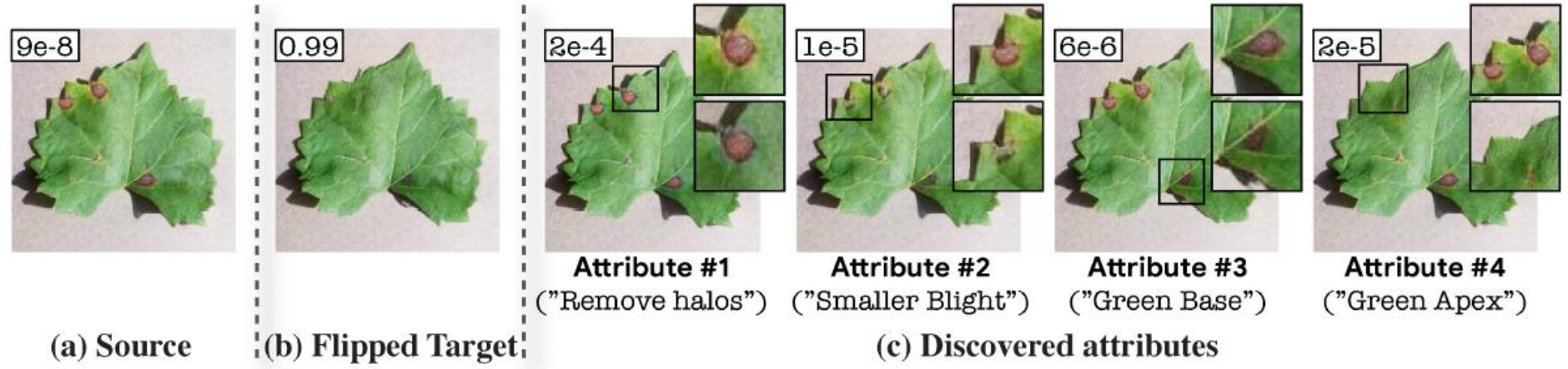


Figure 8: **Combining several attributes**. Attributes in **(c)** are selected by the **Subset** method to inflict the largest accumulated effect on the classification of the image in **(a)**. Interventions on individual attributes result in a small change of the classifier output, yet intervening on all of them results in image **(b)** where the classification is flipped. The classifier probabilities of healthy are presented in the top-left corner.

- Figure 8: **Combining several attributes**
  - (c) 에서의 Attributes 는 (a) 이미지의 classification 에 영향을 크게 주는 **Subset** method 에 의해 선택된다.
  - 각 attributes 결과에서 개입을 하면 classifier 출력에 작은 변화를 일으키지만, 모든 attributes 에서 개입을 하는 것은 classification 이 뒤집히는 이미지 (b) 를 초래한다.
  - Healthy 에서 classifier 확률은 top-left corner 에 나타냈다.

4. Evaluation and Results

- StylEx 방법을 다양한 도메인 집합의 다양한 classifiers 에서 테스트 했다. (Table 1)
- Classifiers 는 MobileNet [13] 구조에 기반한다.

Dataset	Classifier
AFHQ [2]	Cats / dogs
AFHQ	Wild cats species
FFHQ [15]	Perceived gender
FFHQ	Perceived age
Plant-Village [14]	Healthy / sick leaves
Retinal Fundus [18]	DME / non-DME
CUB-2011 [34]	Bird species

Table 1: List of datasets used in this paper.

4.1. Qualitative Evaluation

- **Visualizing StylEx Attributes**
  - StylEx 에서 발견된 attributes 는 실제로 일관성 있는 semantic 개념에 해당한다. 그리고 AttFind 에 의해 자동으로 발견된 top attributes 이다.
  - Figures 4, 5
  - StylEx 에서 추출된 각 top attributes 는 시각적으로 해석가능하다고 보여진다. 추가적으로, 각 attribute 를 수정하는 것은 classifier 의 결과에서 상당한 변화를 이끈다.

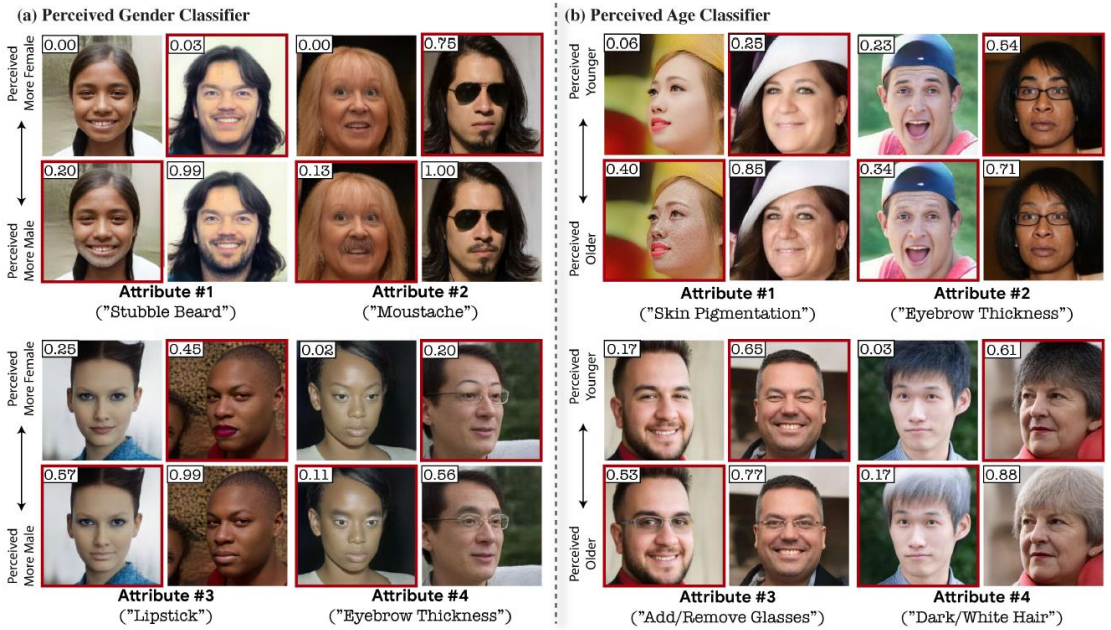


Figure 4

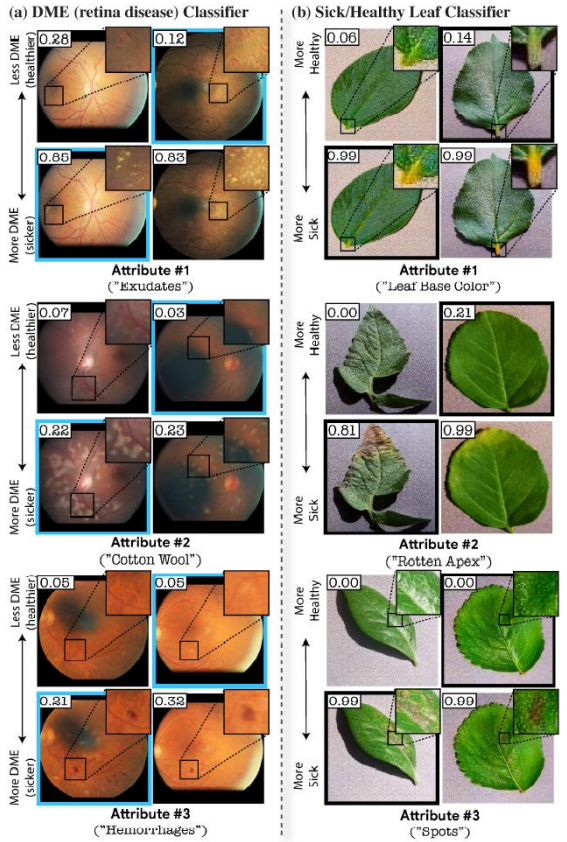
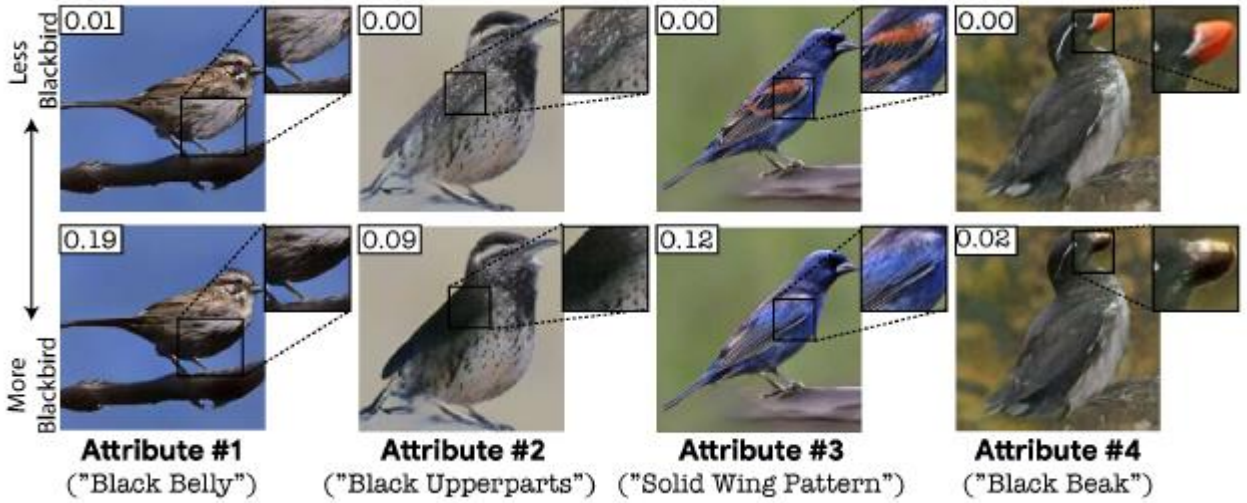


Figure 5

- **Explaining multi-class classifiers**
  - AttFind 는 multi-class 문제에도 적용할 수 있다.
  - Fig. 6 은 CUB-2011 (200 classes) [34] 에서 훈련된 classifier 에서 설명한다. 실제로, StylEx 가 CUB 분류 체계에 해당하는 attributes 를 감지하는 것을 볼 수 있다.



- **Providing Image-Specific Explanations**
  - StylEx attribute 를 특정 이미지들을 위한 counterfactual explanations 를 만드는 데에 사용할 수 있다.
  - 주어진 이미지에 대해서 다음과 같은 명령을 할 수 있다 : **“attribute #1과 #3 를 변경했다면, classifier 결과가 변경되었을 것이다.”**

- Fig. 7 은 “Perceived Age” classifier 에 대한 예시를 보여주고, **Independent** selection method 를 사용한다.
- Fig. 8 은 Plants domain 에 대해서 예시를 보여주고, 우리는 **Subset** selection method 를 사용한다.

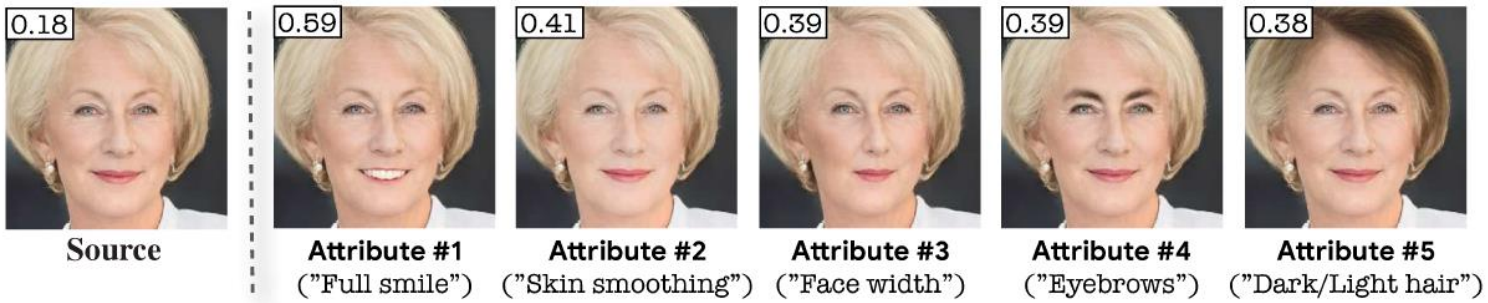


Figure 7: **Image-specific explanations:** Top-5 automatically detected attributes for explaining a perceived-age classifier for a specific image using the **Independent** selection strategy. Attributes are sorted by their effect on the classification of the specific image, resulting in different attributes from those presented in Fig. 4 which have the largest average effect over the entire dataset. The classifier probabilities of young are shown in the top-left corner.

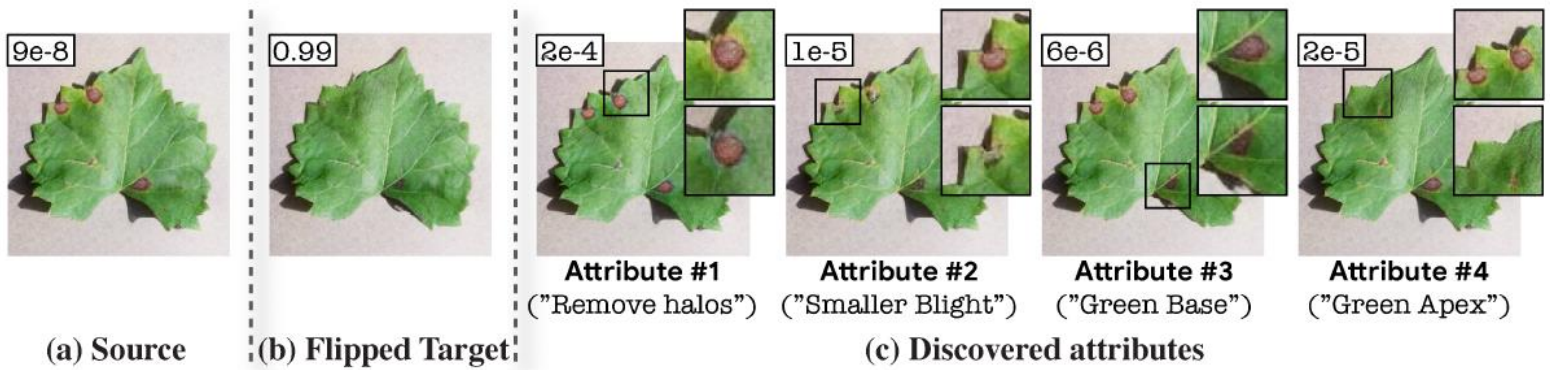


Figure 8: **Combining several attributes**. Attributes in **(c)** are selected by the **Subset** method to inflict the largest accumulated effect on the classification of the image in **(a)**. Interventions on individual attributes result in a small change of the classifier output, yet intervening on all of them results in image **(b)** where the classification is flipped. The classifier probabilities of healthy are presented in the top-left corner.

4.2. Quantitative Evaluation

- Multi-attribute counterfactual explanations 를 평가하는 방법은 명확하지 않지만, 중요한 세 가지 기준이 있다.
- **Visual Coherence**
    - Attributes 는 인간에게 식별가능해야 한다.
- **Distinctness**
    - 추출된 attributes 는 구별되어야만 한다.
- **Effect of Attributes on Classifier Output**
    - 이미지에서 attributes 의 값을 바꾸면 classifier 결과가 바뀌어야 한다.

4.2.1 Baselines and Model Variants

- 두 개의 구성요소의 비교를 위한 baseline 으로 [35] 를 사용했다.
  - [35] Wu, Z., Lischinski, D., & Shechtman, E. (2021). Stylespace analysis: Disentangled controls for stylegan image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 12863-12872)
- (1) classifier-specific training (CST) of the StyleGAN and (2) the AttFind method for finding classifier-related coordinates in StyleSpace
- 이 두 가지의 contributions 의 중요성을 테스트하기 위해, CST 없는 StylEx 와 [35] 에서 StyleSpace selection method 를 사용하여 비교한다.

4.2.2 A User Study for Coherence and Distinctness

- Coherence 와 distinctness 를 평가하기 위해 user study 를 수행했다.

	Wu *et al.*	**Ours**
Perceived Gender	0.783(±0.186)	0.96 (±0.047)
Perceived Age	0.85 (±0.095)	0.983 (±0.037)
Plants	0.91(±0.081)	0.916 (±0.068)
Cats/Dogs	0.65 (±0.18)	0.933 (±0.05)

Table 2: **User study results.** Fraction of correct answers on identification of the top-6 extracted attributes.

4.2.3 ‘Sufficiency’: Effect of Attributes on Classifier Output

- Classifier 에서 attributes 의 효과를 테스트하기 위해, 작은 attributes 집합에서 간섭이 classifier 의 결정을 뒤집을 수 있는지 확인하고 싶다.
- k = 10 인 top k attributes 에서 수정을 한다.

- Table 3 은 1000개의 랜덤으로 선택된 이미지들에서 측정한 것을 나타낸다. StylEx 가 대부분의 도메인에서 높은 explanation percentages 를 보여주는것을 볼 수 있다.
- Table 3 은 또한 conditional training 과 classifier loss 가 없는 StylEx 에서의 결과를 보여준다.
  - w/o CST 일 때, retina 와 plants 에서 극적인 효과를 보인다.

	Wu *et al.*	**Ours w/o CST**	**Ours**
Perceived Gender	14.3%	82.7%	83.2%
Perceived Age	16.9%	93.0%	93.9%
Cats/Dogs	1.0%	15.7%	25.0%
Wild Cats	11.8%	18.9%	66.7%
Plants	14.6%	58.2%	91.2%
Retina	0.0%	0.0%	100%

Table 3: **Effect of Top-10 Attributes on the Classifier**. The fraction of images for which the classification flipped when modifying top-k attributes up to $k = 10$ (see Sec. 4.2.3). Attributes discovered by StylEx affect classification results for a much larger percentage of images than the baseline methods. On the face domains, *AttFind* finds sufficient attributes even on standard StyleGAN2, while in other domains, classifier-specific training is required. On the Cats/Dogs classifier, due to the large visual differences between the two classes, top-10 attributes are not enough. 40 attributes are required to flip the classifier in 94% of the images.

Thank you for listening