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

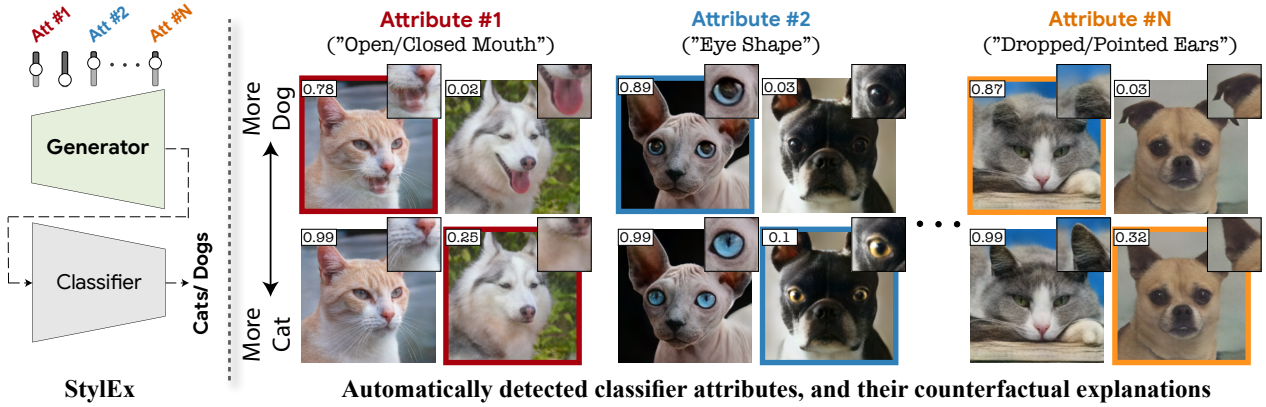

Figure 1. **Classifier-specific interpretable attributes emerge in the StylEx StyleSpace.** Our system, StylEx, explains the decisions of a classifier by discovering and visualizing multiple attributes that affect its prediction. (Left) StylEx achieves this by training a StyleGAN specifically to explain the classifier (e.g., a "cat vs. dog" classifier), thus driving latent attributes in the GAN's StyleSpace to capture *classifier-specific* attributes. (Right) We automatically discover top visual attributes in the StyleSpace coordinates, which best explain the classifier's decision. For each discovered attribute, StylEx can then provide an explanation by generating a *counterfactual example*, i.e., visualizing how manipulation of this attribute (style coordinate) affects the classifier output probability. The generated counterfactual examples are marked in the figure by the color of their attribute. The degree to which manipulating this attribute affects the classifier probability is shown in the top-left of each image. The top attributes found by our method indeed correspond to coherent semantic properties that affect perception of cats vs. dogs (e.g. open or closed mouth, eye shape, and pointed or dropped ears).

Abstract

Image classification models can depend on multiple different semantic attributes of the image. An explanation of the decision of the classifier needs to both discover and visualize these properties. Here we present StylEx, a method for doing this, by training a generative model to specifically explain multiple attributes that underlie classifier decisions. A natural source for such attributes is the StyleSpace of StyleGAN, which is known to generate semantically meaningful dimensions in the image. However, because standard GAN training is not dependent on the classifier, it may not represent these attributes which are important for the classifier decision, and the dimensions of StyleSpace may represent irrelevant attributes. To overcome this, we propose a training procedure for a StyleGAN, which incorporates the classifier model, in order to learn a classifier-specific StyleSpace. Explanatory attributes are then selected from this space. These can be used to visualize the effect of changing multiple attributes per image, thus providing image-specific explanations. We apply StylEx to multiple domains, including animals, leaves, faces and retinal images. For these, we show how an image can be modified in different ways to change its classifier output. Our results show that the method finds attributes that align well with semantic ones, generate meaningful image-specific explanations, and are human-interpretable as measured in user-studies. [1]

1. Introduction

Deep net classifiers are often described as "black boxes" whose decisions are opaque and hard for humans to understand. This significantly holds back their usage, especially in areas of high societal impact, such as medical imaging and autonomous driving, where human oversight is critical.

[*] indicates equal contributions; Work performed by authors while working at Google.

[1] Project website: https://explaining-in-style.github.io/

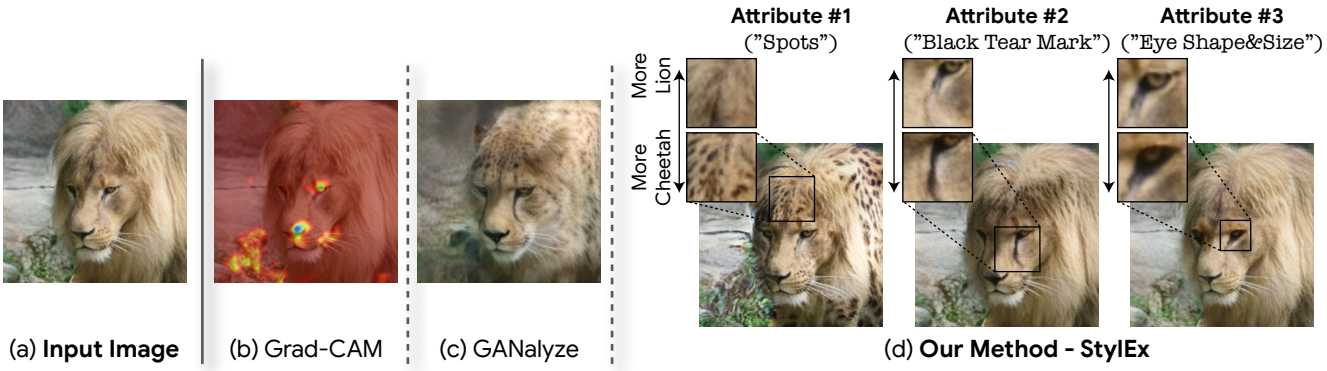

Figure 2. **Comparison to other visual explanation methods for a Lion vs. Cheetah classifier**. (b) Grad-CAM [24] and other heat-map based methods are limited in their ability to visualize attributes that are not spatially localized (e.g., eye size). (c) GANalyze [7] produces a possible counterfactual 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.

Explaining the decision of classifiers can reveal model biases [16], provide support to downstream human decision makers, and even aid scientific discovery [19].

Among the different forms of explanations, counterfactual explanations are gaining increasing attention [18, 11, 32]. A *counterfactual explanation* is a statement of the form "Had the input $\mathbf{x}$ been $\tilde{\mathbf{x}}$ then the classifier output would have been $\tilde{y}$ instead of y", where the difference between $\mathbf{x}$ and $\tilde{\mathbf{x}}$ is easy to explain. For instance, consider a classifier trained to distinguish between cat and dog images. A counterfactual explanation for an image classified as a cat can be "If the pupils were made larger, then the output of the classifier for probability of cat would decrease by 10%". A key advantage of this approach is that it provides per-example explanations, pinpointing which parts of the input are salient towards the classification and also how they can be changed in order to obtain an alternative outcome.

The effectiveness of counterfactual explanations strongly depends on how intuitive is the difference between $\mathbf{x}$ and $\tilde{\mathbf{x}}$ to human observers. For instance, if $\tilde{\mathbf{x}}$ is an arbitrary image classified as a dog, then it is not useful as a counterfactual since it usually changes all features of $\mathbf{x}$ and therefore does not isolate the critical features the classifier depends on. Adversarial examples [9], on the other hand, can change a small amount of features, yet they are also not useful in explaining classifications to humans, as the changes are not interpretable. In non vision applications, a common approach is to change $\mathbf{x}$ in only a few features, where the features themselves are interpretable. For example, when refusing a loan, an explanation could be: "had your income (i.e., feature x_3) been higher, you would have been granted the loan.". However, in computer vision the raw features are the pixels, and changing a few pixels is unlikely to correspond to a meaningful semantic feature.

Therefore to form a useful counterfactual explanation we must discover interpretable features, or *attributes*; in the case of the cats vs. dogs classifier, these might be "pupil size" or "open mouth". To visualize them we further need to enable control of these attributes in the image, a task most suited to generative models. This is an inherently different task from visualizing a smooth transformation between one class and the other, as done for instance in [7, 30]. Such transformations change all attributes at once, making it difficult to isolate fine grained attributes. Both defining and visualizing interpretable attributes is challenging since in many domains, e.g. medical imaging, we may not know the salient visual attributes, or do not have labeled examples of them.

A natural candidate for finding visual attributes and visualizing them is generative models, such as StyleGAN2 [15], where it was shown that it is possible to find disentangled latent variables that control semantic attributes of the images they generate. This approach has been used to create powerful interfaces for image editing and data visualization [3, 25, 7, 27, 26]. Our method builds upon a recent observation by [33] that StyleGAN2 tends to contain a *disentangled* latent space (i.e., the "StyleSpace") which can be used to extract individual attributes. However, as we show here, this approach will not necessarily discover classifier-related attributes since standard StyleGAN2 training does not involve the classifier. Instead, we propose our StylEx framework to overcome this difficulty and promote classifier explainability by: (i) incorporating the classifier into the StyleGAN training procedure to obtain a classifier-specific StyleSpace, and then (ii) mining this StyleSpace for a concise set of attributes that affect the classifier prediction.

Adding the classifier into the training process of the GAN turns out to be crucial in domains where the classification depends on fine details, e.g., in retinal fundus images. A generator unaware of the importance of these subtle details, may fail to generate them. We demonstrate StylEx on a variety of domains, and show it extracts semantic attributes

that are salient for classification in each domain.

Our contributions are as follows:

- We propose the StylEx model for classifier-based training of a StyleGAN2, thus driving its StyleSpace to capture *classifier-specific* attributes.
- A method to discover classifier-related attributes in StyleSpace coordinates, and use these for counterfactual explanations.
- StylEx is applicable for explaining a large variety of classifiers and real-world complex domains. We show it provides explanations that are understood by human users.

2. Related Work

The most widespread visual explanation methods are based on heatmaps. These maps highlight regions of the image that are salient towards the decision, or towards the activation of a hidden unit of the classifier (e.g., [24, 23, 34]). Heatmaps are useful to understand things like which objects in an image contributed to a classification. However they are limited in their ability to visualize attributes that are not simply spatially localized, like size, color, etc. In addition they can show which area of the image may be changed in order to affect classification, but not how it should be changed.

Counterfactual explanations address these limitations by providing alternative inputs, where a small set of attributes is changed and the different classification outcomes are observed. Generative models are natural candidates to produce visual counterfactual explanations, and indeed recent works have shown progress towards this goal. In [30, 7, 31, 1] generative counterfactual explanations are produced, yet their visualization changes all relevant attributes at once, as shown in Fig. 2. Another related approach offered in [28] is to use deep representations from a classifier to manipulate generated images at different granularities. Yet these may involve properties that do not affect the classification outcome and also combine several attributes. Hence these methods do not allow the analysis of a classifier in terms of atomic attributes and their effect on classifications. Other explanation methods generate counterfactuals using attributes, where full or partial supervision for the desired attributes is available [10, 5]. [11] propose a counterfactual explanation method that is not based on a generative model, and instead replaces a small number of patches from one image into another. Their method does decompose the counterfactual generation into several patch replacements, though the counterfactuals are often not realistic images and the method does not explicitly define a set of controllable attributes. The closest method in spirit to the explanations provided by our StylEx approach is [20], though their method only works on small images, and

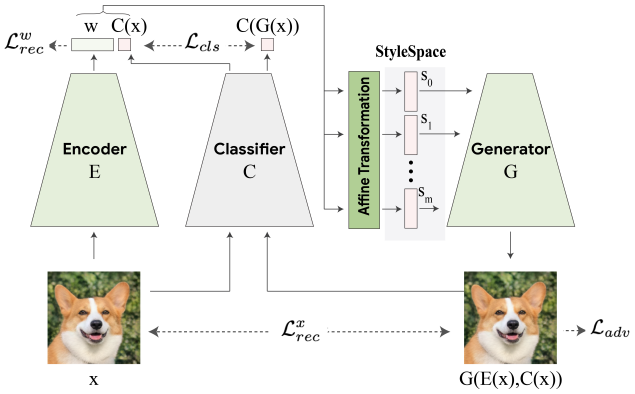

Figure 3. **StylEx architecture**. We jointly train the generator G, discriminator D, and encoder E. During the training phase, an input image is transformed via the encoder into a latent vector w. w is then concatenated to the output $C(x)$ of the classifier C on the image x. The result is transformed via affine transformations to the style vectors $s_0, ..., s_n$, which are then used to generate an image close to the original image. A reconstruction loss is applied between the generated image and the original image, as well as between the corresponding encoder outputs. A GAN loss is applied on the generated image, and a KL loss is applied between the output of the classifier C on the generated image and the input condition.

they do not demonstrate explanations that consist of more than a single change of attributes (nor do they claim to find multiple semantic attributes that affect the classification).

Explanations based on multiple attributes that are extracted in an unsupervised manner are given in [35, 6]. They extract attributes based on superpixels, or activations in the mid-layers of the classifier, and do not use a generative model. Hence they do not create images that serve as counterfactuals, and their attributes are demonstrated by showing relevant image patches or superpixels. In terms of visualization, representative patches are limited in their ability to visualize attributes that are not spatially localized. Furthermore, counterfactual images let us observe how the classification changes under interventions on a combination of attributes.

3. Method

We next describe our approach for discovering classifier-related attributes and modifying these attributes in real images. Our approach consists of two key steps. First, we train a StyleGAN model in a way that incorporates the classifier, thus encouraging the StyleSpace of the StyleGAN to accommodate classifier-specific attributes (Sec. 3.2). Then, we search the StyleSpace of the trained GAN to automatically discover coordinates that correspond to classifier-specific attributes (Sec. 3.3). Finally, we show how to use these attributes in order to visually explain a classifier's decision for a given input image (Sec. 3.4).

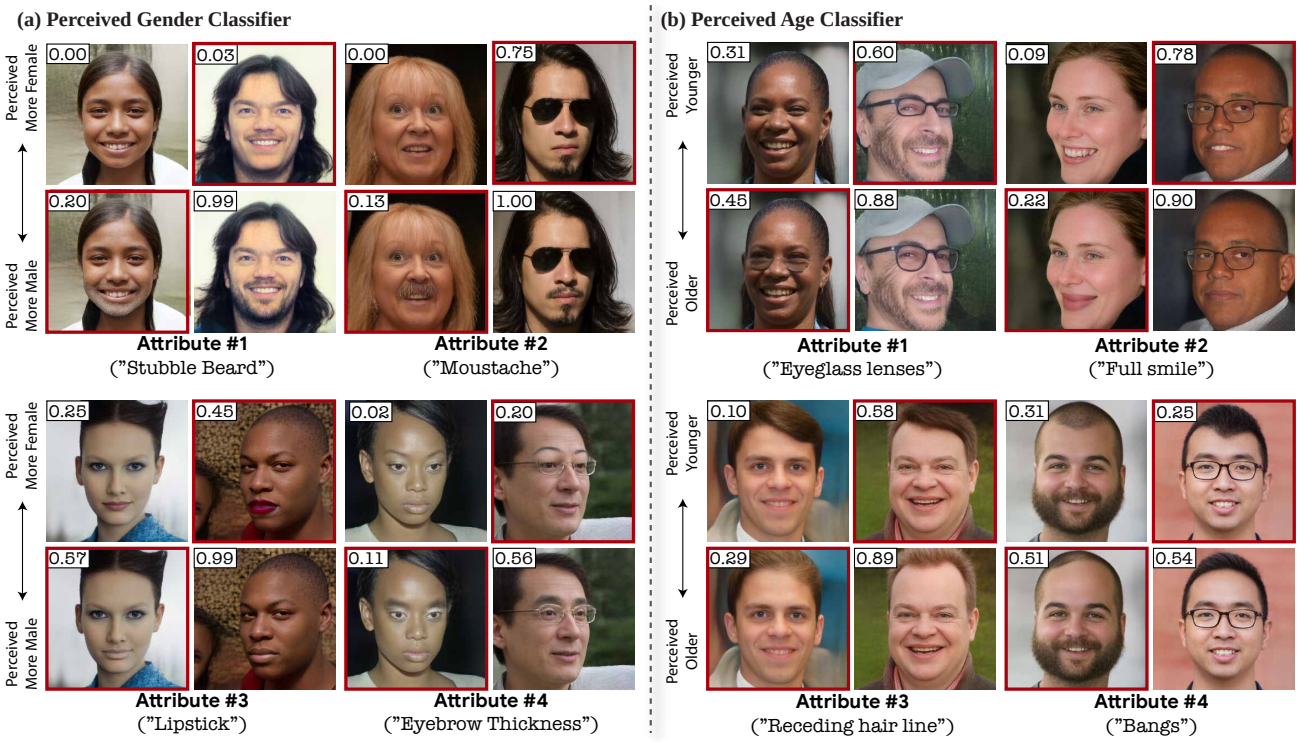

Figure 4. **Top-4 automatically detected attributes for perceived-gender and perceived-age classifiers**. The corresponding modifications are visually coherent between different images, represent diverse semantic attributes, and effect the classifiers predictions (presented in the top-left corner of each image) towards the wanted directions. The generated counterfactual examples are marked by a frame. Please refer to the Supplemental for more attributes and animated-GIFs to view these counterfactual changes (explanations) dynamically.

3.1. StylEx Architecture

Recall that our goal is to explain the classification of a given image by changing certain attributes in the image, and to show they affect the classifier output. We achieve this by combining the following components: a) A conditional generative model that maps an embedding w into an output image. b) An encoder that maps any input image into an embedding w, so that the generator can modify attributes in real images. c) A mechanism for "intervening" with the generation process to change visual attributes in the image.

For the generative model we use StyleGAN2 [15]. This architecture was shown to generate realistic images in multiple domains. But more important to our goal is the observation recently made by [33] that StyleGAN2 tends to inherently contain a disentangled StyleSpace space, which can be used to extract individual attributes. Thus, we argue that modifying coordinates of StyleSpace is a natural approach to our problem of modifying classifier-related attributes. In [33] the authors extracted coordinates of StyleSpace that corresponded to *known attributes* in a pre-trained StyleGAN2. In general, however, StyleGAN2 is not trained to discover classifier-related attributes of an arbitrary classifier, since standard StyleGAN training does not involve that classifier in any way (as shown in Sec. 4.2.3).

To overcome the above problem and allow for the StyleSpace to contain classifier-related attributes, we train our GAN to explicitly explain the classifier, by conditioning the model on classifier output and using a classifier loss (see below). As we shall see this will result in Classifier-specific disentangled attributes that emerge in our StyleSpace.

Finally, we add an encoder which is trained to predict the latent vector w from the image (see Fig. 3). This is necessary for two reasons. First, it allows us to explain the classifier output on any given input image (and not only on GAN-generated images). Second, it allows us to ensure that the generative model captures classifier-related attributes by using the Classifier-Loss (see below).

3.2. Training StylEx

Fig. 3 shows the components of StylEx and its training procedure. The training method of our generative model is based on the standard GAN training procedure, but adds to it several modifications required for guiding the StyleSpace to contain classifier-related attributes. The basic GAN training recipe is to train the generator G and an adversarial discriminator D simultaneously [8]. Additionally, we jointly train an encoder E with the generator G, using a reconstruction loss (i.e., the Encoder and Generator function together as an autoencoder). Finally, we add two components that introduce the classifier into the training procedure.

Conditional Training: We provide the generator with the

intended value of the classifier output on the generated image. Adding this condition helps the StyleSpace to contain more attributes that effect the classifier's decision (as the StyleSpace coordinates become an affine transformation of the classifier output).

Classifier Loss: A GAN trained on a set of images will not necessarily capture visual structures related to a particular classifier. For example, a GAN trained on retinal images will not necessarily visualize pathologies corresponding to a particular disease. In this case, it will be impossible to visually explain a classifier for this pathology using this GAN. To overcome this difficulty, we add a Classifier-Loss on the images generated by the GAN, during the GAN training. This loss is the KL-divergence between the classifier output on the generated image, and the classifier output on the original input image. This loss ensures that the generator does not ignore important details which are meaningful for the classification, or collapse into only one of the labels.

The overall StylEx training loss is the sum of the losses:

$$Loss = \mathcal{L}_{adv} + \mathcal{L}_{reg} + \mathcal{L}_{rec} + \mathcal{L}_{cls} \ , \tag{1}$$

where $\mathcal{L}_{adv}$ is the logistic adversarial loss [8], and $\mathcal{L}_{reg}$ is the path regularization described in [15]. The encoding reconstruction loss, $\mathcal{L}_{rec}$, is given by $\mathcal{L}^x_{rec} + \mathcal{L}_{LPIPS} + \mathcal{L}^w_{rec}$ where the first two terms are calculated between the input image X and the conditioned reconstructed image $X' = G[E(X), C(X)]$. More specifically, $\mathcal{L}^x_{rec} = \|X' - X\|_1$ and $\mathcal{L}_{LPIPS}$ is the LPIPS distance between X and X' as described in [36]. The third term, $\mathcal{L}^w_{rec}$, is adapted from the style reconstruction loss in [2]: $\mathcal{L}^w_{rec} = \|E(X') - E(X)\|_1$. Finally, the Classifier-Loss is $\mathcal{L}_{cls} = D_{KL}\left[C(X')||C(X)\right]$. In Sec. 4 we provide ablation results on these losses.

3.3. Extracting Classifier-Specific Attributes

Thus far we trained a generative model that is constrained to capture classifier-related information. We next turn to finding coordinates in the StyleSpace of the model, which encode classifier-specific attributes. Namely, we seek specific coordinates in the StyleSpace such that changing them will change the generated image in a way that alters the classifier output in a non-negligible way. This will enable generating counterfactual explanations for a given image.

Algorithm 1 describes the *AttFind* procedure for discovering classifier-specific attributes. Denote by K the dimension of the style vector (across all layers), and by $C(I)$ the vector of classifier logits (pre-softmax probabilities) for image I. *AttFind* takes as input the trained model and a set of N images. For each class y (e.g., y="cat" or y="dog") *AttFind* then finds a set S_y of M style coordinates (i.e., $S_y \subset [1, \ldots, K]$ and $|S_y| = M$), such that changing these coordinates decreases the average probability of the class y on these images.[2] Additionally it finds a set of "directions" $D_y \in \{\pm 1\}^M$ that indicates in which direction these coordinates need to be changed to decrease the probability of y.

AttFind proceeds as follows: At each iteration it considers all K style coordinates and calculates their effect on the probability of y.[3] It then selects the coordinate with largest effect, and removes all images where changing this coordinate had a large effect on their probability to belong to class y (i.e., this coordinate suffices to "explain" those images; no need to proceed to other coordinates). This is repeated until no images are left, or until M attributes are found. The process is summarized in Algorithm 1. Examples of these automatically detected attributes, for a variety of different classifiers, can be found in Figures 4 and 5.

Algorithm 1: AttFind

Result: Top M style coordinates S_y & directions D_y
Data: Classifier C. A set of N images I_y predicted as class y by C . Generative model G. Threshold t.
Initialization : S_y, D_y = empty.
while $|S_y| < M$ *or* $|I_y| > 0$ **do**
  for I *in* I_y **do**
    for *style coordinate* $s \notin S_y$ **do**
      Set $\tilde{I}$ to be the image I after changing coordinate s in direction $d \in \{\pm 1\}$;
      Set $\Delta[I, s, d] = C_y(I) - C_y(\tilde{I})$;
    end
  end
  Set $\bar{\Delta}[s, d]$ = Mean $(\Delta[I, s, d])$ over all $I \in I_y$;
  for *style coordinate* $s \notin S_y$ **do**
    if $\bar{\Delta}[s, 1] > 0$ *and* $\bar{\Delta}[s, -1] > 0$ **then**
      Set $\bar{\Delta}[s, \pm 1] = 0$;
  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 $I_{explained}$ be all $I \in I_y$ s.t. $\Delta[I, s_{max}, d_{max}] > t$;
  Set $I_y = I_y \setminus I_{explained}$;
end

3.4. Generating Image-Specific Explanations

StylEx provides a natural mechanism for explaining the decision of a classifier on a specific image: simply find StylEx attributes that affect the classifier's decision on this image, and visualize the effect of changing those.

There are various strategies for finding a set of image-specific attributes. The simplest is to iterate over StylEx attributes, calculate the effect of changing each on the classifier output for this image, and return the top-k of these. We

[2] This may be viewed as an estimate of the Average Causal Effect [21].

[3] More precisely, we use logits instead of probabilities, as often preferred in classifier explanations, e.g. [29, 4]. If a coordinate has an inconsistent change direction it is discarded.

can then visualize the resulting k modified images. We refer to this strategy as **Independent** selection. Alternatively, it could be that individual attributes do not have a large effect, and thus we can search for a set of k StylEx attributes, whose *joint* modification maximizes classifier change. In order to avoid checking all possible $O(2^k)$ subsets, we perform a greedy search (i.e., at each step find the next most influential attribute for this image, given the subset of attributes selected so far; halt once the classifier has flipped its classification). We can then visualize the effect of modifying this subset. We refer to this as **Subset** selection.

Figures 6 and 7 show examples of image-specific explanations (one per selection strategy; see Sec. 4). We ask what has led the classifier to classify this person as "Perceived Old" and not "Perceived Young", or this leaf as "Sick" and not "Healthy". The figures show the top *image-specific* attributes that drove the classifier to its prediction on this image.

Dataset	Classifier
AFHQ [2]	Cats / dogs
AFHQ	Wild cats species
FFHQ [14]	Perceived gender
FFHQ	Perceived age
Plant-Village [13]	Healthy / sick leaves
Retinal Fundus [17]	DME / non-DME

Table 1. List of datasets used in this paper.

4. Evaluation and Results

We test our StylEx method on a variety of classifiers from a diverse set of domains, as listed in Table 1. The classifiers are based on the MobileNet [12] architecture, and achieve a high accuracy of at least 95% on their test sets.

4.1. Qualitative Evaluation

We first demonstrate that each StylEx attribute corresponds to clear visual concepts, and then show that these can be used to explain classifier outputs on specific images.

Visualizing StylEx Attributes. We begin by showing that the attributes discovered by StylEx indeed correspond to coherent semantic notions (see Sec. 4.2.2 for user studies that further demonstrate this). We emphasize that we do not choose attributes by manual inspection but rather use the top attributes found by the *AttFind* procedure.

The semantic meaning of the attribute already becomes clear even when inspecting its effect on a pair of images: one where its effect is increased, and one where it's decreased. Figures 4 and 5 demonstrate this for four domains. For each domain we consider the top StylEx attributes. For each attribute, we find an image in each class where attribute modification was significant, and display the original and modified image. See results and animated-GIFs in Supplemental to view the these counter-factual explanations *dynamically*.

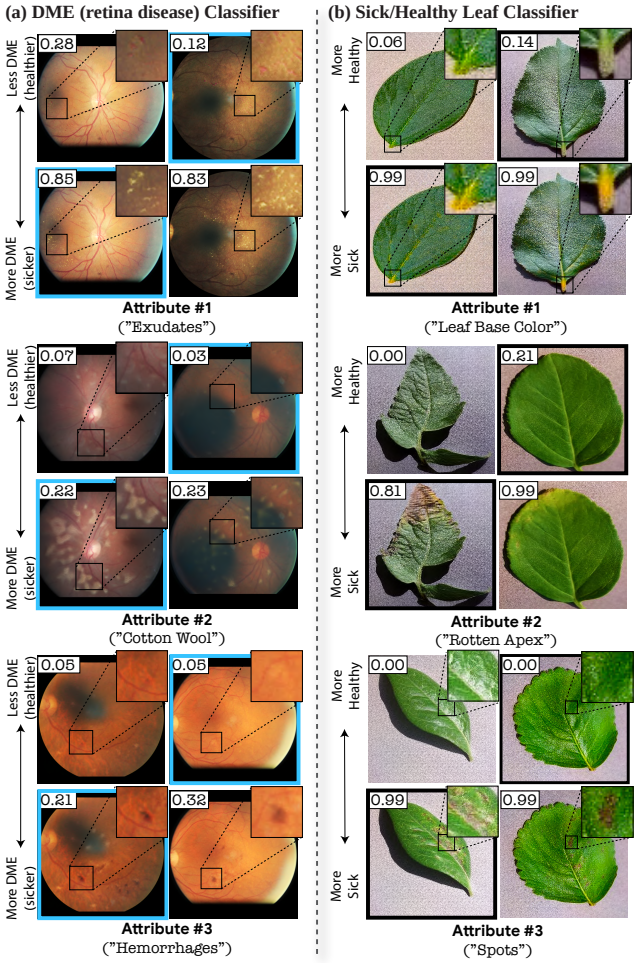

Figure 5. **Top-3 automatically detected attributes for (a) DME classifier of retina images and (b) Classifier of Sick/Healthy leaf images.** The respective classifier scores are presented in their top left corner. The generated counterfactual examples are marked by a frame. The top discovered attributes for both classifiers turn out to be well aligned with known disease indicators ([22], [13])..
Please see animated-GIFs in Supplemental to view these counterfactual changes (explanations) dynamically.

It can be seen that each of the top attributes extracted by StylEx is visually-interpretable. Additionally, modifying each attribute leads to a significant change in the classifier output. The quality of the explanations that StylEx provides can be demonstrated in the cases of the healthy/sick leaves classifier and the retina DME classifier (Fig. 5) where the top discovered attributes are aligned with disease indicators that appear in the corresponding datasets ([13, 22]).

Providing Image-Specific Explanations. Thus far we showed that each StylEx attribute corresponds to an interpretable visual concept. We can now use these to provide counterfactual explanations for specific images. Namely, for a given image we can provide statements such as: "had you

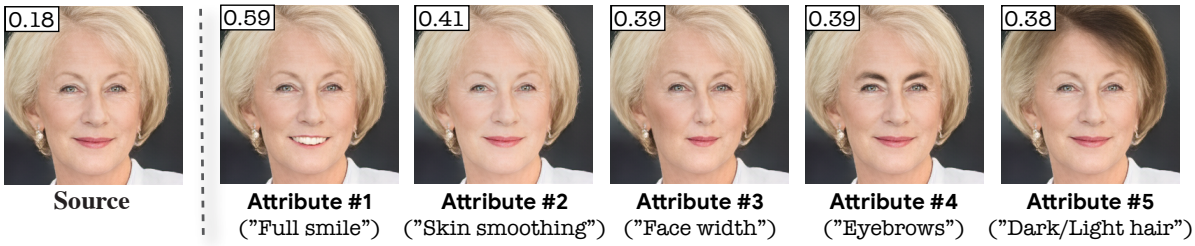

Figure 6. **Image-specific explanations:** Top-5 automatically detected attributes for explaining a perceived-age classifier for a specific image using the **Independent** selection strategy. Attributes are sorted by their effect on the classification of the specific image, resulting in different attributes 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.

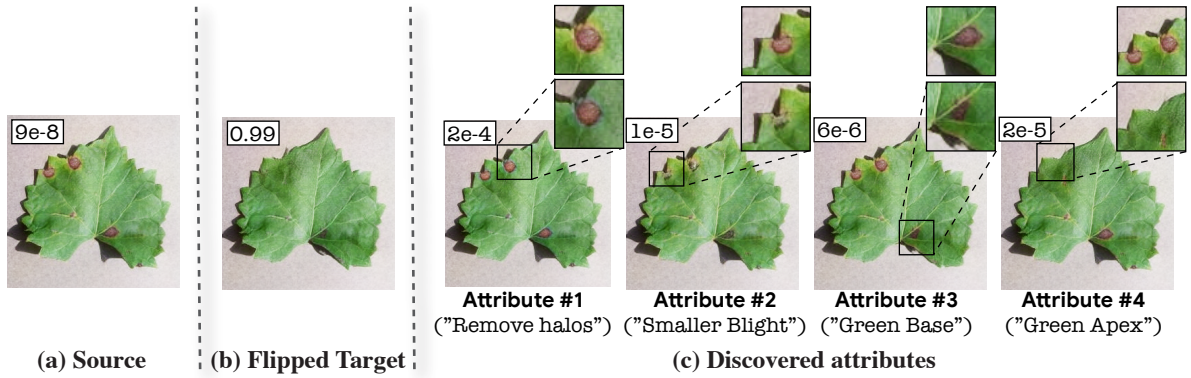

Figure 7. **Combining several attributes**. Attributes in **(c)** are selected by the **Subset** method to inflict the largest accumulated effect on the classification of the image in **(a)**. Interventions on individual attributes result in a small change of the classifier output, yet intervening on all of them results in image **(b)** where the classification is flipped. The classifier probabilities of healthy are presented in the top-left corner.

changed attribute #1 and #3, the classifier output would have changed". Since attribute #1 and #3 have clear semantics (e.g., #1 is adding a moustache) the counterfacutal explanation is easy to understand and is informative. To find a counterfactual explanation, we use the method in Sec. 3.4.

Fig. 6 illustrates this for the "Perceived Age" classifier, where we use the **Independent** selection method. It can be seen that there are five attributes that individually affect that classifier output considerably. Also, these are not the attributes with largest average effect across many images (see Fig. 4), but rather those that most affect this *specific* image. Fig. 7 shows an example for the Plants domain, where we use the **Subset** selection method. Here each of the selected attributes has a smaller effect (though the change in logit is significant), but the combined effect of changing the three attributes results in flipping the classifier decision. These examples nicely demonstrate that StylEx can be used to "decompose" classifier decisions into a set of visual attributes.

4.2. Quantitative Evaluation

It is not immediately clear how to evaluate multi-attribute counterfactual explanations. However, the three following criteria seem key to any such method:

Visual Coherence. Attributes should be identifiable by humans. For instance, the effect of a coordinate that controls pupil dilation in cats can be easily understood by humans after seeing a few examples. On the other hand, if the coordinate changes different visual attributes for each image (e.g. dilates pupils in some images, while shortening ears in others) then understanding its effect is a more difficult task, resulting in a less coherent visual attribute.

Distinctness. Extracted attributes should be distinct. Having distinct attributes lets us compose several counterfactual explanations that expose different elements underlying classifier decisions (e.g., as opposed to GANalyze [7]).

Effect of Attributes on Classifier Output. Changing the value of attributes in an image should result in a change in classifier output. Furthermore, different attributes should have complementary effects so that modifying multiple attributes will result in flipping the decision of the classifier on most images.

4.2.1 Baselines and Model Variants

As discussed in Sec. 2, to-date, multi-attribute counterfactual explanations of visual classifiers have not yet been achieved for real images. Thus, there are no directly comparable baselines in the literature. However, most closely related to our method is the original StyleSpace defined in [33]. While [33] was not proposed as a method to explain a classifier, we can use it as a baseline to test two key components our method adds to the StyleSpace framework: (i)

classifier-specific training (CST) of the StyleGAN and (ii) the *AttFind* method for finding classifier-related coordinates in StyleSpace. To test the importance of these two contributions, we compare against StylEx without CST and also against using the StyleSpace selection method in [33]:

- **StylEx w/o Classifier-Specific Training (CST):** The training procedure for StylEx incorporates the classifier into the StyleGAN training procedure to obtain a classifier-specific StyleSpace. Here we consider the effect of using our *AttFind* procedure with a standard StyleGAN2 that does not involve the classifier.

- **Wu *et al.* [33]**: [33] proposed identifying StyleSpace coordinates that relate to known visual labels. These coordinates can be identified by measuring the normalized difference between the coordinate values on each of the labels. As a baseline we use their method to find the top-M classifier-related coordinates in standard StyleGAN2 StyleSpace and compare against our method, which instead uses *AttFind* and a StyleSpace trained with CST.

4.2.2 A User Study for Coherence and Distinctness

To evaluate coherence and distinctness we conducted a two-part user study. The first part uses a setup similar to [35]. Users are shown four animated GIFs, each corresponding to modifying an attribute for a given image. The left two GIFs are produced from an attribute i and the right two GIFs are from attributes i and j. The user is asked to identify the right GIF corresponding to attribute i (see supplementary for more information). We use the top 6 style coordinates for StylEx and for Wu *et al.*, and perform the experiment using each of these sets. A correct answer shows that attributes are distinct, since users are able to distinguish between the extracted attributes. It also establishes visual coherence since users are able to classify animations as belonging to the attribute based only on two examples. The results in Table 2 show that StylEx achieves high accuracy in this task, suggesting attributes are distinct and coherent. It also outperforms Wu *et al.* on all domains but Plants. However, on Plants we show in Sec. 4.2.3 that Wu *et al.* attributes have little effect on the classifier.

	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.

For the second part of the study (performed on a different set of users) we show 4 GIFs demonstrating an intervention on a single style coordinate. Users are then asked to describe in 1-4 words the single most prominent attribute they see changing in the image. We perform these experiments for Face classifiers and Cats/Dogs. These datasets are chosen since they are more familiar to a layperson, making it more likely that users write similar words when describing an attribute. We provide the full answers of users in the supplementary material, yet a qualitative assessment of the responses leads to similar conclusions as in the first part of the study regarding the distinctness and coherence. For instance, all users used the word "glasses" when describing the top coordinate extracted by StylEx for the perceived age classifier. In general for all coordinates extracted by StylEx, except one, there is a common word shared by all descriptions, and only in two coordinates the most common word is the same. On the other hand, for Wu *et al.* less than half of the coordinates have a common word in all their descriptions, and two pairs of coordinates have the same most common word.

4.2.3 'Sufficiency': Effect of Attributes on Classifier Output

To test the effect of the attributes on the classifier, we ask if interventions on a small set of attributes can flip the classifier decision. Specifically, we try modifications on top k attributes up-to $k = 10$. We then measure the fraction of images that can be flipped (hence explained) in this way.

	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.

Table 3 presents this measure on 1000 randomly chosen images. It can be seen that StylEx achieves high explanation percentages on most domains. Table 3 also reports results of StylEx without the classifier-specific training (i.e. without conditional training and classifier loss). Note that this component has a dramatic effect on performance in the retina and

plants domains. This is in line with the fact that the classes in these cases correspond to features that are more subtle and localized, thus less likely to be captured by a GAN that is oblivious to the classifier. Specifically, we verified that when training without classifier information, the generated images in the retina domain collapse to one class ("healthy").

5. Conclusion

We introduced a new technique for generating different counterfactual explanations for a given classifier on a given image. Our results demonstrate that these attributes correspond to clear visual concepts *and* directly affect classifier decisions. We believe that StylEx is a promising step towards detection and mitigation of previously unknown biases in classifiers. Additionally, our focus on multiple-attribute based counterfactuals is key to providing new insights about previously opaque classification processes and aiding in the process of scientific discovery.

References

[1] Javier Antorán, Umang Bhatt, Tameem Adel, Adrian Weller, and José Miguel Hernández-Lobato. Getting a clue: A method for explaining uncertainty estimates. *arXiv preprint arXiv:2006.06848*, 2020. 3

[2] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8188–8197, 2020. 5, 6

[3] Edo Collins, Raja Bala, Bob Price, and Sabine Susstrunk. Editing in style: Uncovering the local semantics of gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5771–5780, 2020. 2

[4] Kedar Dhamdhere, Mukund Sundararajan, and Qiqi Yan. How important is a neuron? *arXiv preprint arXiv:1805.12233*, 2018. 5

[5] Patrick Esser, Robin Rombach, and Bjorn Ommer. A disentangling invertible interpretation network for explaining latent representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9223–9232, 2020. 3

[6] Amirata Ghorbani, James Wexler, James Zou, and Been Kim. Towards automatic concept-based explanations. *arXiv preprint arXiv:1902.03129*, 2019. 3

[7] Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. Ganalyze: Toward visual definitions of cognitive image properties. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5744–5753, 2019. 2, 3, 7

[8] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 2672–2680, Cambridge, MA, USA, 2014. MIT Press. 4, 5

[9] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 2

[10] Yash Goyal, Amir Feder, Uri Shalit, and Been Kim. Explaining classifiers with causal concept effect (cace). *arXiv preprint arXiv:1907.07165*, 2019. 3

[11] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In *International Conference on Machine Learning*, pages 2376–2384. PMLR, 2019. 2, 3

[12] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR*, abs/1704.04861, 2017. 6

[13] David Hughes, Marcel Salathé, et al. An open access repository of images on plant health to enable the development of mobile disease diagnostics. *arXiv preprint arXiv:1511.08060*, 2015. 6

[14] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 6

[15] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020. 2, 4, 5

[16] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018. 2

[17] Jonathan Krause, Varun Gulshan, Ehsan Rahimy, Peter Karth, Kasumi Widner, Greg S Corrado, Lily Peng, and Dale R Webster. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. *Ophthalmology*, 125(8):1264–1272, 2018. 6

[18] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 607–617, 2020. 2

[19] Arunachalam Narayanaswamy, Subhashini Venugopalan, Dale R Webster, Lily Peng, Greg S Corrado, Paisan Ruamviboonsuk, Pinal Bavishi, Michael Brenner, Philip C Nelson, and Avinash V Varadarajan. Scientific discovery by generating counterfactuals using image translation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 273–283. Springer, 2020. 2

[20] Matthew O'Shaughnessy, Gregory Canal, Marissa Connor, Mark Davenport, and Christopher Rozell. Generative causal explanations of black-box classifiers. *arXiv preprint arXiv:2006.13913*, 2020. 3

[21] Judea Pearl. *Causality*. Cambridge university press, 2009. 5

[22] Alexander Rakhlin. Diabetic retinopathy detection through integration of deep learning classification framework. *bioRxiv*, page 225508, 2018. 6

[23] Sylvestre-Alvise Rebuffi, Ruth Fong, Xu Ji, and Andrea Vedaldi. There and back again: Revisiting backpropagation saliency methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8839–8848, 2020. 3

[24] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 2, 3

[25] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of GANs for semantic face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9243–9252, 2020. 2

[26] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by GANs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 2

[27] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in GANs. *arXiv preprint arXiv:2007.06600*, 2020. 2

[28] Assaf Shocher, Yossi Gandelsman, Inbar Mosseri, Michal Yarom, Michal Irani, William T Freeman, and Tali Dekel. Semantic pyramid for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7457–7466, 2020. 3

[29] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International Conference on Machine Learning*, pages 3145–3153. PMLR, 2017. 5

[30] Sumedha Singla, Brian Pollack, Junxiang Chen, and Kayhan Batmanghelich. Explanation by progressive exaggeration. *arXiv preprint arXiv:1911.00483*, 2019. 2, 3

[31] Sumedha Singla, Brian Pollack, Stephen Wallace, and Kayhan Batmanghelich. Explaining the black-box smoothly-a counterfactual approach. *arXiv preprint arXiv:2101.04230*, 2021. 3

[32] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017. 2

[33] Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for StyleGAN image generation. *arXiv preprint arXiv:2011.12799*, 2020. 2, 4, 7, 8

[34] Shawn Xu, Subhashini Venugopalan, and Mukund Sundararajan. Attribution in scale and space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9680–9689, 2020. 3

[35] Chih-Kuan Yeh, Been Kim, Sercan Arik, Chun-Liang Li, Pradeep Ravikumar, and Tomas Pfister. On completeness-aware concept-based explanations in deep neural networks. 2020. 3, 8

[36] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5