Vision Model Diagnosis: A Generative Perspective

Jinqi Luo

July 26, 2023

The Robotics Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, Pennsylvania, U.S.A.

Thesis Committee:
Prof. Fernando De la Torre, *co-chair*
Dr. Dong Huang, *co-chair*
Prof. Jun-Yan Zhu
Prof. Zachary Lipton
Sheng-Yu Wang

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Robotics.

Abstract

In the evolving landscape of computer vision, deep learning has emerged as a transformative force, enhancing a myriad of societal facets. The deployment of these models necessitates rigorous evaluation and analysis, particularly when outcomes bear significant societal implications, such as their influence across varied ethnicities and genders. This imperative forms the nucleus of trustworthy deep learning, which aims to equip scientists and engineers with an understanding of the robustness, interpretability, fairness, safety, and tractability of these models.

Traditionally, Model Diagnostics [7] refers to the validity assessment of a regression model, including the exploration of assumptions and the structural examination. As we transit into the era of deep learning for computer vision, we reinterpret Vision Model Diagnosis (VMD) as the systematic analysis and evaluation of deep vision models. As we increasingly delegate decision-making power to deep learning vision systems, their output can significantly impact individuals and society. Hence the process of VMD, which has attracted increasing attention from the research community, can enable us to comprehend the deep vision model's behavior, interpret its performance, and fix potential shortcomings and biases.

The main goal of this thesis is to provide a thorough understanding from a generative perspective: how generative models can help diagnose the decision-making process of a model, its fairness, and its robust behavior under various conditions. The use of various generative models with different paradigms, including conditional VAE and CLIP-guided StyleGAN, can empower VMD with rich semantic spaces that provide analysis for attributional fairness and visualize where the model fails. We hope that, with this thesis, we can provide valuable insights into how a diagnostic process should be constructed and raise attention in the research community to address issues of model trustworthiness and alignments. How to accurately uncovers a model's potential limitations and weaknesses is essential for securely publicizing deep learning models, and we envision the significant importance of this theme that will be growing fast in the upcoming decades.

Acknowledgments

I would like to express my deepest gratitude to my advisors, Prof. Fernando De la Torre and Dr. Dong Huang, for their invaluable guidance, support, and mentorship throughout my research at the Robotics Institute of Carnegie Mellon University. I sincerely appreciate my research committee members, Prof. Jun-Yan Zhu, Prof. Zachary Lipton, and Sheng-Yu Wang, for their intellectual assistance and resource support. I deeply thank all of my research collaborators for their continuous help and encouragement, without which this thesis would not have been possible. Finally, I am very grateful to my family and friends for their constant love and support throughout my academic journey. Their firm belief in me has always been my source of strength.

Contents

Chapter 1

Introduction

Deep learning models have revolutionized various fields, including computer vision, by providing sophisticated capabilities for tasks like image recognition, object detection, and semantic segmentation. However, these deep models can often unwittingly perpetuate existing biases present in the data they are trained on. Consequently, these biases may be amplified or suppressed depending on the model's architecture and optimization strategy, which can lead to biased decision-making, particularly in sensitive fields like healthcare, law enforcement, and autonomous driving.

Given the significant implications of these biases and the growing need for more fair, transparent, and robust models, extensive testing and evaluation of deep learning models are becoming imperative. In traditional model evaluation, collecting and labeling large-scale datasets that capture all possible attributes of interest has been a common practice. However, this approach is fraught with challenges, including high costs, time intensiveness, and potential errors in labeling. Moreover, acquiring a balanced dataset that is uniformly distributed across all attributes of interest is often impractical due to its combinatorial nature, and even if attained, it cannot guarantee absolute fairness or robustness due to potential discrepancies between test and real distributions.

With these challenges in mind, this thesis aims to advance the field of vision model diagnosis by proposing and exploring techniques of utilizing generative models with manipulable semantic space, thereby seeking to democratize model diagnosis as tool of interpretability, fairness, and robustness. The two primary methods under

investigation are Semantic Image Attack for Visual Model Diagnosis [26] and Zero-shot Model Diagnosis [27]. Our works make use of generative models with different paradigms, including conditional Variational AutoEncoder (VAE) [14] and CLIP-guided StyleGAN [20, 31], to provide rich semantic analysis for a vision model's sensitivities across attributes and visualize where the model fails.

Semantic Image Attack performs model diagnosis with joint optimization in controllable attribute space and pixel space constructed by an attribute-conditioned VAE. Zero-shot Model Diagnosis, on the other hand, generates counterfactual images with more advanced generative backbones, which can visualize the sensitive factors of an input image that can influence the model's outputs, thereby identifying key factors where the model fails. This approach leverages the zero-shot capabilities of CLIP [34] and StyleGAN for text-driven applications and semantic attribute editing.

Both methods share a common goal: to offer ways to diagnose and understand model behavior without the need for costly, time-consuming, and potentially error-prone test set collection and annotation. Furthermore, these methods provide practical solutions for diagnosing new models or exploring new user-defined attribute spaces without the need for system re-training.

In this thesis, we will discuss these methodologies in detail, their advantages, and potential applications. We will also present comparative analyses and case studies to demonstrate their effectiveness.

By synthesizing insights from these two methods, we aspire to contribute a meaningful advancement to the field of model diagnosis, promoting more fairness, transparency, and robustness in deep learning models.

In the forthcoming chapters, we will delve deeper into the technical background, related works, the methodologies of Zero-shot Model Diagnosis and Semantic Image Attack, comparative analysis of these techniques, and the future prospects of model diagnosis.

Chapter 2

Semantic Image Attack for Visual Model Diagnosis

In practice, metric analysis on a specific train and test dataset does not guarantee reliable or fair ML models. This is partially due to the fact that obtaining a balanced, diverse, and perfectly labeled dataset is typically expensive, time-consuming, and error-prone. Rather than relying on a carefully designed test set to assess ML models' failures, fairness, or robustness, this chapter proposes Semantic Image Attack (SIA), a method based on the adversarial attack that provides semantic adversarial images to allow model diagnosis, interpretability, and robustness. Traditional adversarial training is a popular methodology for robustifying ML models against attacks. However, existing adversarial methods do not combine the two aspects that enable the interpretation and analysis of the model's flaws: semantic traceability and perceptual quality. SIA combines the two features via iterative gradient ascent on a predefined semantic attribute space and the image space. We illustrate the validity of our approach in three scenarios for keypoint detection and classification. (1) Model diagnosis: SIA generates a histogram of attributes that highlights the semantic vulnerability of the ML model (i.e., attributes that make the model fail). (2) Stronger attacks: SIA generates adversarial examples with visually interpretable attributes that lead to higher attack success rates than baseline methods. The adversarial training on SIA improves the transferable robustness across different gradient-based attacks. (3) Robustness to imbalanced datasets: we use SIA to augment the underrepresented

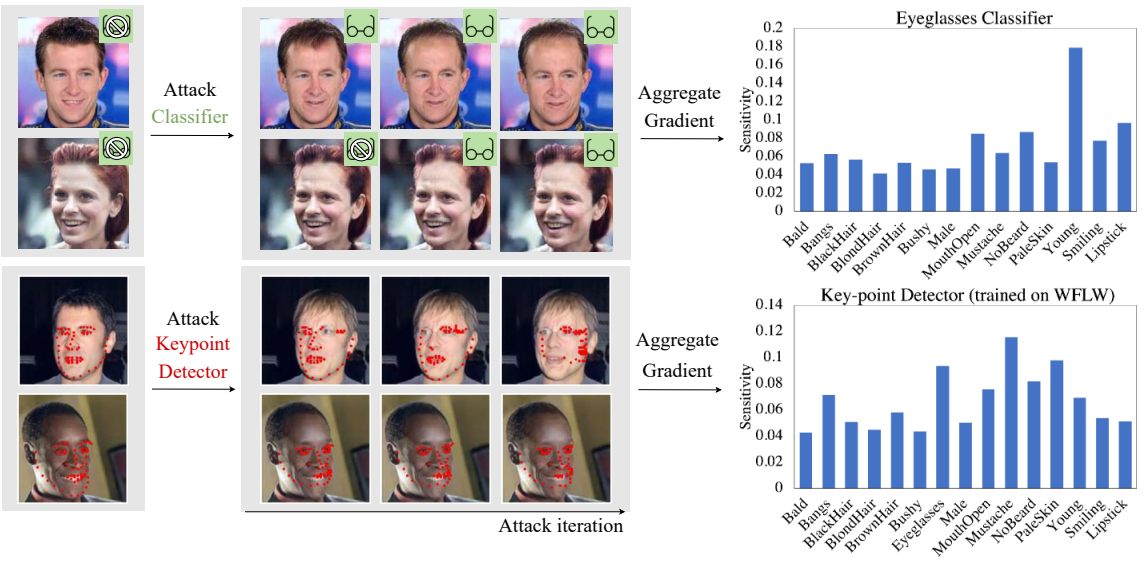


Figure 2.1: Model diagnosis by SIA. The models to be diagnosed are an eyeglasses classifier (top two rows) and a keypoint detector (bottom two rows). SIA reveals that the eyeglasses classifier is more sensitive to lipstick and age, whereas the keypoint detection tends to fail on people with mustaches and pale skin. See the text for an explanation of the figure.

classes, which outperforms strong augmentation and re-balancing baselines.

2.1 Introduction

In Machine Learning (ML), error analysis of train and test data is a critical stage in model assessment and debugging. However, the conclusions extracted from the metric analysis on the train or test data do not guarantee reliability nor fairness, partially due to the fact that datasets are imperfect [37, 35]. Even with careful collection and filtering, data naturally contain biases. Furthermore, in the case of computer vision learning systems, having a uniform distribution over all conceivable variability of an object in an image (e.g., position, lighting, background) is typically impractical (i.e., exponential) and labels are prone to errors. The issue only grows more severe with large-scale datasets. ML models trained on these datasets inevitably inherit these imbalances and biases. These limitations also apply to test sets that are typically used for model evaluation. Such a vulnerability is a landmine that must be recognized and processed in order for ML applications to succeed. The question we strive to

address in this study is whether there are alternative/better methods for discovering biases and performing model diagnostics in computer vision models instead of only relying on a test set.

Fig. 2.1 illustrates the problems that this section tries to address. Given an eyeglasses classifier (top two rows) or a keypoint detector (bottom two rows), which kind of face images will lead to misclassification or misdetection? How can we automatically discover these failure cases and robustify the model? How can we perform visual model diagnosis in a semantic attribute space? To accomplish these, we propose Semantic Image Attack (SIA), a new adversarial attack in a generative model of faces parameterized by attributes. In top left in Fig. 2.1, we see two images of faces without eyeglasses, and the model classifies them correctly. After several iterations of SIA (right column), our model is able to modify facial attributes (e.g., smile, eye color, facial hair) to mislead the eyeglass classifier. Also, our model builds a histogram of the sensitivity across attributes (i.e., visual model diagnosis). While evaluating the model resilience on a single attribute can be relatively straightforward, evaluating the model robustness for combinations of attributes can be quite challenging (due to the exponential nature of attribute combinations). SIA is able to *jointly* search over the space of attributes, and hence performs a multi-attribute attack for model diagnosis. Similarly, in Fig. 2.1, the bottom two rows illustrate the model diagnosis results for keypoint detection.

In addition to model diagnosis, SIA is able to robustify the target model by re-training the model on adversarial examples (see Fig. 2.1 middle columns). In our experiments, we also show the robustness from SIA is more transferable to other types of attacks than other competing attack methods. Finally, we show that SIA outperforms popular image augmentation techniques [6, 51] and re-balancing baselines when learning from imbalanced datasets.

2.2 Related Work

2.2.1 Adversarial Attacks

Gradient-guided image space perturbation attacks have been popular in robustifying ML models [11, 28]. The image perturbations generated by such attacks are small

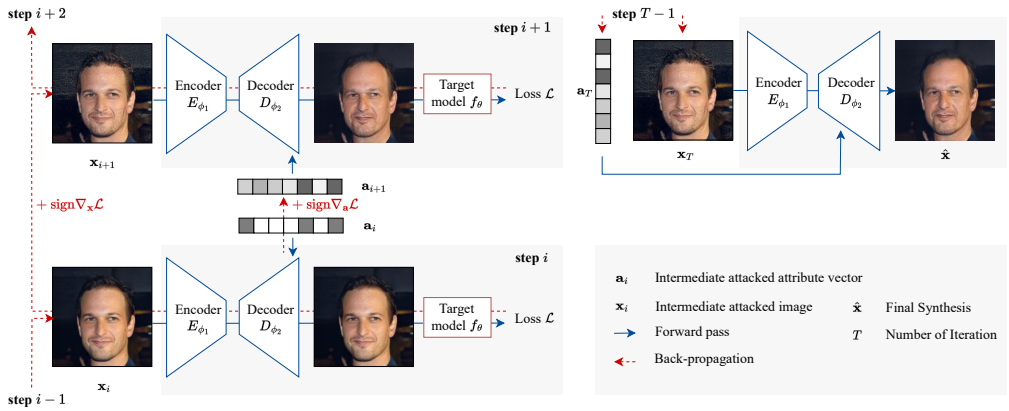


Figure 2.2: The SIA framework uses an encoder-decoder GAN $\mathcal{G}_\phi = \{E_{\phi_1}, D_{\phi_2}\}$ to attack a target model f_θ. In each iteration i, we update the image $\mathbf{x}_i$ and the attribute vector $\mathbf{a}_i$ using the gradients from the loss $\mathcal{L}$ (see Eq. (2.1)). Finally, the encoder-decoder GAN projects the attacked image $\mathbf{x}_T$ and attributes $\mathbf{a}_T$ in the last iteration back to the image manifold to produce adversary $\hat{\mathbf{x}} = \mathcal{G}_\phi(\mathbf{x}_T, \mathbf{a}_T)$. Solid lines stand for forward passes, and dashed lines stand for backpropagation.

image changes typically imperceptible to humans. [54, 50] adopted such attacks on keypoint detectors to robustify detectors against adversarial perturbations. [49] was pioneering in using Generative Adversarial Networks (GANs) [10] to generate adversarial attacks. However, [49] only allowed a limited perturbation bound and required individually trained GANs for every target ML model. A major issue of previous methods is the lack of interpretability of the attack. To address this issue, [33] used the interpolation of semantic feature maps to generate attacks, and showed the effectiveness in terms of the attack's success rate in classification and detection problems. [9] also modeled the perturbations in the attribute space, and showed that the attribute space can improve model robustness. However, this work aims to find perturbations in samples that do not change labels, and their model is not robust to small perturbation attacks in the image space. Moreover, [9] did not provide interpretability into the failures of the computer vision model. Similarly, [23] sampled images in the latent space of a GAN to generate strong attacks, but their attacks are not interpretable in the attribute space. [17] conducted model attacks only in the attribute space using the attribute-assisted GAN (AttGAN) [14]. This approach does not attack the image space and does not constrain the scale of parametric gradients, which leads to generating unrealistic images.

Unlike previous work in the adversarial attack literature, SIA performs gradient-guided attack simultaneously in the image and a pre-defined attribute space. As we will show in the experimental section, performing gradient ascent only in the attribute space leads to unstable results. In addition, our approach only uses *one* GAN backbone [14] to attack all target models (i.e., AttGAN can be used to evaluate any computer vision model). Finally, our method provides a histogram of the sensitivity of the target models across attributes of interest. This information can be critical to gather insights into the fairness and robustness of the model.

2.2.2 Bias and Fairness Analysis

[2, 8] showed that by traversing images in the GAN latent space, one can visualize the attribute-wise sensitivity of a target classifier. But such a process requires manual annotation of the generated images, which is expensive and infeasible for large attribute spaces. Recently, [21] used StyleGAN [20] to learn a target-model-dependent style/attribute space, which allows a human to interpret the target models' behavior in terms of attributes. Furthermore, several previous works proposed fairness metrics to evaluate a model without a fair test set [13, 35, 52]. While previous fairness metrics focus on a model's statistical behavior across attributes, SIA focuses on the model's decision for each instance (though individual sensitivities can be further aggregated to get sub-population sensitivity, see Fig. 2.1). Moreover, SIA is able to search over attribute combinations.

2.3 Method

This section describes our SIA algorithm starting with the notation.

Target model (f_θ): Let f_θ, parameterized by θ, be the target model that we want to improve or perform model diagnosis on. In this section, we cover two types of neural network models f_θ: an attribute classifier and a keypoint detector.

An attribute classifier takes an image $\mathbf{x}$ as input and outputs $f_\theta(\mathbf{a}|\mathbf{x})$, the conditional probability of attribute $\mathbf{a} \in \mathcal{A}$ given $\mathbf{x}$, where $\mathcal{A}$ is the attribute space. Without loss of generality, we consider binary classifiers. Given the ground truth class label c of the image $\mathbf{x}$, the classification loss is defined as the binary cross-entropy

$\mathcal{L}_\theta = -(c \log f_\theta(c|\mathbf{x}) + (1-c)(\log(1 - f_\theta(c|\mathbf{x}))))$.

The keypoint detector takes an image $\mathbf{x}$ as input and outputs $f_\theta(\mathbf{p}|\mathbf{x})$, the probability heatmap of the keypoints $\mathbf{p} \in \mathcal{P}$, where $\mathcal{P}$ is the 2D pixel coordinate space. Given a training image $\mathbf{x}$ with ground truth facial keypoints $\mathbf{c}$, the loss $\mathcal{L}_\theta$ is defined as the mean squared error between the predicted heatmap and the ground-truth heatmap corresponding to $\mathbf{c}$, see [43] for details.

Adversary ($\hat{\mathbf{x}}$): For each input image $\mathbf{x}$, an adversarial example $\hat{\mathbf{x}}$ is a synthesized image that misleads the target model f_θ to produce outputs that are far away from the ground truth $\mathbf{c}$ or changes the label of the classifier. Different from traditional adversarial attack methods, SIA generates adversarial examples under a combination of perturbations in the attribute and image spaces.

SIA consists of two main components: (1) an AttGAN $\mathcal{G}_\phi = \{E_{\phi_1}, D_{\phi_2}\}$, $\mathcal{G}_\phi(\mathbf{x}, \mathbf{a}) = D_{\phi_2}([E_{\phi_1}(\mathbf{x}); \mathbf{a}])$, where the encoder E_{ϕ_1} maps an input image $\mathbf{x}$ to a latent vector, the decoder D_{ϕ_2} takes as an input the concatenation of $E_{\phi_1}(\mathbf{x})$ and the attribute vector $\mathbf{a}$ to generate an image; (2) a pretrained target model f_θ to be diagnosed.

2.3.1 Generating Iterative Adversaries

Our framework uses both the attribute space and the image space to iteratively generate adversarial images $\hat{\mathbf{x}}$. We iteratively compute gradient ascent in the attribute space and the image space. An advantage of optimizing over the attribute and image space is an improved adversarial space, that leads to a better generation of adversarial examples (see experiment section).

The procedure to *jointly* update the attribute vectors and images is as follows:

$$\begin{aligned} \mathbf{a}_i &= \Pi_{\mathcal{B}(\epsilon_\mathbf{a})}(\mathbf{a}_{i-1} + \eta \operatorname{sign}[\nabla_\mathbf{a}(\mathcal{L}_\theta(f_\theta(\mathcal{G}_\phi(\mathbf{x}_{i-1}, \mathbf{a}_{i-1}))))]), \\ \mathbf{x}_i &= \Pi_{\mathcal{B}(\epsilon_\mathbf{x})}(\mathbf{x}_{i-1} + \eta \operatorname{sign}[\nabla_\mathbf{x}(\mathcal{L}_\theta(f_\theta(\mathcal{G}_\phi(\mathbf{x}_{i-1}, \mathbf{a}_{i-1}))))]). \end{aligned} \tag{2.1}$$

The adversarial example $\hat{\mathbf{x}}$ is an image space projection of a fine-grained perturbation of the original input image $\mathbf{x}$ at both pixel and attribute levels. During the process, our SIA framework manipulates the attribute vector in a predefined attribute space such that the target model is compromised. Note that each iteration of the updates will be clipped with a radius ϵ to make sure that the perturbation is bounded and valid. The pixel-level perturbed image is fed into $\mathcal{G}_\phi$ to encode the adversarial information into the latent vector, which is concatenated with the perturbed attribute vector.

Specifically, instead of directly perturbing the output image, which may significantly harm the perceptual quality, we perturb the input attribute and the image and let $\mathcal{G}_\phi$ project the perturbed image and attribute back to the image manifold. To prevent synthesis collapse, we adopt the projection Π onto the ℓ_∞ ball $\mathcal{B}$ of radius ϵ to constrain the optimization. The projection to generate the final adversarial example is formulated as $\hat{\mathbf{x}} = \mathcal{G}_\phi(\mathbf{x}_T, \mathbf{a}_T)$. An overview of our SIA framework is shown in Figure 2.2.

At this point, it is important to notice that perturbing in both the image space and attribute space produces higher attack success rate and finer visual adversarial images. Also, we do it for a fair comparison with traditional methods. Recall that directly perturbing the semantic space limits the attacking capability. Our hybrid attack gives us the flexibility to analyze both the semantic and pixel-level robustness of the model. In fact, SIA's pixel-level perturbation helps to avoid exaggerated semantic variation that makes the image generation collapse. An ablation study that illustrates the advantages of perturbing in both the image and attribute space is included in the experimental section.

2.3.2 Interpreting and Improving the Target Model

Given a set of image-attribute pairs $(\mathbf{x}^{(p)}, \mathbf{a}^{(p)})$ $(p = 1, \ldots, N)$, we run T iterations of Eq. 2.1 and store all the generated adversaries. By calculating the absolute variation of attributes during the generation of adversaries $\hat{\mathbf{x}}^{(p)}$, we can discover the most sensitive attribute(s) to the target model $f_\theta(\cdot)$ in the $\mathcal{G}_\phi$'s attribute space. We define the sensitivity vector containing sensitivities (in the range of $[0, 1]$) of the target model on each attribute as follows:

$$\mathbf{s} = \frac{1}{N} \sum_{p=1}^{N} (|\mathbf{a}_T^{(p)} - \mathbf{a}_1^{(p)}|), \tag{2.2}$$

Each value in $\mathbf{s}$ will represent the average perturbation of the corresponding attribute across all sampled images. Note that this method can be extended to select top-k attributes that have a greater influence on the prediction of the target model. The generated adversaries $\hat{\mathbf{x}}^{(p)}$ are associated with more diverse attribute vectors $\hat{\mathbf{a}}$, which can be considered as an augmented dataset for adversarial training. See Algorithm 1

Algorithm 1 SIA to generate adversarial examples and sensitivity analysis.

Input: A set of image-attribute pairs $\{(\mathbf{x}_0^{(p)}, \mathbf{a}_0^{(p)})\}_{p=1}^N$; target model $f_\theta(\cdot)$
Output: Model sensitivity $\mathbf{s}$; a set of adversaries $\{\hat{\mathbf{x}}^{(p)}\}_{p=1}^N$
for $p \in \{1, \ldots, N\}$ **do**
  for $i \in \{1, \ldots, T\}$ **do**
    $\mathbf{a}_i^{(p)} \leftarrow \mathbf{a}_{i-1}^{(p)} + \eta\, \text{sign}[\nabla_{\mathbf{a}}(\mathcal{L}_\theta(f_\theta(\mathcal{G}_\phi(\mathbf{x}_{i-1}^{(p)}, \mathbf{a}_{i-1}^{(p)}))))]$
    $\mathbf{a}_i^{(p)} \leftarrow \Pi_{\mathcal{B}(\epsilon_{\mathbf{a}})}(\mathbf{a}_i^{(p)})$
    $\mathbf{x}_i^{(p)} \leftarrow \mathbf{x}_{i-1}^{(p)} + \eta\, \text{sign}[\nabla_{\mathbf{x}}(\mathcal{L}_\theta(f_\theta(\mathcal{G}_\phi(\mathbf{x}_{i-1}^{(p)}, \mathbf{a}_{i-1}^{(p)}))))]$
    $\mathbf{x}_i^{(p)} \leftarrow \Pi_{\mathcal{B}(\epsilon_{\mathbf{x}})}(\mathbf{x}_i^{(p)})$
  end for
  $\hat{\mathbf{x}}^{(p)} \leftarrow \mathcal{G}_\phi(\mathbf{x}_T^{(p)}, \mathbf{a}_T^{(p)})$
end for
$\mathbf{s} = \frac{1}{N}\sum_{p=1}^N (|\mathbf{a}_T^{(p)} - \mathbf{a}_1^{(p)}|)$

for more details on how to generate adversaries and sensitivity analysis.

2.4 Experimental Results

This section explains the experimental validation to demonstrate the benefits of SIA for visual model diagnostics, improved robustness against visual attacks, and imbalanced robustness.

2.4.1 Experimental Setups

Attribute-assisted GAN: Our backbone of AttGAN $\mathcal{G}_\phi$ is trained on the whole CelebA dataset [25], using 15 attributes[1]. Images are center cropped, resized to $(224, 224)$, and normalized using the ImageNet normalization. $\mathcal{G}_\phi$'s encoding and decoding dimensions are both 64. Shortcuts and inject layers are activated, and the Wasserstein loss [1] is used. We used the codes provided by [14][2].

Attribute Classifier: Our classifiers are fine-tuned from TorchVision's pre-trained ResNet50. Unless otherwise stated, we trained binary classifiers on the CelebA training set [25]. For training, we used the Adam optimizer with a learning

[1]we used Bald, Bangs, Black_Hair, Blond_Hair, Brown_Hair, Bushy_Eyebrows, Eyeglasses, Male, Mouth_Slightly_Open, Mustache, No_Beard, Pale_Skin, Young, Smiling, Wearing_Lipstick
[2]https://github.com/elvisyjlin/AttGAN-PyTorch

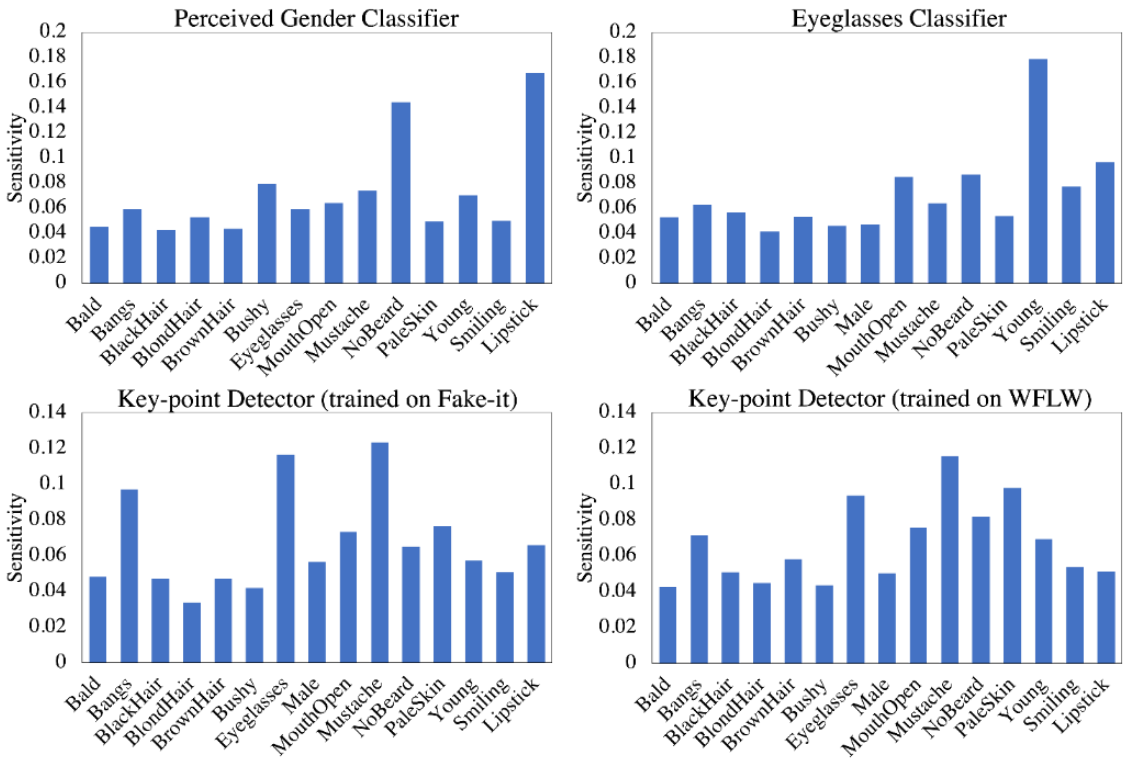


Figure 2.3: Attribute sensitivity analysis generated by SIA for different classifiers (top) and keypoint detectors (bottom). Perceived gender and eyeglasses classifiers are sensitive to different attributes. However, the keypoint detectors trained on synthetic (left) and real (right) data are sensitive to similar attributes, but the one trained on synthetic data is slightly more sensitive than the one trained on real data.

rate of 0.001 and batch size of 128. The seed for random number generation is 42 for Numpy and PyTorch.

Keypoint Detector: We used the HR-Net architecture [43]. We trained two models, one trained on the Wilder Facial Landmark in the Wild (WFLW) dataset [46] and the other on the Microsoft (Fake-it) synthetic dataset [45]. To train the two keypoint detectors, we used all images $(10,000)$ from the WFLW dataset and the first $10,000$ images from the Fake-it dataset, respectively. We trained with 98 keypoints on the WFLW dataset and 68 keypoints on the Fake-it dataset.

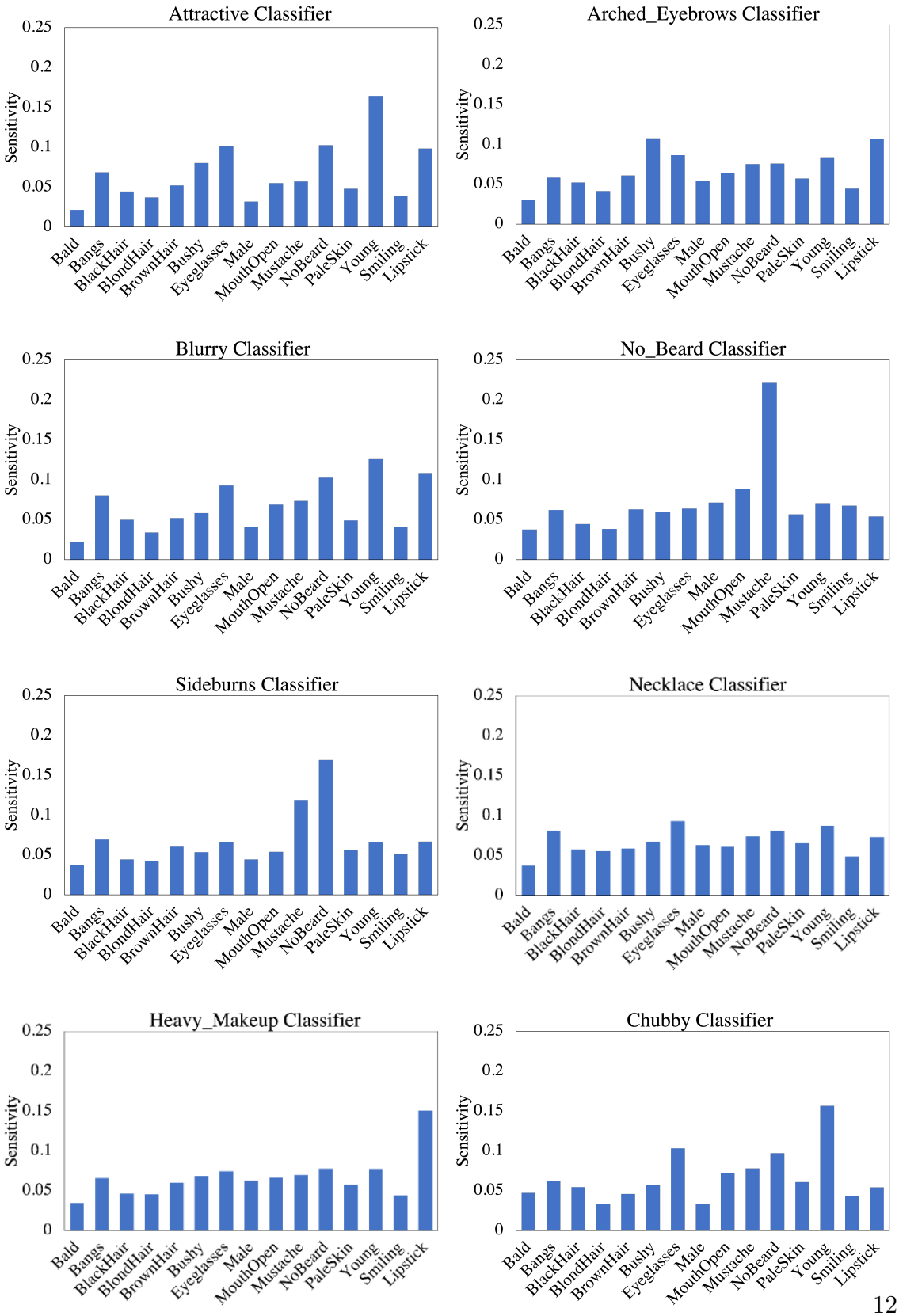
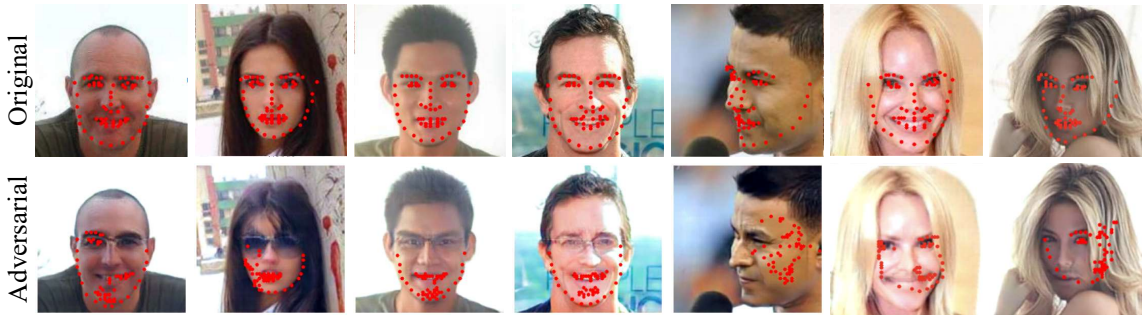


Figure 2.4: Attribute sensitivity analysis generated by SIA for more target classifiers.

(a) Examples of SIA on perceived-gender classifier



(b) Examples of SIA on Fake-it [45] keypoint detector

Figure 2.5: SIA adversarial examples on different target models.

(a) SIA on eyeglass classifier.

(b) SIA on WFLW [46] keypoint detector.

Figure 2.6: (Cont.) SIA adversarial examples on different target models.

2.4.2 Visual Model Diagnosis

After training a deep learning model and tuning hyper-parameters of the model on a validation set, an important step is error analysis. The error analysis includes analyzing where the model fails on test data and making systematic changes based on the insights. However, in some scenarios, it is difficult to collect test data across all possible attributes of interest in a uniform manner. Instead of collecting test data, this section describes how SIA can be used for model diagnosis and provides insights into the image attributes that make the model fail.

Diagnosis visualization

We trained 8 binary classifiers on the following attributes from CelebA: *Attractive, Arched_Eyebrows, Blurry, Chubby, Eyeglasses, Male, No_Beard, Sideburns* with the setup mentioned in Section 4.1. In addition, we trained two keypoint detection algorithms, one on real images and another one on synthetic images, using the same architecture HR-Net [43]. SIA reports the sensitivity of the target model w.r.t. different attributes, which is formalized in Eq. 2.2. We selected the first $10,000$ images in CelebA to evaluate the sensitivities. Fig. 2.3 illustrates the histogram for the classifier (first row) and keypoint detector (second row) towards different attributes, according to Eq. 2.2. For clearer visualization, we have normalized the sensitivity for each attribute by the sum of sensitivities. We can see that for the perceived-gender classifier, lipstick and beard are the most sensitive attributes. Similarly, we discovered that changing specific attributes can largely affect the outcome of a well-trained keypoint detection model. Interestingly, both keypoint detectors are very sensitive to mustache and eyeglasses, and not very sensitive to hair color or perceived gender. This is expected, since keypoints have a higher density around the eyes and mouth region, and modification of these regions can be critical to the accuracy. Fig. 2.4 shows the histograms of the sensitivity across attributes generated for additional attribute classifiers in Section 4.2.

Fig. 2.5 shows example images of SIA attacking the two target models. For the perceived-gender classifier in Fig. 2.5 (a), we can see from the first four columns that mutating the lipstick and beard attributes will influence the model's prediction. The last three columns show that mutating other attributes including hair color,

	PSNR (↑)						SSIM (↑)					
	SPT-50	SIA-50 (Attr)	SIA-50 (Full)	SPT-200	SIA-200 (Attr)	SIA-200 (Full)	SPT-50	SIA-50 (Attr)	SIA-50 (Full)	SPT-200	SIA-200 (Attr)	SIA-200 (Full)
$\eta = \frac{0.25}{255}$	26.63	41.98	**42.24**	19.43	31.21	**33.94**	0.9083	0.9929	**0.9930**	0.7732	0.9602	**0.9718**
$\eta = \frac{4}{255}$	14.18	22.36	**28.25**	13.59	20.16	**25.75**	0.6385	0.8573	**0.9285**	0.6230	0.8037	**0.8926**

Table 2.1: Image quality evaluation for SIA and SPT.

skin color, and bangs can also affect the model decision. Fig. 2.5 (b) shows that SIA changes attribute such as eyeglasses, pale skin, or mustache to cause keypoints misdetection in facial images. This sensitivity analysis and adversarial examples can provide insights into the kind of images where the keypoint detector or classifier fails, and generate adversaries to improve performance. More adversarial examples and histograms for the remaining attributes are shown in Appendix A and B.

Image quality evaluation

We evaluated the image perceptual quality for adversarial examples generated by SPT [17] and SIA. To interpret Table 2.1, SPT-50 ($\eta = \frac{0.25}{255}$) stands for the adversarial examples generated by SPT with 200 iterations and step size of $\frac{0.25}{255}$. The tables show that SIA's image quality is better than SPT under both PSNR (Peak Signal to Noise Ratio) and SSIM (Structured Similarity Indexing Method) [44] metrics. We can see that perturbing in both image space and attribute space produces visually finer adversarial images. In fact, SIA's pixel-level perturbation helps to avoid exaggerated semantic variation that makes the image generation collapse.

Sensitivity by single-attribute optimization

We can also perform SIA independently for every single attribute and organize the sensitivities as the histogram on the right in Fig. 2.7. We can see that SIA's histograms, no matter multi-attribute or single-attribute, support consistent analysis of most sensitive attributes. However, it is worth noting that a greedy single attribute perturbation can be computationally expensive for a large attribute space (e.g., 15 attributes). It is very time-consuming to adversarially traverse a single attribute over the dataset and repeat 15 times (i.e., repeat for each attribute) in a grid-search manner. Jointly optimizing all attributes is more time-effective and comprehensive

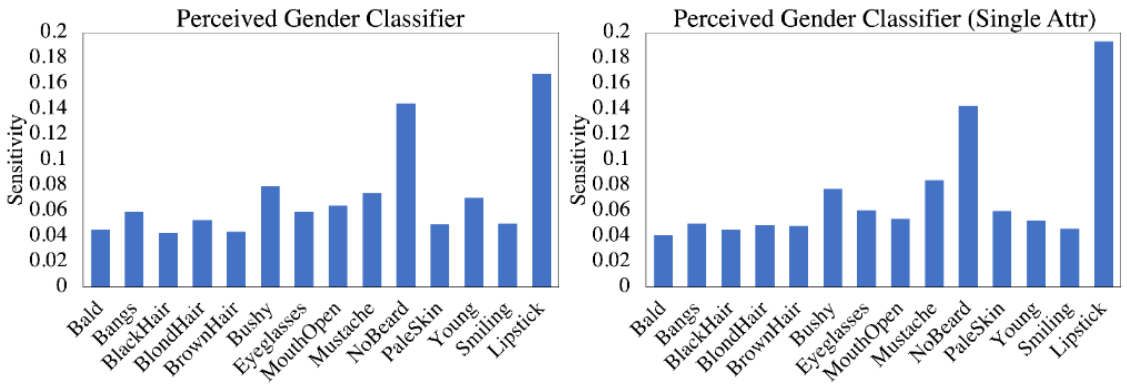


Figure 2.7: Histogram for attribute sensitivities (under multi- and single-attribute optimization) for the perceived-gender classifier.

(i.e., exploring a continuous space across all attributes) as the histogram on the left.

2.4.3 Attack Effectiveness

This section compares SIA to popular gradient-based adversarial attacks in a white-box setting for the attribute classifiers. Then an ablation study is conducted to demonstrate the effectiveness of various components in SIA, including the use of attribute and image perturbations.

Attack success rate

SIA constrains the perturbation bounds of attribute space ($\mathbf{a}$) and image space ($\mathbf{x}$) separately. The attributes that do not overlap with the target model range between $[0, 1]$ with a step size $\eta = \frac{0.25}{255}$. The attribute that is equivalent to the target classifier is constrained to be a small constant depending on the attribute being classified. We iteratively perturbed the input image bounded by $\epsilon = \frac{1.5}{255}$ with $\eta = \frac{0.25}{255}$. The number of steps for both the attribute space and image space will be 200. The evaluated subset in CelebA corresponds to the first $10,000$ images.

We used the FGSM [11] and PGD [28] under l_∞ norm of different perturbation bounds as baseline methods. For PGD, the iteration step size $\eta = \frac{0.25}{255}$ with 200 steps in total. For FGSM, the attack will iterate once within a bounded perturbation. In the adversarial training experiment, we additionally compared with SPT [17] using the same attribute space as SIA. However, we did not compare with [23] since

ϵ	Classifier	FGSM	PGD	SIA
1.5/255	Eyeglasses	28.01	49.85	**68.20**
	Perceived Gender	56.31	65.32	**88.19**
2/255	Eyeglasses	42.83	87.79	**94.82**
	Perceived Gender	75.27	87.19	**92.44**
4/255	Eyeglasses	78.90	99.94	**99.99**
	Perceived Gender	97.33	97.41	**98.53**

Table 2.2: Success rate (%) for different adversarial attack methods with different perturbation bounds.

their method samples adversarial examples from StyleGAN, which does not support attacking existing images. We also did not compare with [21] because their method requires training a separate model on StyleGAN's original training data for each target model.

Table 2.2 shows the attack success rates for different perturbation-based attacks on multiple target classifiers. Notably, we can see that SIA achieves performance comparable to traditional attacks with smaller perturbations.

2.4.4 Baseline Comparison

We implement SPT [17], CW [3], and Face-Manifold (FM) [23] attacks and evaluate the attack success rates (ASR) on our facial eyeglass classifier. The classifier is trained with the setup mentioned in section 4.1. For all listed setups (unless otherwise stated), the images for evaluation is the first $2,000$ images from CelebA test set.

In SPT attack, we use the attribute space consisting the same 15 attributes as SIA. The optimizer is RMSProp with two learning rates $\eta = 0.25/255$ and $\eta = 4/255$. Table 2.3 shows the ASR with different attack iterations. Under the same setup of 200 iterations with $\eta = 0.25/255$, our SIA *attribute-only* ASR (in section 4.3 ablation study table) is 32.67% which outperforms SPT. This shows that the use of signed gradient to update the attribute space, which stabilizes the optimization, can improve the attack effectiveness.

In CW attack, we fix the attack iteration same as SIA's 200 and evaluate the ASR under different box-constrain parameters. Table 2.4 shows that by relaxing the box-constrain, the ASR can hit to 52.6% which is higher than our PGD baseline of

Iterations	2	5	10	15	20	50	100	150	200
$\eta = \frac{0.25}{255}$	0.3%	0.3%	0.4%	0.7%	0.9%	3.1%	12.3%	19.4%	**22.9%**
$\eta = \frac{4}{255}$	1.2%	5.0%	14.6%	21.7%	25.4%	28.3%	29.6%	**30.3%**	29.7%

Table 2.3: ASR for SPT attack with different step size η on eyeglasses classifier

Box Contrain	0.10	0.25	0.50	0.75	1.0	1.5	2.0	5.0	10.0
ASR	9.1%	16.8%	25.0%	29.4%	31.9%	35.1%	37.2%	45.4%	**52.6%**

Table 2.4: ASR for CW attack with different box-constrains setup on eyeglasses classifier

49.85%. The default setting of SIA where image and attribute spaces are co-updated has an ASR of 68.20% which are much more effective than CW and PGD. Note that during SIA's adversarial optimization, we can obtain attribute sensitivity which provides intuitive model interpretation to users. Pure image space attacks cannot support such features.

In FM attack, we follow the setup as specified in the original section to make sure that we can re-implement the high ASR reported in their section. We sample $2,000$ images from the style space to experiment on different settings of ϵ_1 (style step size) and ϵ_2 (noise step size). Table 2.5 shows the ASR of different ϵ_1 and ϵ_2 settings. We find out that noise vectors have a superior effect on flipping the prediction of our eyeglasses ResNet classifier. With increasing the strength of injected noises during the generation, the image quality will significantly decrease. Note that SIA and PGD can also achieve similar ASR (99.99% and 99.94% correspondingly) by relaxing the image space constraint.

	$\epsilon_2 = 0$ (no noise)	$\epsilon_2 = 0.01$	$\epsilon_2 = 0.02$	$\epsilon_2 = 0.03$	$\epsilon_2 = 0.04$	$\epsilon_2 = 0.05$
$\epsilon_1 = 0.004$	2.5%	70.8%	96.6%	99.7%	100.0%	99.9%
$\epsilon_1 = 0.01$	2.2%	19.3%	55.9%	78.8%	91.9%	97.3%
$\epsilon_1 = 0.05$	0.3%	1.9%	3.95%	8.2%	19.3%	20.5%
$\epsilon_1 = 0.1$	0.2%	0.8%	1.3%	2.9%	7.0%	12.1%

Table 2.5: ASR for FM attack on eyeglasses classifier

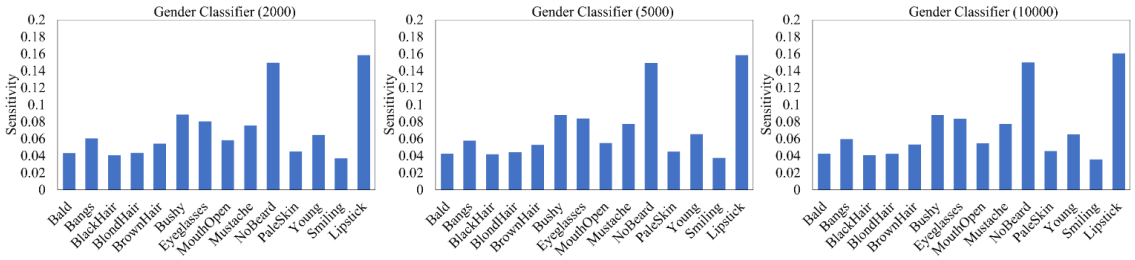


Figure 2.8: Histogram visualization of attribute sensitivities (under different SIA data amount) for the gender classifier.

	Eyeglasses	Goatee	Age	Sideburns
OR	0.02	3.49	0.15	3.32
I	1.83	13.58	6.10	12.51
I + PA	18.25	26.54	30.16	22.46
A	32.67	83.59	**90.90**	67.97
A + PI	44.26	79.82	85.98	70.65
Full	**68.20**	**90.51**	90.38	**87.08**

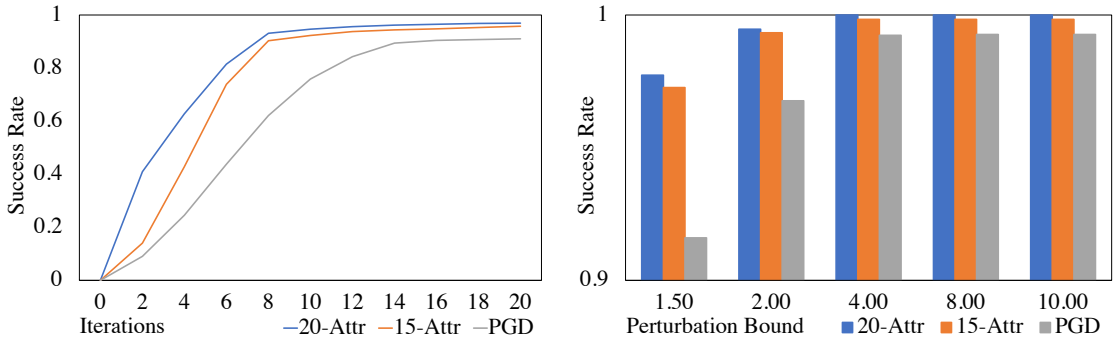
Table 2.6: Ablation study on the success rate (%) of attacks.

Ablation for image v.s. attribute space

This experiment analyzes the attack success rate when attacking the image and/or the attribute space (see Table 2.6). Original reconstruction refers to the images reconstructed by $\mathcal{G}_\phi$ without any perturbations. I/A refers to only updating the image/attribute space during the attack. PI/PA refers to partially updating the image/attribute space in the first 20 iterations of the total 200 iterations. Note that the Attr-space setting is different from SPT [17] since SIA uses sign linearization to constrain the gradient updates to stabilize the attack. As expected, the attack effectiveness is much higher regardless of using attribute space alone or in combination.

Extending attribute space

We experimented with an alternative attribute space of 20 attributes for $\mathcal{G}_\phi$. We removed Black_Hair, Brown_Hair, Bushy_Eyebrows, Eyeglasses, Male(perceived), No_Beard, Young(perceived), Wearing_Lipstick which are either attribute of target classifier or attributes with overlapped concepts. Then we added Narrow_Eyes, Oval_Face, Pale_Skin, Pointy_Nose, Receding_Hairline, Rosy_Cheeks, Sideburns,



(a) Ablation for attack iterations

(b) Ablation for attack bound

Figure 2.9: The effect of different attribute spaces used in SIA. We compare 15 and 20 attributes and show that a larger space of attributes leads to faster attack convergence (left) and a higher success rate with various bounds (right).

Straight_Hair, Big_Lips, Big_Nose, Chubby, Goatee, Heavy_Makeup, High_Cheekbones. Compared with the attribute space used in our main experiment, this alternative attribute space covers more semantic variations in facial data. Fig. 2.9(a) shows the success rate for the attractive classifier. PGD refers to the implemented PGD attack [28]. The larger the attribute space, the higher the success rate, and the attack converges in fewer iterations. This is not surprising because the larger semantic space helps $\mathcal{G}_\phi$ to search the combination of adversarial attributes more effectively. Fig. 2.9(b) shows that with the same perturbation bound setting, the extended $\mathcal{G}_\phi$ will give a stronger attack on the target classifier.

2.4.5 Adversarial Training

In this experiment, we evaluated the effectiveness of SIA to improve adversarial robustness. We adopted the setting such that the target model is fine-tuned with adversarial examples for one epoch. Table 2.7 shows how SIA can be used effectively for re-fitting adversarial examples generated by Algorithm 1. SIA-Adv, PGD-Adv, SPT-Adv are eyeglasses classifiers adversarially trained with $30,000$ adversarial examples generated by the corresponding attack method from the first $30,000$ images of CelebA. The perturbation bound is $\epsilon = 1.5/255$. Non-adversarial training means the regular classifier trained in Section 2.4.2. All models are evaluated on the first $10,000$ images from the CelebA test set that the models have never seen before.

	Non-Adv	PGD-Adv [28]	SPT-Adv [17]	SIA-Adv
Clean Test Set	99.63	99.54	99.51	99.52
FGSM (1.5/255)	73.97	86.55	77.76	**94.01**
PGD (1.5/255)	50.63	81.98	16.52	**86.90**
SIA (1.5/255)	12.59	27.90	**74.01**	67.07
FGSM (4/255)	22.47	22.09	41.78	**45.51**
SIA (4/255)	4.56	10.85	10.55	**12.57**

Table 2.7: Adversarial training. The reported numbers represent the accuracy (%) for adversaries.

Results show that the robustness of SIA adversarial training is transferable to other attack methods, but not vice versa (i.e., see how the column SIA-Adv works well across all the attacks). This is because our adversarial example constructs both conceptual shifts in the semantic space and noise shift in the image space, which introduces richer information during the adversarial training compared to traditional perturbation attacks. Fig. 2.10 shows the visual comparison of SIA and SPT adversaries on the eyeglasses classifier. SPT generates less fine-controlled semantic changes because updating only the attribute space results in large changes across many attributes. More visual comparisons of different baselines for adversarial training are reported in Appendix C.

Standard deviation of attribute robustness

We established a measure named Standard Deviation of Attribute Robustness (SDAR) to understand the final variance of our model across attributes. For a given classifier f_θ, SIA generates the sensitivity histogram based on the attribute perturbation vector $\mathbf{s}$ of length L. The SDAR metric $\sigma_{\boldsymbol{s}}$ is defined as the standard deviation of the sensitivity values $\sigma_{\boldsymbol{s}} = \sqrt{\frac{1}{L}\sum_{i=1}^{L}(s_i - \bar{\mathbf{s}})^2}$.

Ideally, an unbiased model should have equal sensitivity across all attributes, hence a decrease in the standard deviation will indicate that the model is less biased. To validate the method, we calculated SDARs after evaluating different models from adversarial training. The test data was the first $10,000$ images of CelebA test set. We evaluated the SDAR metric under two bounds (ϵ) of SIA. Table 2.8 shows the results.

	Non-Adv	PGD-Adv [28]	SPT-Adv [17]	SIA-Adv (Ours)
$\epsilon = \frac{1.5}{255}$	0.1145	0.0917	0.0852	**0.0781**
$\epsilon = \frac{4}{255}$	0.1419	0.1270	0.1253	**0.1056**

Table 2.8: SDAR metric on different adversarially-trained models from Section 4.4. A lower value indicates a less biased model.

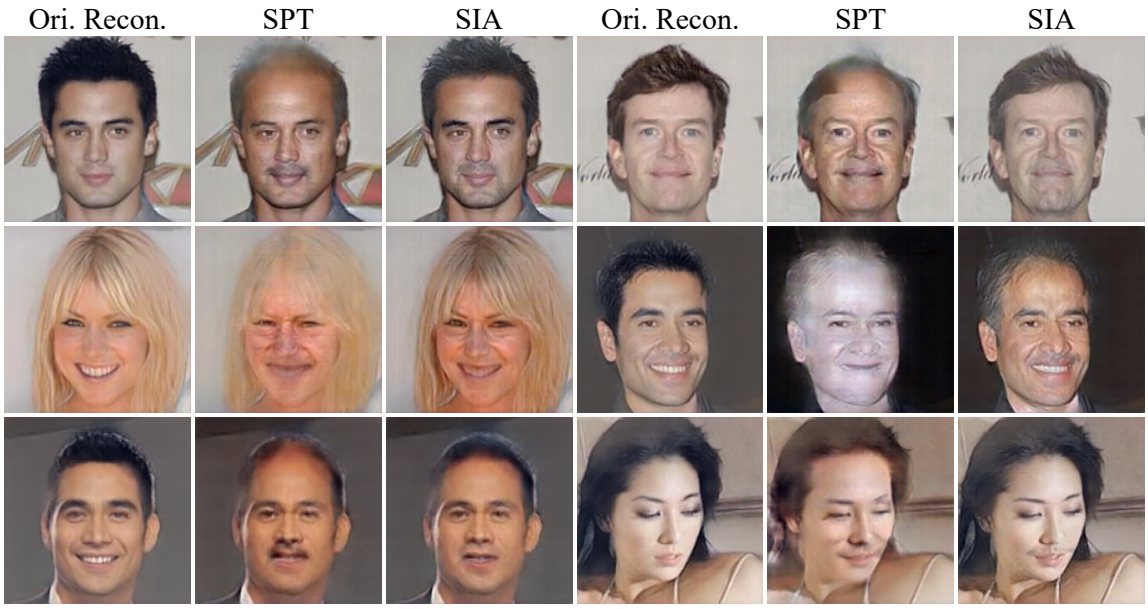


Figure 2.10: Demonstration of SIA and SPT [17] adversarial examples on the eyeglasses classifier. Results show that SPT generates unrealistic images, while SIA generates realistic images with small but semantic modifications of the original image.

We can see that the non-adv classifiers will have larger σ and the adversarially-trained models have smaller σ since the model generalization is improved by the adversarial training process. By optimizing both the attribute and the image space, SIA-Adv better generalizes over attributes than regular classifiers.

2.4.6 Robustness to Imbalanced Datasets

This section reports experiments to evaluate the robustness of SIA in learning from imbalanced datasets. In these situations, it is vital to develop algorithms that are not biased toward the majority class. While data augmentation and re-weighting are

commonly used techniques, we show how SIA provides an alternative that generates semantically meaningful augmentation with high visual quality.

We trained two attribute classifiers, for eyeglasses and bangs, using the ResNet50 architecture. We generated a synthetically unbalanced dataset to produce a controlled imbalanced environment. For training, we randomly sampled $30,000$ images from CelebA training set such that 1% are positive and 99% are negative. We trained the classifiers from random initialization. For testing, we use balanced test data including random $2,500$ positive-label images and $2,500$ negative-label images from CelebA test set.

Table 2.9 shows the precision, recall, and accuracy for several imbalanced learning strategies. We compared SIA to five data augmentation and re-balancing approaches. The Non-Adv attribute classifiers are trained on the synthetic data with 1:99 CelebA training set. The CutMix [51] baselines are augmented with action probability $p = 0.5$ and learning rate $\alpha = 0.001$. We followed PyTorch's implementation on AutoAugment [6]. We also included two commonly used baselines to deal with imbalanced data: Reweighting and Resampling. Reweighting means upweighting the under-represented samples based on the proportion of class samples. Resampling means duplicating the under-represented samples until different classes have the same number of samples. SIA refers to classifiers augmented by randomly sampling $30,000$ our adversarial images. SIA + Reweight is the scheme where the reweighting is performed on our SIA-augmented dataset. Results show that SIA can effectively be used to augment imbalanced datasets, outperforming other widely used augmentation methods. One possible reason is that SIA generates semantically meaningful augmentations, different from CutMix and AutoAugment. Finally, we conduct a similar experiment with pre-trained classifiers in Appendix D. We show that the difference in accuracy between the methods narrows down considerably if we pre-train the classifiers. This is not surprising, since pre-training with sufficient data provides robust features that are less prone to imbalance.

	Strategy	Prec. ↑	Recall ↑	Acc. (%) ↑
Eyeglasses	Non-adv classifier	0.9985	0.8052	90.20
	Reweighting	**0.9995**	0.8368	91.82
	Resample	0.9984	0.7700	88.44
	CutMix	0.9963	0.3236	66.12
	AutoAugment	0.9975	0.8004	89.92
	SIA-Adv (ours)	0.9991	**0.8864**	**94.28**
	SIA-Adv + Reweight (ours)	0.9991	0.8856	94.24
Bangs	Non-adv classifier	0.9847	0.2576	62.68
	Reweighting	0.9912	0.2708	63.42
	Resample	0.9935	0.1840	59.14
	CutMix	**1.0000**	0.0000	50.00
	AutoAugment	0.9701	0.0260	51.26
	SIA-Adv (ours)	0.9791	0.4116	70.14
	SIA-Adv + Reweight (ours)	0.9854	**0.5960**	**79.36**

Table 2.9: Comparison of different strategies for learning from imbalanced datasets. See text.

2.4.7 Image Synthesis Analysis

Visual Comparison of Adversarial Examples

Fig. 2.11 shows more visual comparisons of the adversarial examples generated by different methods. As we can see, SIA adds perturbation in the image space and the attribute space, generating photo-realistic fine-grained adversarial examples. Perturbing in both image space and attribute space produces finer visual adversarial images. In fact, SIA 's pixel-level perturbation helps to avoid exaggerated semantic variation that makes the image generation collapse.

AttGAN's Reconstruction and Semantic Editing

AttGAN ($\mathcal{G}_\phi$) is capable of editing both fine-level semantics (e.g., beard) and complex concepts (e.g., age). The reconstruction loss during the training of $\mathcal{G}_\phi$ guarantees the preservation of facial details. As stated in [14], the use of shortcut layers [36] improves the quality of image translation. During the SPT and SIA attack, we constrained all mutated attributes in the range of [0,1] to make sure that the transformed attribute vector for $\mathcal{G}_\phi$ is valid. The style intensity hyper-parameter is set to 1, and the number of encoder layers and decoder layers are both 5.

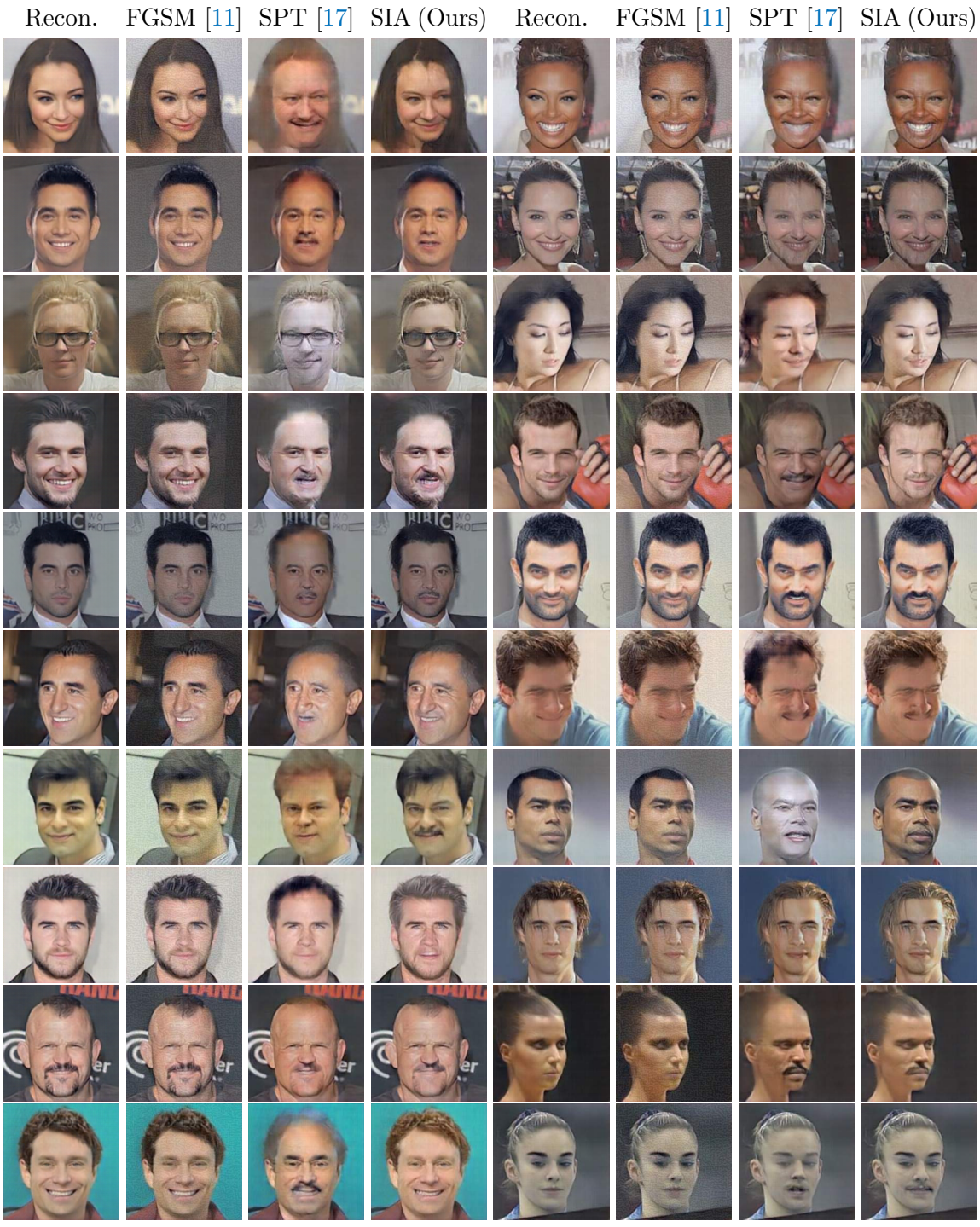


Figure 2.11: Adversarial examples by FGSM [11], SPT [17], and SIA (Ours) on the eyeglasses classifier.

2.5 Discussion and Future Work

This section introduced SIA, an attribute-assisted adversarial method with applications in model diagnosis, improving target model robustness, and increasing the success of visual attacks. A major appeal of our technique is the capacity of analyzing a deep learning model without a carefully designed test set. SIA reveals the dependencies between attributes and model outputs, which helps interpret the biases learned by models during prediction. We hope our results pave the way for new tools to analyze models and inspire future work on mitigating such biases.

While we showed the benefits of our technique in two computer vision problems, our approach is applicable to any end-to-end differentiable target deep learning model. It is unclear how to extend this approach to non-differentiable ML models, and more research needs to be done. Our method works with white-box attacks since our primary motivation is to diagnose a known model. More research needs to be done to address black-box attacks. Furthermore, we have illustrated the power of SIA only in the context of faces, but our method can extend to generative models that have been trained with other attributes of interest and can be applied to other visual domains.

Chapter 3

Zero-Shot Model Diagnosis

When it comes to deploying deep vision models, the behavior of these systems must be explicable to ensure confidence in their reliability and fairness. A common approach to evaluate deep learning models is to build a labeled test set with attributes of interest and assess how well it performs. However, creating a balanced test set (i.e., one that is uniformly sampled over all the important traits) is often time-consuming, expensive, and prone to mistakes. The question we try to address is: can we evaluate the sensitivity of deep learning models to arbitrary visual attributes **without an annotated test set**?

This chapter argues the case that **Z**er**o**-sh**o**t **M**odel Diagnosis (ZOOM) is possible without the need for a test set nor labeling. To avoid the need for test sets, our system relies on a generative model and CLIP. The key idea is enabling the user to select a set of prompts (relevant to the problem) and our system will automatically search for semantic counterfactual images (i.e., synthesized images that flip the prediction in the case of a binary classifier) using the generative model. We evaluate several visual tasks (classification, key-point detection, and segmentation) in multiple visual domains to demonstrate the viability of our methodology. Extensive experiments demonstrate that our method is capable of producing counterfactual images and offering sensitivity analysis for model diagnosis without the need for a test set.

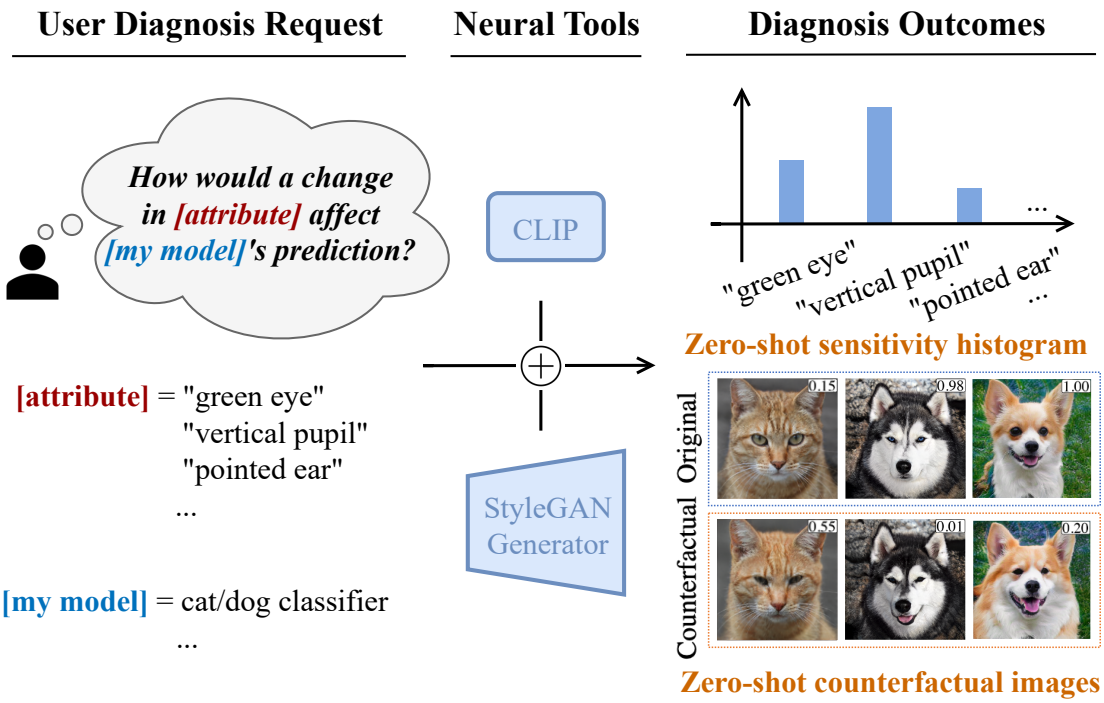


Figure 3.1: Given a differentiable deep learning model (e.g., a cat/dog classifier) and user-defined text attributes, how can we determine the model's sensitivity to specific attributes without using labeled test data? Our system generates counterfactual images (bottom right) based on the textual directions provided by the user, while also computing the sensitivity histogram (top right).

3.1 Introduction

Deep learning models inherit data biases, which can be accentuated or downplayed depending on the model's architecture and optimization strategy. Deploying a computer vision deep learning model requires extensive testing and evaluation, with a particular focus on features with potentially dire social consequences (e.g., non-uniform behavior across gender or ethnicity). Given the importance of the problem, it is common to collect and label large-scale datasets to evaluate the behavior of these models across attributes of interest. Unfortunately, collecting these test datasets is extremely time-consuming, error-prone, and expensive. Moreover, a balanced dataset, that is uniformly distributed across all attributes of interest, is also typically impractical to acquire due to its combinatorial nature. Even with careful metric analysis in this test set, no robustness nor fairness can be guaranteed since there

can be a mismatch between the real and test distributions [35]. This research will explore model diagnosis without relying on a test set in an effort to *democratize* model diagnosis and lower the associated cost.

Counterfactual explainability as a means of model diagnosis is drawing the community's attention [30, 12]. Counterfactual images visualize the sensitive factors of an input image that can influence the model's outputs. In other words, counterfactuals answer the question: *"How can we modify the input image* $\mathbf{x}$ *(while fixing the ground truth) so that the model prediction would diverge from* $\mathbf{y}$ *to* $\hat{\mathbf{y}}$*?"*. The parameterization of such counterfactuals will provide insights into identifying key factors of where the model fails. Unlike existing image-space adversary techniques [11, 28], counterfactuals provide semantic perturbations that are interpretable by humans. However, existing counterfactual studies require the user to either collect uniform test sets [18], annotate discovered bias [24], or train a model-specific explanation every time the user wants to diagnose a new model [21].

On the other hand, recent advances in Contrastive Language-Image Pretraining (CLIP) [34] can help to overcome the above challenges. CLIP enables text-driven applications that map user text representations to visual manifolds for downstream tasks such as avatar generation [15], motion generation [53] or neural rendering [32, 42]. In the domain of image synthesis, StyleCLIP [31] reveals that text-conditioned optimization in the StyleGAN [20] latent space can decompose latent directions for image editing, allowing for the mutation of a specific attribute without disturbing others. With such capability, users can freely edit semantic attributes conditioned on text inputs. This section further explores its use in the scope of model diagnosis.

The central concept of the this section is depicted in Fig. 3.1. Consider a user interested in evaluating which factors contribute to the lack of robustness in a cat/dog classifier (target model). By selecting a list of keyword attributes, the user is able to (1) see counterfactual images where semantic variations flip the target model predictions (see the classifier score in the top-right corner of the counterfactual images) and (2) quantify the sensitivity of each attribute for the target model (see sensitivity histogram on the top). Instead of using a test set, we propose using a StyleGAN generator as the picture engine for sampling counterfactual images. CLIP transforms user's text input, and enables model diagnosis in an open-vocabulary setting. This is a major advantage since there is no need for collecting and annotating

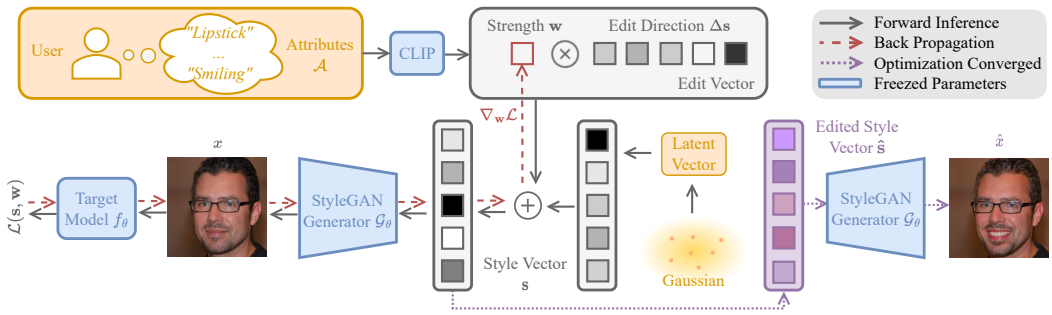


Figure 3.2: The ZOOM framework. Black solid lines stand for forward passes, red dashed lines stand for backpropagation, and purple dashed lines stand for inference after the optimization converges. The user inputs single or multiple attributes, and we map them into edit directions with the method in Sec. 3.3.2. Then we assign to each edit direction (attribute) a weight, which represents how much we are adding/removing this attribute. We iteratively perform adversarial learning on the attribute space to maximize the counterfactual effectiveness.

images and minimal user expert knowledge. In addition, we are not tied to a particular annotation from datasets (e.g., specific attributes in CelebA [25]).

To summarize, our proposed work offers three major improvements over earlier efforts:

- The user requires neither a labeled, balanced test dataset, and minimal expert knowledge in order to evaluate where a model fails (i.e., model diagnosis). In addition, the method provides a sensitivity histogram across the attributes of interest.
- When a different target model or a new user-defined attribute space is introduced, it is not necessary to re-train our system, allowing for practical use.
- The target model fine-tuned with counterfactual images not only slightly improves the classification performance, but also greatly increases the distributional robustness against counterfactual images.

3.2 Related Work

This section reviews prior work on attribute editing with generative models and recent efforts on model diagnosis.

3.2.1 Attribute Editing with Generative Models

With recent progress in generative models, GANs supports high-quality image synthesis, as well as semantic attributes editing [48]. [14, 4] edit the images by perturbing the intermediate latent space encoded from the original images. These methods rely on images to be encoded to latent vectors to perform attribute editing. On the contrary, StyleGAN [20] can produce images by sampling the latent space. Many works have explored ways to edit attributes in the latent space of StyleGAN, either by relying on image annotations [39] or in an unsupervised manner [40, 16]. StyleSpace [47] further disentangles the latent space of StyleGAN and can perform specific attribute edits by disentangled style vectors. Based upon StyleSpace, StyleCLIP [31] builds the connection between the CLIP language space and StyleGAN latent space to enable arbitrary edits specified by the text. Our work adopts this concept for fine-grained attribute editing.

3.2.2 Model Diagnosis

To the best of our knowledge, model diagnosis without a test set is a relatively unexplored problem. In the adversarial learning literature, it is common to find methods that show how image-space perturbations [11, 28] flip the model prediction; however, such perturbations lack visual interpretability. [49] pioneers in synthesizing adversaries by GANs. More recently, [17, 38, 33] propose generative methods to synthesize semantically perturbed images to visualize where the target model fails. However, their attribute editing is limited within the dataset's annotated labels. Instead, our framework allows users to easily customize their own attribute space, in which we visualize and quantify the biased factors that affect the model prediction. On the bias detection track, [21] co-trains a model-specific StyleGAN with each target model, and requires human annotators to name attribute coordinates in the Stylespace. [24, 8, 22] synthesize counterfactual images by either optimally traversing the latent space or learning an attribute hyperplane, after which the user will inspect the represented bias. Unlike previous work, we propose to diagnose a deep learning model without any model-specific re-training, new test sets, or manual annotations/inspections.

3.3 Method

This section firstly describes our method to generate counterfactual images guided by CLIP in a zero-shot manner. We then introduce how we perform the sensitivity analysis across attributes of interest. Fig. 3.2 shows the overview of our framework.

3.3.1 Notation and Problem Definition

Let f_θ, parameterized by θ, be the target model that we want to diagnose. In this section, f_θ denotes two types of deep nets: binary attribute classifiers and face keypoint detectors. Note that our approach is extendable to any end-to-end differentiable target deep models. Let $\mathcal{G}_\phi$, parameterized by ϕ, be the style generator that synthesizes images by $\mathbf{x} = \mathcal{G}_\phi(\mathbf{s})$ where $\mathbf{s}$ is the style vector in Style Space $\mathcal{S}$ [47]. We denote a counterfactual image as $\hat{\mathbf{x}}$, which is a synthesized image that misleads the target model f_θ, and denote the original reference image as $\mathbf{x}$. a is defined as a single user input text-based attribute, with its domain $\mathcal{A} = \{a_i\}_{i=1}^N$ for N input attributes. $\hat{\mathbf{x}}$ and $\mathbf{x}$ differs only along attribute directions $\mathcal{A}$. Given a set of $\{f_\theta, \mathcal{G}_\phi, \mathcal{A}\}$, our goal is to perform counterfactual-based diagnosis to interpret where the model fails without manually collecting nor labeling any test set. Unlike traditional approaches of image-space noises which lack explainability to users, our method adversarially searches the counterfactual in the user-designed semantic space. To this end, our diagnosis will have three outputs, namely counterfactual images (Sec. 3.3.3), sensitivity histograms (Sec. 3.3.4), and distributionally robust models (Sec. 3.3.5).

3.3.2 Extracting Edit Directions

This section examines the terminologies, method, and modification we adopt in ZOOM to extract suitable global directions for attribute editing. Since CLIP has shown strong capability in disentangling visual representation [29], we incorporate style channel relevance from StyleCLIP [31] to find edit directions for each attribute.

Given the user's input strings of attributes, we want to find an image manipulation direction $\Delta\mathbf{s}$ for any $\mathbf{s} \sim \mathcal{S}$, such that the generated image $\mathcal{G}_\phi(\mathbf{s} + \Delta\mathbf{s})$ *only* varies in the input attributes. Recall that CLIP maps strings into a text embedding $\mathbf{t} \in \mathcal{T}$,

the text embedding space. For a string attribute description a and a neutral prefix p, we obtain the CLIP text embedding difference $\Delta\mathbf{t}$ by:

$$\Delta\mathbf{t} = \text{CLIP}_{\text{text}}(p \oplus a) - \text{CLIP}_{\text{text}}(p) \tag{3.1}$$

where $\oplus$ is the string concatenation operator. To take 'Eyeglasses' as an example, we can get $\Delta\mathbf{t} = \text{CLIP}_{\text{text}}(\text{'a face with Eyeglasses'}) - \text{CLIP}_{\text{text}}(\text{'a face'})$.

To get the edit direction, $\Delta\mathbf{s}$, we need to utilize a style relevance mapper $\mathbf{M} \in \mathbb{R}^{c_\mathcal{S} \times c_\mathcal{T}}$ to map between the CLIP text embedding vectors of length $c_\mathcal{T}$ and the Style space vector of length $c_\mathcal{S}$. StyleCLIP optimizes $\mathbf{M}$ by iteratively searching meaningful style channels: mutating each channel in $\mathcal{S}$ and encoding the mutated images by CLIP to assess whether there is a significant change in $\mathcal{T}$ space. To prevent undesired edits that are irrelevant to the user prompt, the edit direction $\Delta\mathbf{s}$ will filter out channels that the style value change is insignificant:

$$\Delta\mathbf{s} = (\mathbf{M} \cdot \Delta\mathbf{t}) \odot \mathbb{1}((\mathbf{M} \cdot \Delta\mathbf{t}) > \lambda), \tag{3.2}$$

where λ is the hyper-parameter for the threshold value. $\mathbb{1}(\cdot)$ is the indicator function, and $\odot$ is the element-wise product operator. Since the success of attribute editing by the extracted edit directions will be the key to our approach, Appendix A will show the capability of CLIP by visualizing the global edit direction on multiple sampled images, conducting the user study, and analyzing the effect of λ.

3.3.3 Style Counterfactual Synthesis

Identifying semantic counterfactuals necessitates a manageable parametrization of the semantic space for effective exploration. For ease of notation, we denote $(\Delta\mathbf{s})_i$ as the global edit direction for i^{th} attribute $a_i \in \mathcal{A}$ from the user input. After these N attributes are provided and the edit directions are calculated, we initialize the control vectors $\mathbf{w}$ of length N where the i^{th} element w_i controls the strength of the i^{th} edit direction. Our counterfactual edit will be a linear combination of normalized edit directions: $\mathbf{s}_{edit} = \sum_{i=1}^{N} w_i \frac{(\Delta\mathbf{s})_i}{||(\Delta\mathbf{s})_i||}$.

The black arrows in Fig. 3.2 show the forward inference to synthesize counterfactual images. Given the parametrization of attribute editing strengths and the final loss

value, our framework searches for counterfactual examples in the optimizable edit weight space. The original sampled image is $\mathbf{x} = G_\phi(\mathbf{s})$, and the counterfactual image is

$$\hat{\mathbf{x}} = G_\phi(\mathbf{s} + \mathbf{s}_{edit}) = G_\phi\left(\mathbf{s} + \sum_{i=1}^{N} w_i \frac{(\Delta\mathbf{s})_i}{||(\Delta\mathbf{s})_i||}\right), \tag{3.3}$$

which is obtained by minimizing the following loss, $\mathcal{L}$, that is the weighted sum of three terms:

$$\mathcal{L}(\mathbf{s}, \mathbf{w}) = \alpha\mathcal{L}_{target}(\hat{\mathbf{x}}) + \beta\mathcal{L}_{struct}(\hat{\mathbf{x}}) + \gamma\mathcal{L}_{attr}(\hat{\mathbf{x}}). \tag{3.4}$$

We back-propagate to optimize $\mathcal{L}$ w.r.t the weights of the edit directions $\mathbf{w}$, shown as the red pipeline in Fig. 3.2.

The targeted adversarial loss $\mathcal{L}_{target}$ for binary attribute classifiers minimizes the distance between the current model prediction $f_\theta(\hat{\mathbf{x}})$ with the flip of original prediction $\hat{p}_{cls} = 1 - f_\theta(\mathbf{x})$. In the case of an eyeglass classifier on a person wearing eyeglasses, $\mathcal{L}_{target}$ will guide the optimization to search $\mathbf{w}$ such that the model predicts no eyeglasses. For a keypoint detector, the adversarial loss will minimize the distance between the model keypoint prediction with a set of *random* points $\hat{p}_{kp} \sim \mathcal{N}$:

$$\text{(binary classifier) } \mathcal{L}_{target}(\hat{\mathbf{x}}) = L_{CE}(f_\theta(\hat{\mathbf{x}}), \hat{p}_{cls}), \tag{3.5}$$

$$\text{(keypoint detector) } \mathcal{L}_{target}(\hat{\mathbf{x}}) = L_{MSE}(f_\theta(\hat{\mathbf{x}}), \hat{p}_{kp}). \tag{3.6}$$

If we only optimize $\mathcal{L}_{target}$ w.r.t the global edit directions, it is possible that the method will not preserve image statistics of the original image and can include the particular attribute that we are diagnosing. To constrain the optimization, we added a structural loss $\mathcal{L}_{struct}$ and an attribute consistency loss $\mathcal{L}_{attr}$ to avoid generation collapse. $\mathcal{L}_{struct}$ [44] aims to preserve global image statistics of the original image $\mathbf{x}$ including image contrasts, background, or shape identity during the adversarial editing. While $\mathcal{L}_{attr}$ enforces that the target attribute (perceived ground truth) be consistent on the style edits. For example, when diagnosing the eyeglasses classifier, ZOOM preserves the original status of eyeglasses and precludes direct eyeglasses

addition/removal.

$$\mathcal{L}_{struct}(\hat{\mathbf{x}}) = L_{SSIM}(\hat{\mathbf{x}}, \mathbf{x}) \tag{3.7}$$

$$\mathcal{L}_{attr}(\hat{\mathbf{x}}) = L_{CE}(\text{CLIP}(\hat{\mathbf{x}}), \text{CLIP}(\mathbf{x})) \tag{3.8}$$

Given a pretrained target model f_θ, a domain-specific style generator G_ϕ, and a text-driven attribute space $\mathcal{A}$, our goal is to sample an original style vector $\mathbf{s}$ for each image and search its counterfactual edit strength $\hat{\mathbf{w}}$:

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\text{argmin}}\, \mathcal{L}(\mathbf{s}, \mathbf{w}). \tag{3.9}$$

Unless otherwise stated, we iteratively update $\mathbf{w}$ as:

$$\mathbf{w} = \text{clamp}_{[-\epsilon,\epsilon]}(\mathbf{w} - \eta \nabla_{\mathbf{w}} \mathcal{L}), \tag{3.10}$$

where η is the step size and ϵ is the clamp bound to avoid synthesis collapse caused by exaggerated edit. Note that the maximum counterfactual effectiveness does not indicate the maximum edit strength (i.e., $w_i = \epsilon$), since the attribute edit direction does not necessarily overlap with the target classifier direction. The attribute change is bi-directional, as the w_i can be negative in Eq. 3.3.

3.3.4 Attribute Sensitivity Analysis

Single-attribute counterfactual reflects the sensitivity of target model on the individual attribute. By optimizing independently along the edit direction for a single attribute and averaging the model probability changes over images, our model generates independent sensitivity score h_i for each attribute a_i:

$$h_i = \mathbb{E}_{\mathbf{x}\sim\mathcal{P}(\mathbf{x}),\hat{\mathbf{x}}=\text{ZOOM}(\mathbf{x},a_i)} |f_\theta(\mathbf{x}) - f_\theta(\hat{\mathbf{x}})|. \tag{3.11}$$

The sensitivity score h_i is the probability difference between the original image $\mathbf{x}$ and generated image $\hat{\mathbf{x}}$, at the most counterfactual point when changing attribute a_i.

We synthesize a number of images from $\mathcal{G}_\phi$, then iteratively compute the sensitivity

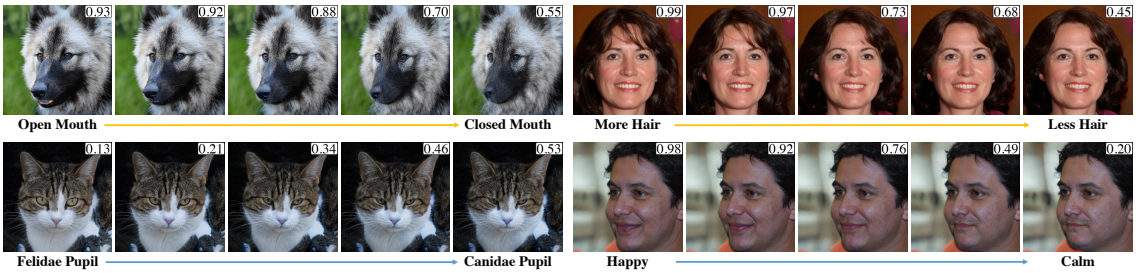


Figure 3.3: Effect of progressively generating counterfactual images on (left) cat/dog classifier (0-Cat / 1-Dog), and (right) perceived age classifier (0-Senior / 1-Young). Model probability prediction during the process is attached at the top right corner.

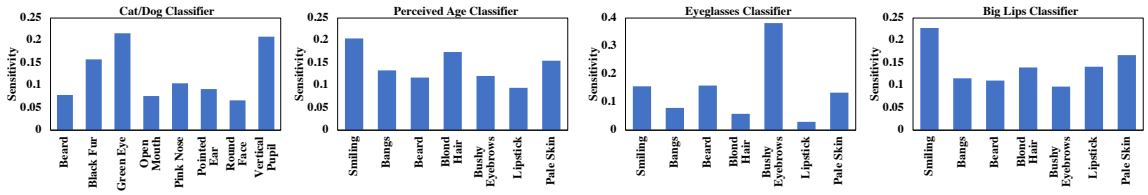
for each given attribute, and finally normalize all sensitivities to draw the histogram as shown in Fig. 3.4. The histogram indicates the sensitivity of the evaluated model f_θ on each of the user-defined attributes. Higher sensitivity of one attribute means that the model is more easily affected by that attribute.

3.3.5 Counterfactual Training

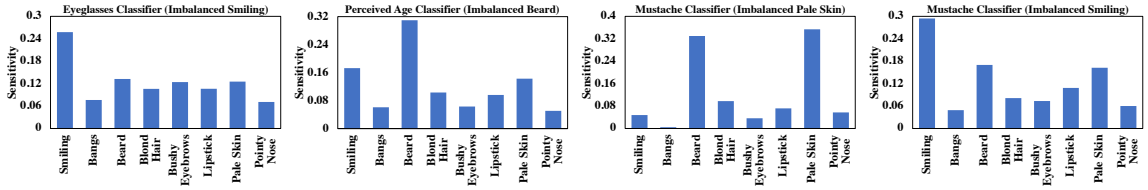
The multi-attribute counterfactual approach visualizes semantic combinations that cause the model to falter, providing valuable insights for enhancing the model's robustness. We naturally adopt the concept of iterative adversarial training [28] to robustify the target model. For each iteration, ZOOM receives the target model parameter and returns a batch of mutated counterfactual images with the model's original predictions as labels. Then the target model will be trained on the counterfactually-augmented images to achieve the robust goal:

$$\theta^* = \underset{\theta}{\operatorname{argmin}}\, \mathbb{E}_{\mathbf{x}\sim\mathcal{P}(\mathbf{x}),\hat{\mathbf{x}}=\mathrm{ZOOM}(\mathbf{x},\mathcal{A})} L_{CE}(f_\theta(\hat{\mathbf{x}}), f_\theta(\mathbf{x})) \tag{3.12}$$

where batches of $\mathbf{x}$ are randomly sampled from the StyleGAN generator $\mathcal{G}_\phi$. In the following, we abbreviate the process as Counterfactual Training (CT). Note that, although not explicitly expressed in Eq. 3.12, the CT process is a min-max game. ZOOM synthesizes counterfactuals to maximize the variation of model prediction (while persevering the perceived ground truth), and the target model is learned with the counterfactual images to minimize the variation.



(a) Model diagnosis histograms generated by ZOOM on four facial attribute classifiers.



(b) Model diagnosis histograms generated by ZOOM on four classifiers trained on manually-crafted imbalance data.

Figure 3.4: Model diagnosis histograms generated by ZOOM. The vertical axis values reflect the attribute sensitivities calculated by averaging the model probability change over all sampled images. The horizontal axis is the attribute space input by user.

3.4 Experimental Results

This section describes the experimental validations on the effectiveness and reliability of ZOOM. First, we describe the model setup in Sec. 3.4.1. Sec. 3.4.2 and Sec. 3.4.3 visualize and validate the model diagnosis results for the single-attribute setting. In Sec. 3.4.5, we show results on synthesized multiple-attribute counterfactual images and apply them to counterfactual training.

3.4.1 Model Setup

Pre-trained models: We used Stylegan2-ADA [19] pretrained on FFHQ [20] and AFHQ [4] as our base generative networks, and the pre-trained CLIP model [34] which is parameterized by ViT-B/32. We followed StyleCLIP [31] setups to compute the channel relevance matrices $\mathcal{M}$.

Target models: Our classifier models are ResNet50 with single fully-connected head initialized by TorchVision[1]. In training the binary classifiers, we use the Adam

[1]https://pytorch.org/blog/how-to-train-state-of-the-art-models-using-torchvision-latest-primitives/

optimizer with learning rate 0.001 and batch size 128. We train binary classifiers for *Eyeglasses, Perceived Gender, Mustache, Perceived Age* attributes on CelebA and for *cat/dog* classification on AFHQ. For the 98-keypoint detectors, we used the HR-Net architecture [43] on WFLW [46].

3.4.2 Visual Model Diagnosis: Single-Attribute

Understanding where deep learning model fails is an essential step towards building trustworthy models. Our proposed work allows us to generate counterfactual images (Sec. 3.3.3) and provide insights on sensitivities of the target model (Sec. 3.3.4). This section visualizes the counterfactual images in which only one attribute is modified at a time.

Fig. 3.3 shows the single-attribute counterfactual images. Interestingly (but not unexpectedly), we can see that reducing the hair length or joyfulness causes the age classifier more likely to label the face to an older person. Note that our approach is extendable to multiple domains, as we change the cat-like pupil to dog-like, the dog-cat classification tends towards the dog. Using the counterfactual images, we can conduct model diagnosis with the method mentioned in Sec. 3.3.4, on which attributes the model is sensitive to. In the histogram generated in model diagnosis, a higher bar means the model is more sensitive toward the corresponding attribute. Fig. 3.8 shows more examples of single-attribute counterfactual images on the Cat/Dog and Perceived Gender classifiers. The output prediction is shown in the top-right corner. It shows that the model prediction is flipped without changing the actual target attribute.

Fig. 3.4a shows the model diagnosis histograms on regularly-trained classifiers. For instance, the cat/dog classifier histogram shows outstanding sensitivity to green eyes and vertical pupil. The analysis is intuitive since these are cat-biased traits rarely observed in dog photos. Moreover, the histogram of eyeglasses classifier shows that the mutation on bushy eyebrows is more influential for flipping the model prediction. It potentially reveals the positional correlation between eyeglasses and bushy eyebrows. The advantage of single-attribute model diagnosis is that the score of each attribute in the histogram are independent from other attributes, enabling unambiguous understanding of exact semantics that make the model fail.

Fig. 3.14 shows more histograms on the classifiers trained on CelebA (top) and the classifiers that are intentionally biased (bottom). The models and datasets are created using the same method described above.

3.4.3 Validation of Visual Model Diagnosis

Evaluating whether our zero-shot sensitivity histograms (Fig. 3.4) explain the true vulnerability is a difficult task, since we do not have access to a sufficiently large and balanced test set fully annotated in an open-vocabulary setting. To verify the performance, we create synthetically imbalanced cases where the model bias is known. We then compare our results with a supervised diagnosis setting [26]. In addition, we will validate the decoupling of the attributes by CLIP.

Creating imbalanced classifiers

In order to evaluate whether our sensitivity histogram is correct, we train classifiers that are highly imbalanced towards a known attribute and see whether ZOOM can detect the sensitivity w.r.t the attribute. For instance, when training the perceived-age classifier (binarized as Young in CelebA), we created a dataset on which the trained classifier is strongly sensitive to Bangs (hair over forehead). The custom dataset is a CelebA training subset that consists of $20,200$ images. More specifically, there are $10,000$ images that have both young people that have bangs, represented as $(1,1)$, and $10,000$ images of people that are not young and have no bangs, represented as $(0,0)$. The remaining combinations of $(1,0)$ and $(0,1)$ have only 100 images. With this imbalanced dataset, bangs is the attribute that dominantly correlates with whether the person is young, and hence the perceived-age classifier would be highly sensitive towards bangs. See Fig. 3.5 (the right histograms) for an illustration of the sensitivity histogram computed by our method for the case of an age classifier with bangs (top) and lipstick (bottom) being imbalanced.

We trained multiple imbalanced classifiers with this methodology, and visualize the model diagnosis histograms of these imbalanced classifiers in Fig. 3.4b. We can observe that the ZOOM histograms successfully detect the synthetically-made bias, which are shown as the highest bars in histograms. See the caption for more information.

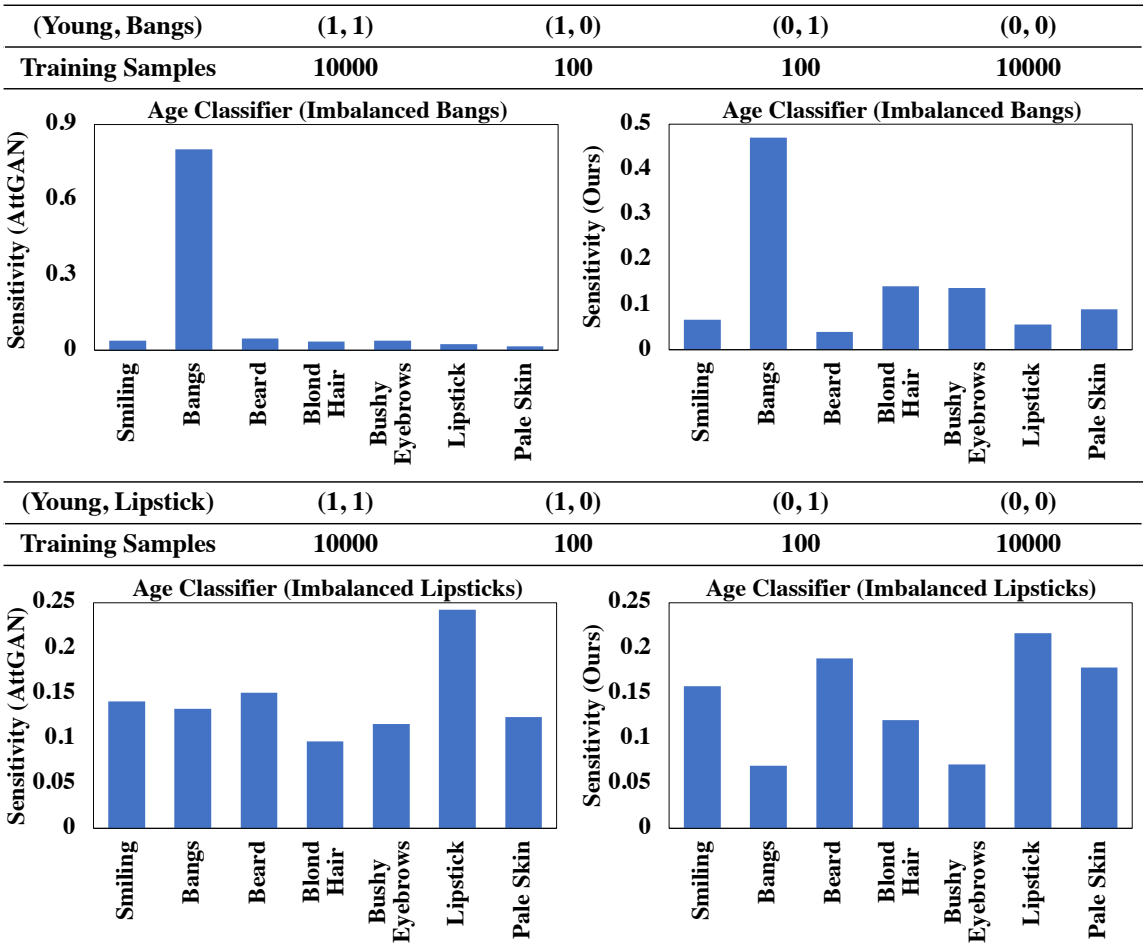
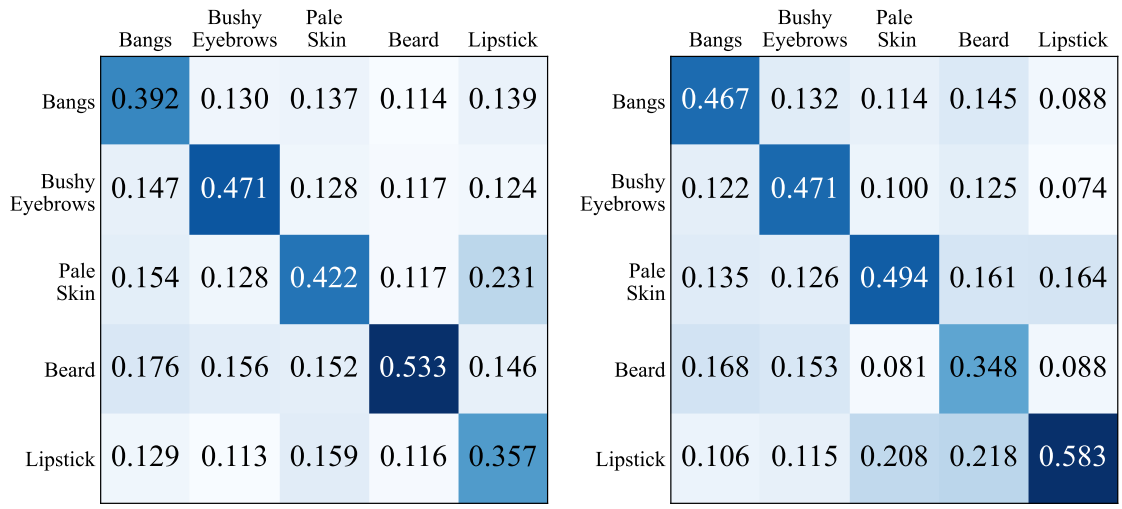


Figure 3.5: The sensitivity of the age classifier is evaluated with ZOOM (right) and AttGAN (left), achieving comparable results.

Comparison with supervised diagnosis

We also validated our histogram by comparing it with the case in which we have access to a generative model that has been explicitly trained to disentangle attributes. We follow the work on [26] and used AttGAN [14] trained on the CelebA training set over 15 attributes[2]. After the training converged, we performed adversarial learning in the attribute space of AttGAN and create a sensitivity histogram using the same approach as Sec. 3.3.4. Fig. 3.5 shows the result of this method on the perceived-age classifier

[2] *Bald, Bangs, Black_Hair, Blond_Hair, Brown_Hair, Bushy_Eyebrows, Eyeglasses, Male, Mouth_Slightly_Open, Mustache, No_Beard, Pale_Skin, Young, Smiling, Wearing_Lipstick.*



(a) Mustache classifier (b) Perceived age classifier

Figure 3.6: Confusion matrix of CLIP score variation (vertical axis) when perturbing attributes (horizontal axis). This shows that attributes in ZOOM are highly decoupled.

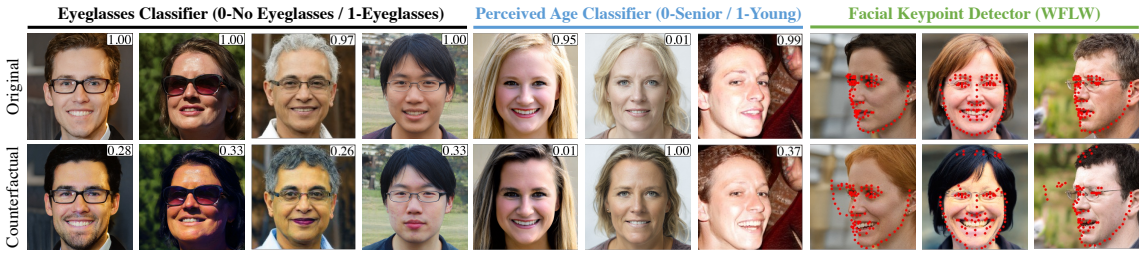


Figure 3.7: Multi-attribute counterfactual in faces. The model probability is documented in the upper right corner of each image.

which is made biased towards bangs. As anticipated, the AttGAN histogram (left) corroborates the histogram derived from our method (right). Interestingly, unlike ZOOM, AttGAN show less sensitivity to remaining attributes. This is likely because AttGAN has a latent space learned in a supervised manner and hence attributes are better disentangled than with StyleGAN. Note that AttGAN is trained with a fixed set of attributes; if a new attribute of interest is introduced, the dataset needs to be re-labeled and AttGAN retrained. ZOOM, however, merely calls for the addition of a new text prompt. More results in Appendix B.

Measuring disentanglement of attributes

Previous works demonstrated that the StyleGAN's latent space can be entangled [39, 5], adding undesired dependencies when searching single-attribute counterfactuals. This section verifies that our framework can disentangle the attributes and mostly edit the desirable attributes.

We use CLIP as a super annotator to measure attribute changes during single-attribute modifications. For $1,000$ images, we record the attribute change after performing adversarial learning in each attribute, and average the attribute score change. Fig. 3.6 shows the confusion matrix during single-attribute counterfactual synthesis. The horizontal axis is the attribute being edited during the optimization, and the vertical axis represents the CLIP prediction changed by the process. For instance, the first column of Fig. 3.6a is generated when we optimize over bangs for the mustache classifier. We record the CLIP prediction variation. It clearly shows that bangs is the dominant attribute changing during the optimization. From the main diagonal of matrices, it is evident that the ZOOM mostly perturbs the attribute of interest. The results indicate reasonable disentanglement among attributes.

Visualization for edited images

Our methodology relies on CLIP-guided fine-grained image editing to provide adequate model diagnostics. It is critical to verify CLIP's ability to link language and visual representations. This section introduces two techniques for validating CLIP's capabilities. In this section, we analyze the decoupling of attribute editing used in StyleCLIP [31] in our domain.

Effect of λ. Fig. 3.16 shows the effect of λ in Equation 2 of the main text [31] . Originally in StyleCLIP, this filter parameter (denoted as β in [31]) helps the style disentanglement for editing. As we have normalized the edit vectors, which contributes to disentanglement in our framework, the impact of λ on style disentanglement is reduced. Consequently, λ primarily influences intensity control and denoising.

Single-attribute editing. Fig. 3.18 and Fig. 3.19 show a set of images of different object categories by editing different attributes extracted with the global edit directions method (as described in Section 3.2 of the main text). By analyzing the user's input attribute string, we can see that the modified image only alters in

the attribute direction while maintaining the other attributes.

Multiple-attribute editing. We demonstrate the validity of our method for editing multiple attributes through linear combination (as outlined in Equation 3 of the main text) by presenting illustrations of combined edits in Figure 3.20.

3.4.4 User study for edited images

To validate that our counterfactual image synthesis preserve fine-grained details and authenticity, we conducted a user study validating two aspects: synthesis fidelity and attribute consistency.

User study for synthesis fidelity. The classification of the counterfactual synthesis image vs real images by the user is employed to confirm that no unrealistic artifacts are introduced throughout the process of our model Fig. 3.15a shows sample questions of this study. In theory, the worst-case scenario is that users can accurately identify the semantic modification and achieve a user recognition rate of 100%. Conversely, the best-case scenario would be that users are unable to identify any counterfactual synthesis and make random guesses, resulting in a user recognition rate of 50%.

User study for attribute consistency. We ask users whether they agree that the counterfactual and original images are consistent on the ground truth w.r.t. the target classifier. For example, during the counterfactual synthesis for the cat/dog classifier, a counterfactual cat image should stay consistent as a cat. Fig. 3.15b shows another sample questions. The worst case is that the counterfactual changes the ground truth label to affect the target model, which makes the user agreement rate very low (even to zero).

The user study statistics are presented in Table 3.1. The study involved 34 participants with at least an undergraduate level of education, who were divided into two groups using separate collector links. The participants themselves randomly selected their group (i.e., the link clicked), and their responses were collected.

The production of high-quality counterfactual images is supported by the difficulty users had in differentiating them. Additionally, the majority of users concurred that the counterfactual images do not change the ground truth concerning the target classifier, confirming that our methodology generates meaningful counterfactuals.

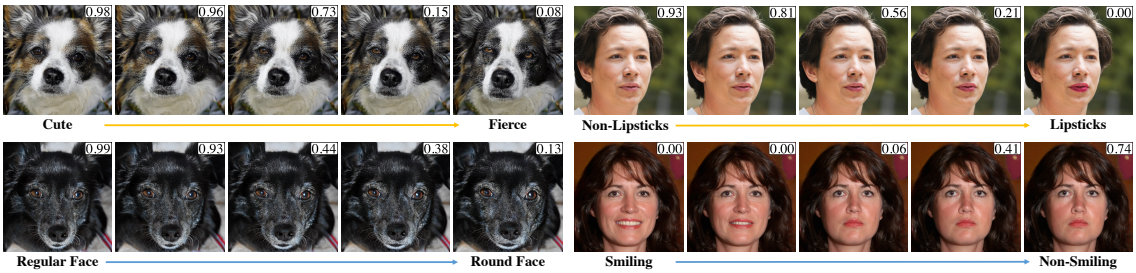


Figure 3.8: Effect of progressively generating counterfactual images on the Cat/Dog classifier (0-Cat / 1-Dog), and the Perceived Gender classifier (0-Female / 1-Male). Model probability prediction during the process is attached at the top right corner.

However, it should be noted that due to the nature of our recognition system, human volunteers are somewhat more responsive to human faces. As a result, we observed a slightly higher recognition rate in the human face (FFHQ) domain than in the animal face (AFHQ) domain.

Stability across CLIP phrasing/wording:

It is worth noting that the resulting counterfactual image is affected by the wording of the prompt used. In our framework, we subtract the neutral phrase (such as "a face") after encoding in CLIP space to ensure that the attribute edit direction is unambiguous enough. Our experimentation has shown that as long as the prompt accurately describes the object, comparable outcomes can be achieved. For instance, we obtained similar sensitivity results on the perceived-age classifier using prompts like "a picture of a person with X," "a portrait of a person with X," or other synonyms. Examples of this are presented in Figure 3.13.

Name of Study	Domain	Group 1	Group 2
Synthesis Fidelity (	FFHQ	62.12	71.79
Recognition Rate ↓, %)	AFHQ	51.30	50.55
Attribute Consistency (	FFHQ	94.12	90.76
Agreement Rate ↑, %)	AFHQ	89.92	88.26

Table 3.1: User study results. We can see from the table that our counterfactual synthesis preserves the visual quality and maintains the ground truth labels from the user's perspective.

Figure 3.9: ZOOM counterfactuals on more tasks (segmentation, multi-class classifier) and additional visual domains (cars, churches). Zoom in for better visibility.

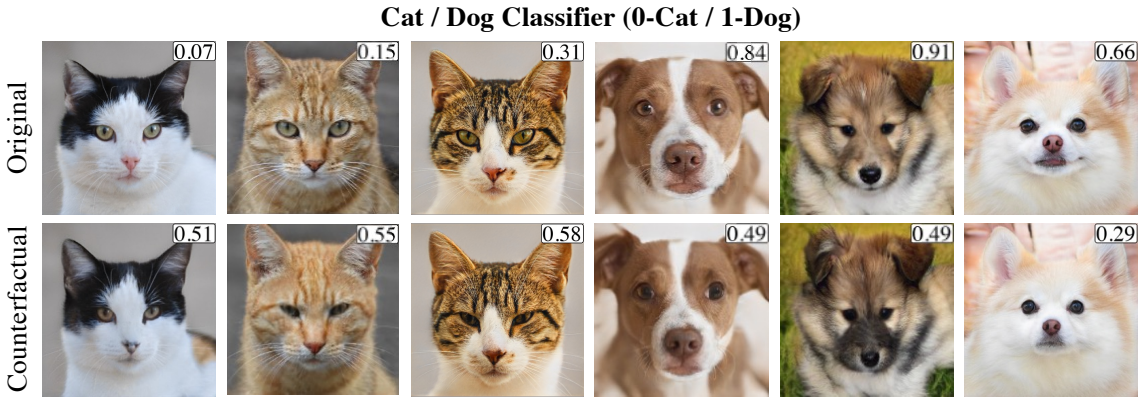


Figure 3.10: Multi-attribute counterfactual on Cat/Dog classifier. The number in each image is the predicted probability of being a dog.

3.4.5 Visual Model Diagnosis: Multi-Attributes

In the previous sections, we have visualized and validated single-attribute model diagnosis histograms and counterfactual images. In this section, we will assess ZOOM's ability to produce counterfactual images by concurrently exploring multiple attributes within $\mathcal{A}$, the domain of user-defined attributes. The approach conducts multi-attribute counterfactual searches across various edit directions, identifying distinct semantic combinations that result in the target model's failure. By doing so, we can effectively create more powerful counterfactuals images (see Fig. 3.11).

Fig. 3.7 and Fig. 3.10 show examples of multi-attribute counterfactual images generated by ZOOM, against human and animal face classifiers. It can be observed

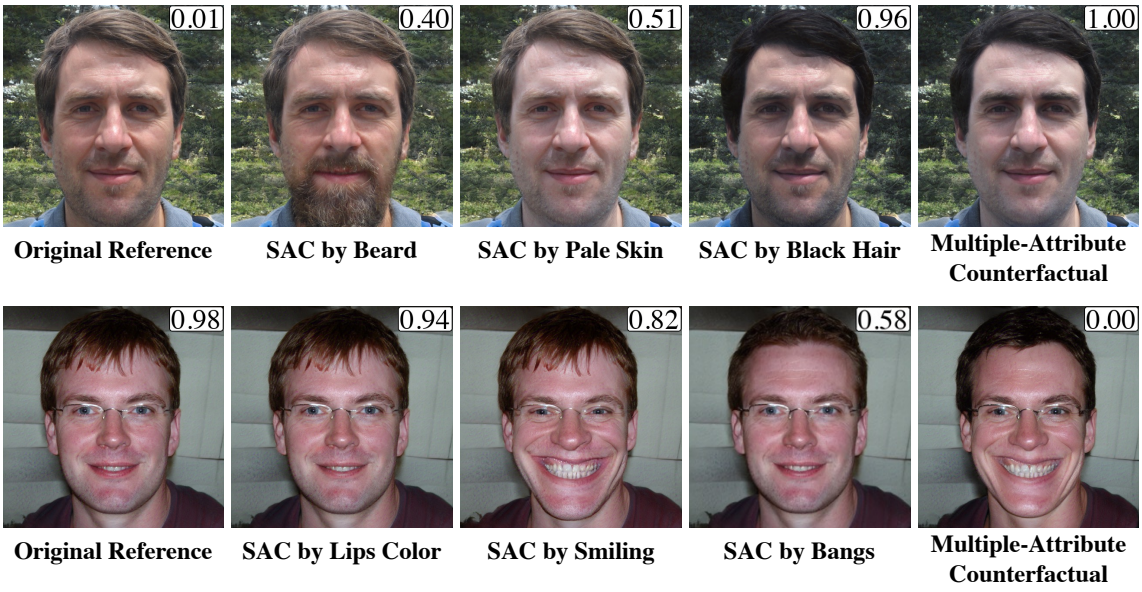
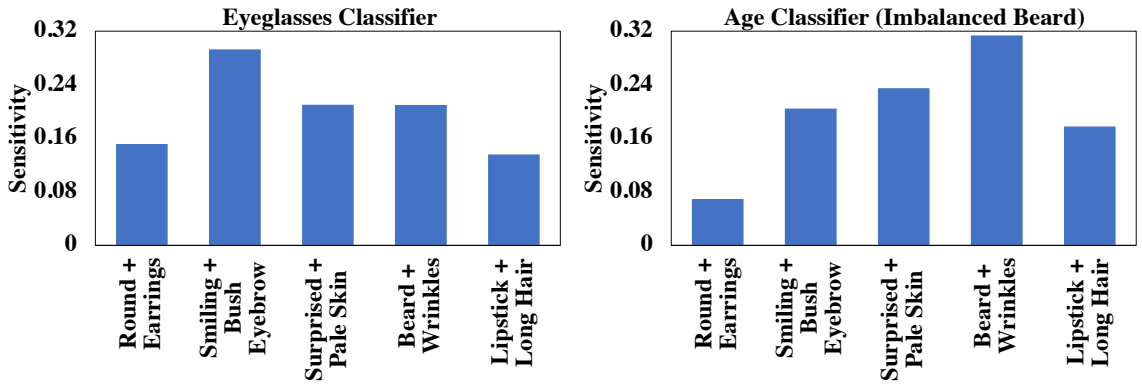


Figure 3.11: Multiple-Attribute Counterfactual (MAC, Sec. 3.4.5) compared with Single-Attribute Counterfactual (SAC, Sec. 3.4.2). We can see that optimization along multiple directions enable the generation of more powerful counterfactuals.

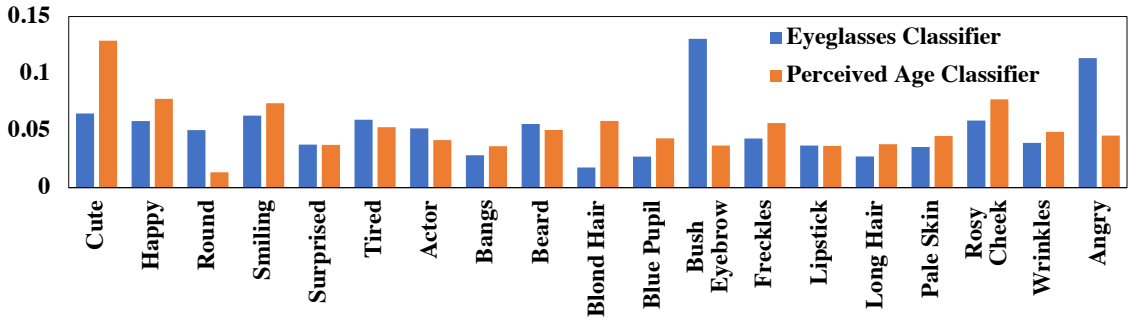
that multiple face attributes such as lipsticks or hair color are edited in Fig. 3.7, and various cat/dog attributes like nose pinkness, eye shape, and fur patterns are edited in Fig. 3.10. These attribute edits are blended to affect the target model prediction. Appendix B further illustrates ZOOM counterfactual images for semantic segmentation, multi-class classification, and a church classifier. By mutating semantic representations, ZOOM reveals semantic combinations as outliers where the target model underfits.

Fig. 3.17 shows more examples of multiple-attribute counterfactual images. In addition to binary classification and key-point detection in our manuscript, we further illustrate the extension of ZOOM counterfactuals on semantic segmentation, multi-class classification, and binary church classifier (BCC) in Fig. 3.9.

In the following sections, we will use the Flip Rate (the percentage of counterfactuals that flipped the model prediction) and Flip Resistance (the percentage of counterfactuals for which the model successfully withheld its prediction) to evaluate the multi-attribute setting.



(a) Sensitivity histograms generated by ZOOM on attribute combinations.



(b) Model diagnosis by ZOOM over 19 attributes. Our framework is generalizable to analyze facial attributes of various domains.

Figure 3.12: Customizing attribute space for ZOOM.

Customizing attribute space

In some circumstances, users may finish one round of model diagnosis and proceed to another round by adding new attributes, or trying a new attribute space. The linear nature of attribute editing (Eq. 3.3) in ZOOM makes it possible to easily add or remove attributes. Table 3.2 shows the flip rates results when adding new attributes into $\mathcal{A}$ for perceived age classifier and big lips classifier. We can observe that a different attribute space will results in different effectiveness of counterfactual images. Also, increasing the search iteration will make counterfactual more effective (see last row). Note that neither re-training the StyleGAN nor user-collection/labeling of data is required at any point in this procedure. Moreover, Fig. 3.12a shows the model diagnosis histograms generated with combinations of two attributes. Fig. 3.12b demonstrates the capability of ZOOM in a rich vocabulary setting where we can

Method	AC Flip Rate (%)	BC Flip Rate (%)
Initialize ZOOM by $\mathcal{A}$	61.95	83.47
+ Attribute: Beard	72.08	90.07
+ Attribute: Smiling	87.47	**96.27**
+ Attribute: Lipstick	90.96	94.07
+ Iterations increased to 200	**92.91**	94.87

Table 3.2: Model flip rate study. The initial attribute space $\mathcal{A}$ = {Bangs, Blond Hair, Bushy Eyebrows, Pale Skin, Pointy Nose}. AC is the perceived age classifier and BC is the big lips classifier.

analyze attributes that are not labeled in existing datasets [25, 41].

Counterfactual training results

This section evaluates regular classifiers trained on CelebA [25] and counterfactually-trained (CT) classifiers on a mix of CelebA data and counterfactual images as described in Sec. 3.3.5. Table 3.3 presents accuracy and flip resistance (FR) results. CT outperforms the regular classifier. FR is assessed over 10,000 counterfactual images, with FR-25 and FR-100 denoting Flip Resistance after 25 and 100 optimization iterations, respectively. Both use $\eta = 0.2$ and $\epsilon = 30$. We can observe that the classifiers after CT are way less likely to be flipped by counterfactual images while maintaining a decent accuracy on the CalebA testset. Our approach robustifies the model by increasing the tolerance toward counterfactuals. Note that CT slightly improves the CelebA classifier when trained on a mixture of CelebA images (original images) and the counterfactual images generated with a generative model trained in the FFHQ [20] images (different domain).

3.5 Discussion and Future Work

In this chapter, we present ZOOM, a zero-shot model diagnosis framework that generates sensitivity histograms based on user's input of natural language attributes. ZOOM assesses failures and generates corresponding sensitivity histograms for each attribute. A significant advantage of our technique is its ability to analyze the failures of a target deep model without the need for laborious collection and annotation of

Attribute	Metric	Regular (%)	CT (Ours) (%)
Perceived Age	CelebA Accuracy	86.10	**86.29**
	ZOOM FR-25	19.54	**97.36**
	ZOOM FR-100	9.04	**95.65**
Big Lips	CelebA Accuracy	74.36	**75.39**
	ZOOM FR-25	14.12	**99.19**
	ZOOM FR-100	5.93	**88.91**

Table 3.3: Results of network inference on CelebA original images and ZOOM-generated counterfactual. The CT classifier is significantly less prone to be flipped by counterfactual images, while test accuracy on CelebA remains performant.

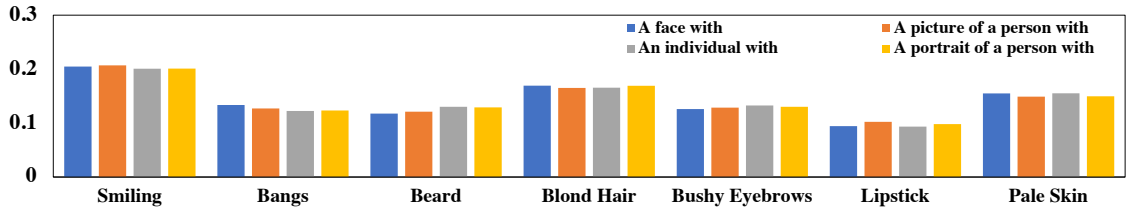


Figure 3.13: Sensitivity histograms when using four instances of phrases with a similar concept. Zoom in for better visibility.

test sets. ZOOM effectively visualizes the correlation between attributes and model outputs, elucidating model behaviors and intrinsic biases.

Our work has three primary limitations. First, users should possess domain knowledge as their input (text of attributes of interest) should be relevant to the target domain. Recall that it is a small price to pay for model evaluation without an annotated test set. Second, StyleGAN2-ADA struggles with generating out-of-domain samples. Nevertheless, our adversarial learning framework can be adapted to other generative models (e.g., stable diffusion), and the generator can be improved by training on more images. We have rigorously tested our generator with various user inputs, confirming its effectiveness for regular diagnosis requests. Currently, we are exploring stable diffusion models to generate a broader range of classes while maintaining the core concept. Finally, we rely on a pre-trained model like CLIP which we presume to be free of bias and capable of encompassing all relevant attributes.

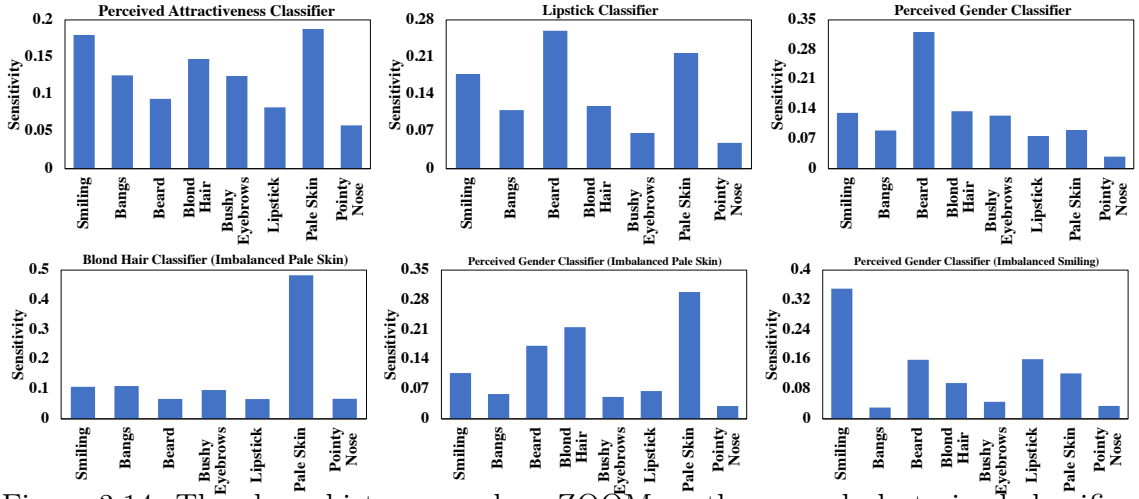
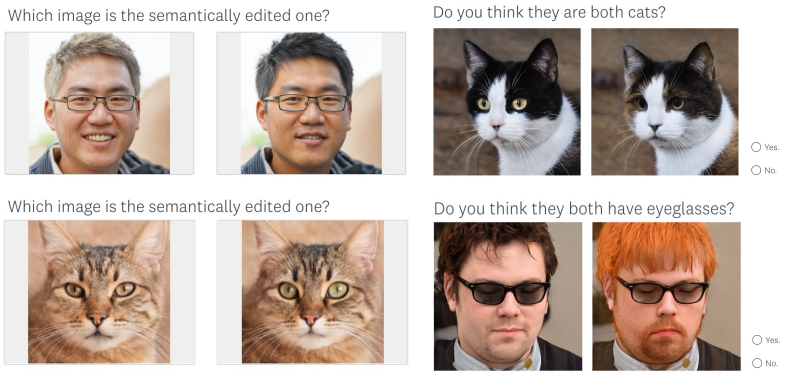


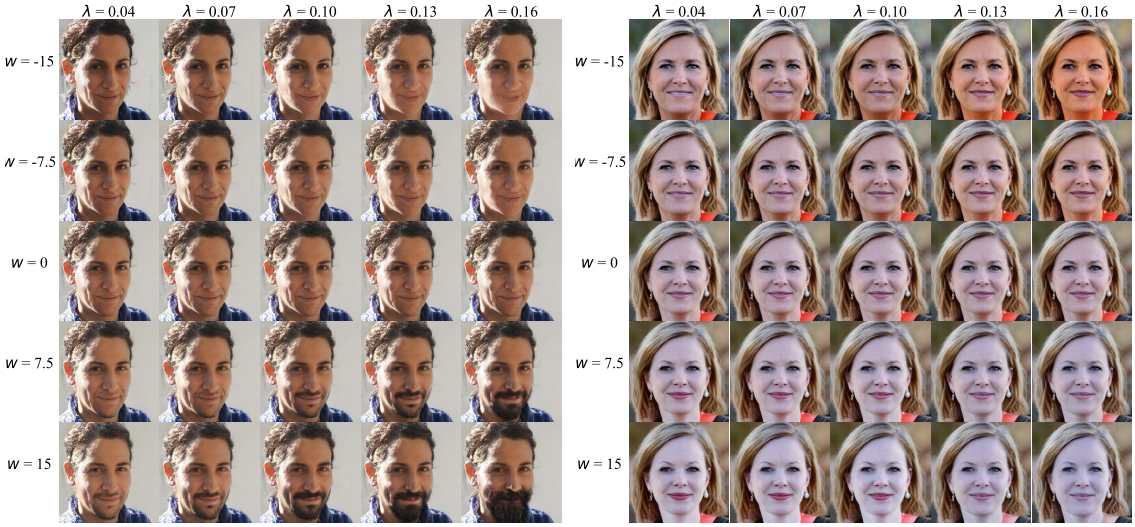
Figure 3.14: The above histograms show ZOOM on three regularly trained classifiers on CelebA, and the bottom histograms show ZOOM successfully detects the bias in the manually-crafted imbalanced classifiers.



(a) Evaluating visual fidelity. We show two images and let users choose the one that they think is edited. The counterfactuals are generated on Eyeglasses classifier and Cat/Dog classifier.

(b) Evaluating attribute consistency. The user classifies whether the ground truth is flipped. Example of counterfactual images on Cat/Dog classifier and Eyeglasses classifier is shown above.

Figure 3.15: Sample questions in the user study. Each user answers 10 questions for each of the two user studies.



(a) Effect of λ values for editing beard. (b) Effect of λ values for editing pale skin.

Figure 3.16: Visualization of the effect of different λ values.

(a) Multiple-attribute counterfactual for cat/dog classifier.

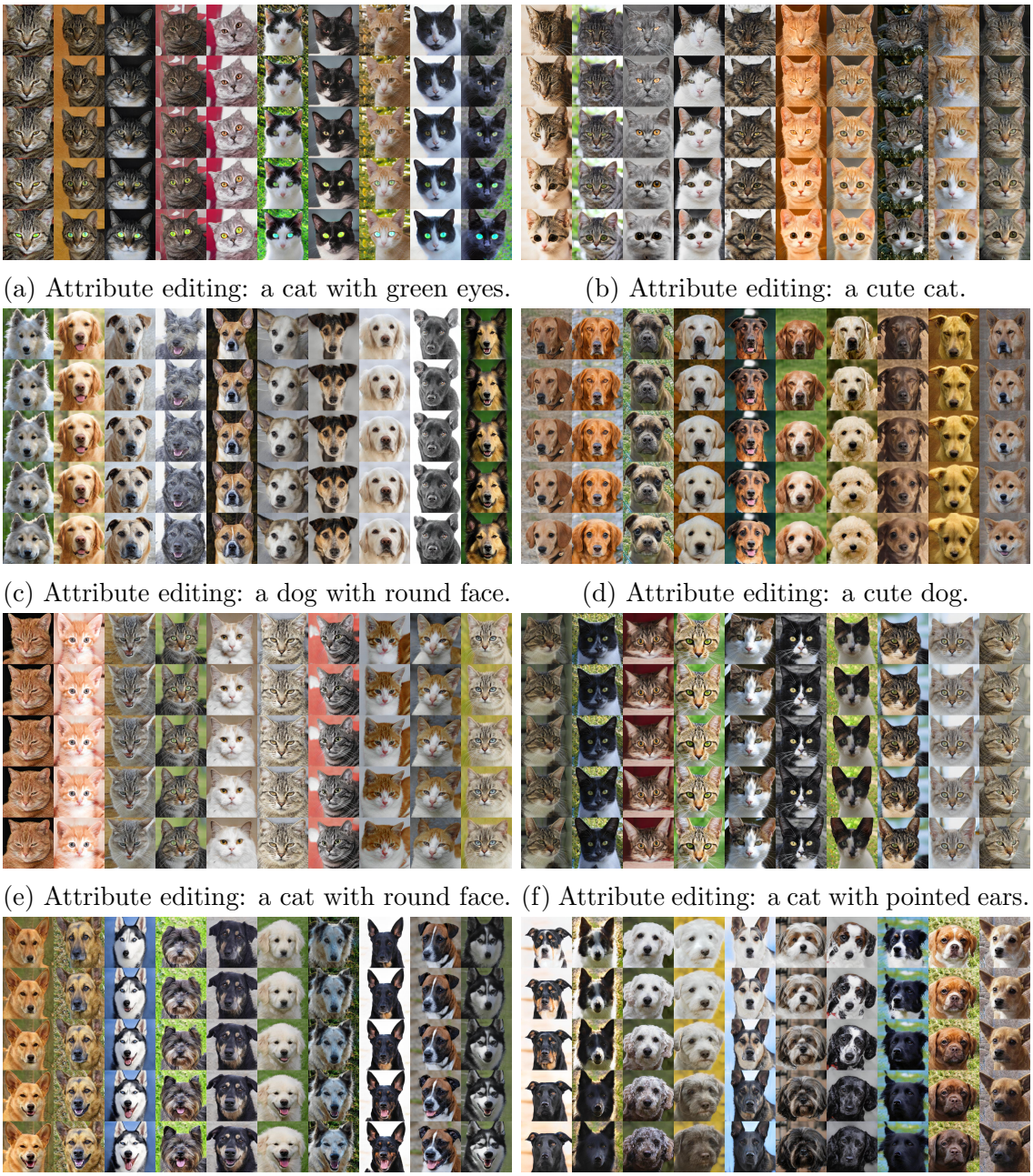
(b) Multiple-attribute counterfactual for eyeglasses classifier.

(c) Multiple-attribute counterfactual for perceived gender classifier.

(d) Multiple-attribute counterfactual for mustache classifier.

(e) Multiple-attribute counterfactual for perceived age classifier.

Figure 3.17: Multi-attribute counterfactual in the human face and animal face domain. The right-up corner of each image records the model output prediction.



(a) Attribute editing: a cat with green eyes.
(b) Attribute editing: a cute cat.
(c) Attribute editing: a dog with round face.
(d) Attribute editing: a cute dog.
(e) Attribute editing: a cat with round face.
(f) Attribute editing: a cat with pointed ears.
(g) Attribute editing: a dog with open mouth.
(h) Attribute editing: a black dog.

Figure 3.18: Visualization of global edit directions by utilizing the StyleCLIP channel relevance matrix. Images are sampled from the AFHQ domain using StyleGAN2-ADA. Every column demonstrates an edited image from edit weight $w = -30$ to $w = 30$. Weights of five images are linearly interpolated as $\{-30, -15, 0, 15, 30\}$. We can see that global edit directions are generalizable on multiple images.

(a) Attribute editing: an angry face.

(b) Attribute editing: a face with eyeglasses.

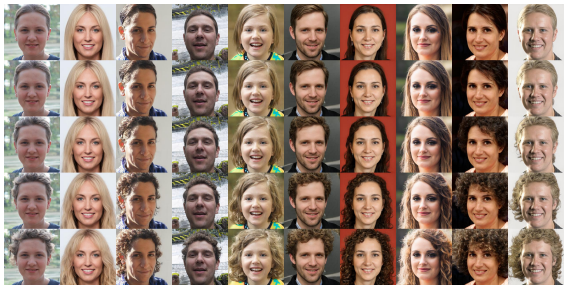
(c) Attribute editing: a cute face.

(d) Attribute editing: a face with blond hair.

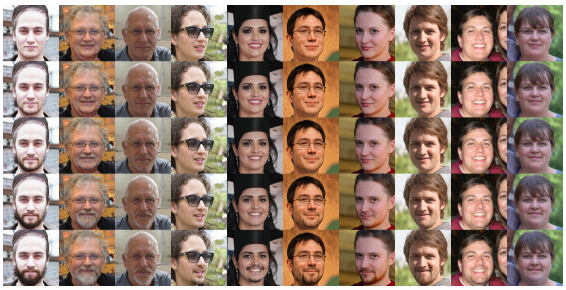
(e) Attribute editing: a face with bangs.

(f) Attribute editing: a smiling face.

(g) Attribute editing: a happy face.



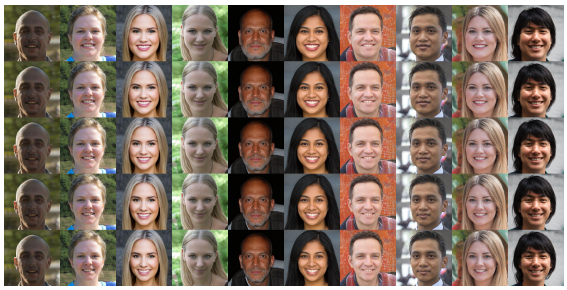
(h) Attribute editing: a face with curly hair.



(i) Attribute editing: a face with beard.

(j) Attribute editing: a face with lipstick.

(k) Attribute editing: a tired face.



(l) Attribute editing: a skinny face.

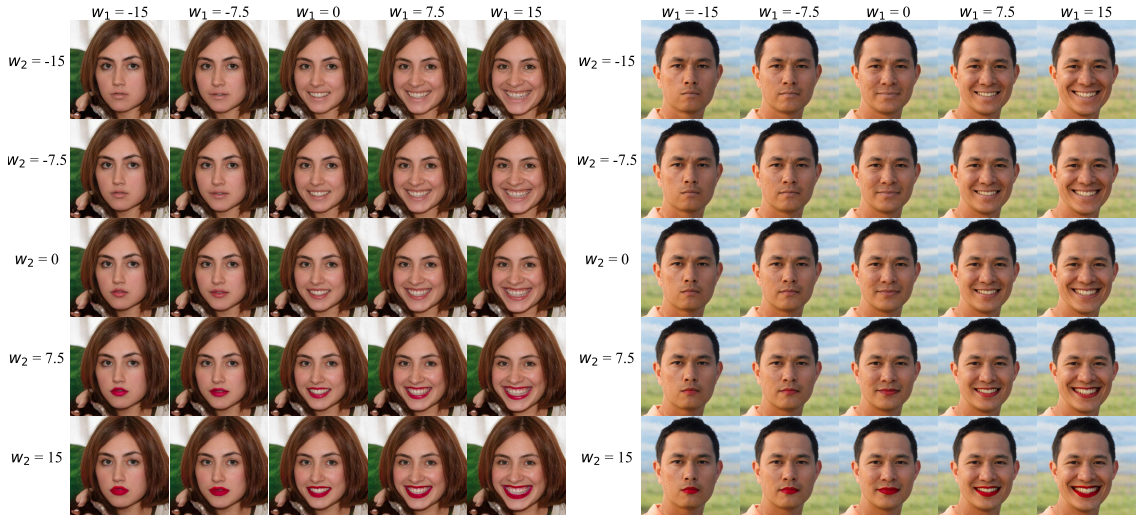
(m) Attribute editing: a male face.

(n) Attribute editing: a surprised face.

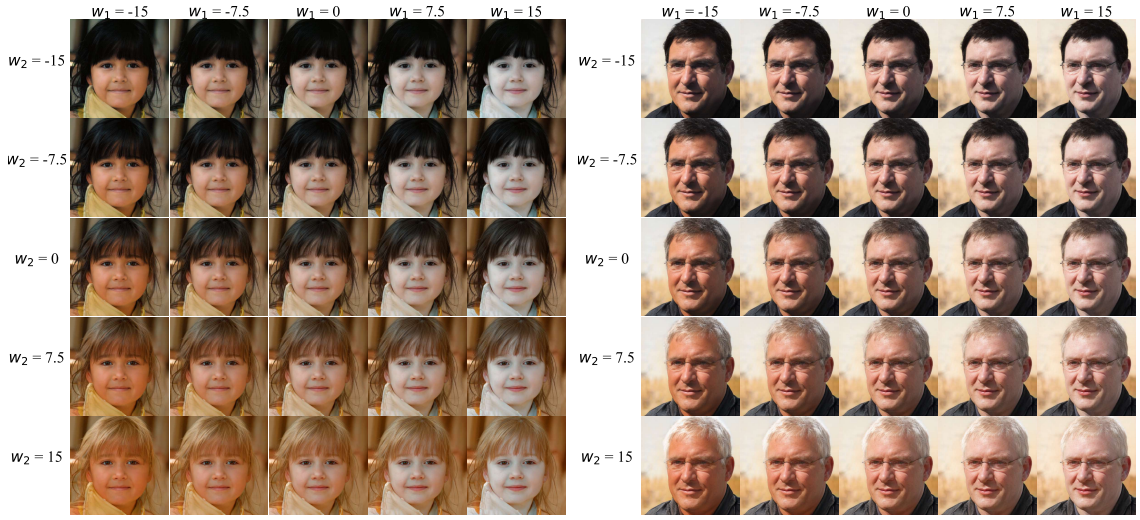
(o) Attribute editing: a face with long hair.

(p) Attribute editing: a face with pale skin.

Figure 3.19: Visualization of global edit directions by utilizing the StyleCLIP channel relevance matrix. Images are sampled from the FFHQ domain using StyleGAN2-ADA. Every column demonstrates an edited image from edit weight $w = -30$ to $w = 30$. Weights of five images are linearly interpolated as $\{-30, -15, 0, 15, 30\}$. We can see that global edit directions are generalizable on multiple images.



(a) Combination of smiling (w_1) and lipstick (w_2).



(b) Combination of pale skin (w_1) and blond hair (w_2).

Figure 3.20: Visualization of traversing on directional (attribute) style vectors to validate the effectiveness of multiple attribute editing.

Chapter 4

Conclusions

This thesis embarks on a journey to explore and elucidate the potential of generative models in the context of vision model diagnosis. Two main methodologies, Semantic Image Attack for Visual Model Diagnosis and Zero-shot Model Diagnosis, undergo extensive investigations and discussions. Their unique yet complementary approaches allow for flexible ways of model diagnosis without the need for time-consuming, costly, and error-prone test set collection and annotation.

Semantic Image Attack for Visual Model Diagnosis conducts the diagnosis through joint adversarial optimization in both controllable attribute space and pixel space. This method demonstrates immense promise in identifying and understanding model weaknesses, shedding light on the regions of the pixel space and attribute space where the model is less adversarially robust.

Zero-shot Model Diagnosis, through the innovative use of StyleGAN and CLIP, facilitates the generation of counterfactual images that visualize the sensitive factors for the target model. It highlights its remarkable ability to identify critical aspects where models fail, providing a substantial foundation for the subsequent improvement of these models.

In both methodologies, the practicality of these techniques in diagnosing new models and exploring user-defined attribute spaces without the necessity of collecting balanced datasets is noteworthy. This potential flexibility implies a high degree of scalability and usability across a wide array of potential application domains.

Looking forward, there is ample scope to refine these methods with better genera-

tive foundations, accommodate the diagnosis philosophy in broader visual domains, and adapt the frameworks to newer optimization strategies. The democratization of model diagnosis through generative models is expected to continue to gain momentum, further fueling interpretability, fairness, and robustness in deep learning.

In conclusion, this thesis serves as an important stepping stone in the field of vision model diagnosis. As the AI landscape continues to evolve, we envision more exploration, development, and deployment of model diagnosis frameworks will be conducted to fulfill the rising quests for more transparent, fair, and robust AI systems.

Bibliography

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. "Wasserstein Generative Adversarial Networks". In: *ICML*. 2017.

[2] Guha Balakrishnan, Yuanjun Xiong, Wei Xia, and Pietro Perona. "Towards Causal Benchmarking of Bias in Face Analysis Algorithms". In: *ECCV*. 2020.

[3] Nicholas Carlini and David Wagner. "Towards Evaluating the Robustness of Neural Networks". In: *IEEE SP*. 2017.

[4] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. "StarGAN v2: Diverse Image Synthesis for Multiple Domains". In: *CVPR*. 2020.

[5] Edo Collins, Raja Bala, Bob Price, and Sabine Süsstrunk. "Editing in Style: Uncovering the Local Semantics of GANs". In: *CVPR*. 2020.

[6] Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V. Le. "AutoAugment: Learning Augmentation Strategies From Data". In: *CVPR*. 2019.

[7] A. C. Davison and C.-L. Tsai. "Regression Model Diagnostics". In: *International Statistical Review*. 1992.

[8] Emily L. Denton, Ben Hutchinson, Margaret Mitchell, Timnit Gebru, and Andrew Zaldivar. "Image Counterfactual Sensitivity Analysis for Detecting Unintended Bias". In: *CVPR*. 2019.

[9] Tejas Gokhale, Rushil Anirudh, Bhavya Kailkhura, Jayaraman J. Thiagarajan, Chitta Baral, and Yezhou Yang. "Attribute-Guided Adversarial Training for Robustness to Natural Perturbations". In: *AAAI*. 2021.

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative Adversarial Nets". In: *NeurIPS*. Ed. by Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger. 2014.

[11] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. "Explaining and Harnessing Adversarial Examples". In: *ICLR*. Ed. by Yoshua Bengio and Yann LeCun. 2015.

[12] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. "Counterfactual Visual Explanations". In: *ICML*. 2019.

[13] Moritz Hardt, Eric Price, and Nathan Srebro. "Equality of Opportunity in Supervised Learning". In: *arXiv preprint arXiv:1610.02413* (2016).

[14] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen. "AttGAN: Facial Attribute Editing by Only Changing What You Want". In: *IEEE TIP*. 2019.

[15] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. "AvatarCLIP: Zero-Shot Text-Driven Generation and Animation of 3D Avatars". In: *ACM TOG*. 2022.

[16] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. "GANSpace: Discovering Interpretable GAN Controls". In: *NeurIPS*. 2020.

[17] Ameya Joshi, Amitangshu Mukherjee, Soumik Sarkar, and Chinmay Hegde. "Semantic Adversarial Attacks: Parametric Transformations That Fool Deep Classifiers". In: *ICCV*. 2019.

[18] Kimmo Karkkainen and Jungseock Joo. "FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation". In: *WACV*. 2021.

[19] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. "Training Generative Adversarial Networks with Limited Data". In: *NeurIPS*. 2020.

[20] Tero Karras, Samuli Laine, and Timo Aila. "A Style-Based Generator Architecture for Generative Adversarial Networks". In: *CVPR*. 2019.

[21] Oran Lang, Yossi Gandelsman, Michal Yarom, Yoav Wald, Gal Elidan, Avinatan Hassidim, William T. Freeman, Phillip Isola, Amir Globerson, Michal Irani, and Inbar Mosseri. "Explaining in Style: Training a GAN To Explain a Classifier in StyleSpace". In: *ICCV*. 2021.

[22] Bo Li, Qiulin Wang, Jiquan Pei, Yu Yang, and Xiangyang Ji. "Which Style Makes Me Attractive? Interpretable Control Discovery and Counterfactual Explanation on StyleGAN". In: *arXiv preprint arXiv:2201.09689* (2022).

[23] Dongze Li, Wei Wang, Hongxing Fan, and Jing Dong. "Exploring Adversarial Fake Images on Face Manifold". In: *CVPR*. 2021.

[24] Zhiheng Li and Chenliang Xu. "Discover the Unknown Biased Attribute of an Image Classifier". In: *ICCV*. 2021.

[25] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. "Deep Learning Face Attributes in the Wild". In: *ICCV*. 2015.

[26] Jinqi Luo, Zhaoning Wang, Chen Henry Wu, Dong Huang, and Fernando De la Torre. "Semantic Image Attack for Visual Model Diagnosis". In: *arXiv preprint arXiv:2303.13010* (2023).

[27] Jinqi Luo, Zhaoning Wang, Chen Henry Wu, Dong Huang, and Fernando De La Torre. "Zero-shot Model Diagnosis". In: *CVPR*. 2023.

[28] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. “Towards Deep Learning Models Resistant to Adversarial Attacks”. In: *ICLR*. 2018.

[29] Joanna Materzynska, Antonio Torralba, and David Bau. “Disentangling Visual and Written Concepts in CLIP”. In: *CVPR*. 2022.

[30] Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. “Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations”. In: *ACM FAccT*. 2020.

[31] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. “StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery”. In: *ICCV*. 2021.

[32] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. “DreamFusion: Text-to-3D using 2D Diffusion”. In: *arXiv preprint arXiv:2209.14988* (2022).

[33] Haonan Qiu, Chaowei Xiao, Lei Yang, Xinchen Yan, Honglak Lee, and Bo Li. “SemanticAdv: Generating Adversarial Examples via Attribute-conditioned Image Editing”. In: *ECCV*. 2020.

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. “Learning Transferable Visual Models From Natural Language Supervision”. In: *ICML*. 2021.

[35] Vikram V. Ramaswamy, Sunnis S. Y. Kim, and Olga Russakovsky. “Fair Attribute Classification through Latent Space De-biasing”. In: *CVPR*. 2021.

[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-Net: Convolutional Networks for Biomedical Image Segmentation”. In: *MICCAI*. 2015.

[37] Prasanna Sattigeri, Samuel C. Hoffman, Vijil Chenthamarakshan, and Kush R. Varshney. “Fairness GAN”. In: *arXiv preprint arXiv:1805.09910* (2018).

[38] Axel Sauer and Andreas Geiger. “Counterfactual Generative Networks”. In: *ICLR*. 2021.

[39] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. “InterFaceGAN: Interpreting the Disentangled Face Representation Learned by GANs”. In: *IEEE TPAMI*. 2020.

[40] Yujun Shen and Bolei Zhou. “Closed-Form Factorization of Latent Semantics in GANs”. In: *CVPR*. 2021.

[41] Philipp Terhörst, Daniel Fährmann, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. “MAAD-Face: A Massively Annotated Attribute Dataset for Face Images”. In: *IEEE TIFS*. 2021.

[42] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. “CLIP-NeRF: Text-and-Image Driven Manipulation of Neural Radiance Fields”. In: *CVPR*. 2022.

[43] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. "Deep High-Resolution Representation Learning for Visual Recognition". In: *IEEE TPAMI*. 2019.

[44] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. "Image Quality Assessment: from Error Visibility to Structural Similarity". In: *IEEE TIP*. 2004.

[45] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Matthew Johnson, Virginia Estellers, Thomas J. Cashman, and Jamie Shotton. "Fake It Till You Make It: Face analysis in the wild using synthetic data alone". In: *arXiv preprint arXiv:2109.15102* (2021).

[46] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. "Look at Boundary: A Boundary-Aware Face Alignment Algorithm". In: *CVPR*. 2018.

[47] Zongze Wu, Dani Lischinski, and Eli Shechtman. "StyleSpace Analysis: Disentangled Controls for StyleGAN Image Generation". In: *CVPR*. 2021.

[48] Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. "GAN Inversion: A Survey". In: *IEEE TPAMI*. 2022.

[49] Chaowei Xiao, Bo Li, Jun-yan Zhu, Warren He, Mingyan Liu, and Dawn Song. "Generating Adversarial Examples with Adversarial Networks". In: *IJCAI*. 2018.

[50] Qingsong Yao, Zecheng He, Hu Han, and S. Kevin Zhou. "Miss the Point: Targeted Adversarial Attack on Multiple Landmark Detection". In: *MICCAI*. 2020.

[51] Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Seong Joon Oh, Youngjoon Yoo, and Junsuk Choe. "CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features". In: *ICCV*. 2019.

[52] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. "Mitigating Unwanted Biases with Adversarial Learning". In: *AAAI*. 2018.

[53] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. "MotionDiffuse: Text-Driven Human Motion Generation with Diffusion Model". In: *arXiv preprint arXiv:2208.15001* (2022).

[54] Congcong Zhu, Xiaoqiang Li, Jide Li, and Songmin Dai. "Improving Robustness of Facial Landmark Detection by Defending Against Adversarial Attacks". In: *ICCV*. 2021.