Class Attribute Inference Attacks: Inferring Sensitive Class Information by Diffusion-Based Attribute Manipulations

Lukas Struppek [*1,2] **Dominik Hintersdorf** [1,2] **Felix Friedrich** [2,3]
Manuel Brack [1,2] **Patrick Schramowski** [1,2,3,5] **Kristian Kersting** [1,2,3,4]

[1]German Research Center for Artificial Intelligence (DFKI)
[2]Technical University of Darmstadt
[3]Hessian Center for AI (Hessian.AI)
[4]Centre for Cognitive Science, Technical University of Darmstadt
[5]LAION

Abstract

Neural network-based image classifiers are powerful tools for computer vision tasks, but they inadvertently reveal sensitive attribute information about their classes, raising concerns about their privacy. To investigate this privacy leakage, we introduce the first **C**lass **A**ttribute **I**nference **A**ttack (CAIA), which leverages recent advances in text-to-image synthesis to infer sensitive attributes of individual classes in a black-box setting, while remaining competitive with related white-box attacks. Our extensive experiments in the face recognition domain show that CAIA accurately infers undisclosed sensitive attributes, such as an individual's hair color, gender, and racial appearance, which are not part of the training labels.

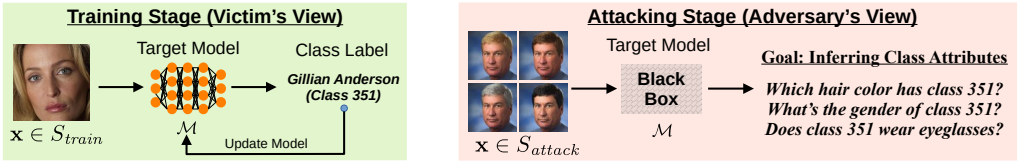
1 Introduction

Classifying images with neural networks is widely adopted in various domains [4, 11, 2, 1]. In the pursuit of enhancing predictive performance, privacy concerns of the acquired knowledge are often disregarded and moved into the background. We investigate the privacy leakage of face recognition models and demonstrate that models indeed leak sensitive identity information even without any specific information about the classes, training samples, or attribute distributions available to the adversary. Our research shows that these models reveal sensitive details about the different identities within their outputs, such as gender, hair color, and racial appearance.

We introduce a ***C**lass **A**ttribute **I**nference **A**ttack* (CAIA), which enables an adversary to infer sensitive attributes of a specific class from a trained image classifier with high accuracy. Phrased differently, CAIA allows the creation of a profile of the individual classes by only interacting with the trained classifier through the extraction of class attributes that have not been explicitly part of the training objective. Fig. 1 illustrates the setting of our investigation, in which the adversary has only black-box access to the target model and no specific information about the appearance of individual identities. The adversary then aims to infer sensitive information about the identities from the target model's training data. For the attack, we utilize recent advances in text-to-image synthesis to craft images that only differ in one attribute by editing real images with textual guidance. We then exploit that image classifiers assign higher logits to inputs that share the same sensitive attribute with the training

[*]corresponding author: struppek@cs.tu-darmstadt.de
Code: `https://github.com/LukasStruppek/Class_Attribute_Inference_Attacks`

AdvML-Frontiers'24: The 3rd Workshop on New Frontiers in Adversarial Machine Learning@NeurIPS'24, Vancouver, CA.



(a) Training on private data.

(b) Inferring sensitive information about the identities.

Figure 1: Overview of the setting of our proposed Class Attribute Inference Attack (CAIA). The victim (1a) first trains a model on a private training set S_{train} to predict the identity of a given input image. The adversary (1b) only has black-box access to the trained model without any information about the individual classes or training samples in S_{train}. The adversary then tries to infer sensitive attributes about the individual class identities, e.g., their hair color or gender appearance. For this, the adversary has access to a separate attack dataset S_{attack} of public facial images.

samples of a class, which allows us to infer class information by observing the input-output relation of the trained target model. Compared to related inference attacks [6, 5, 24, 19], CAIA is model-agnostic and requires only black-box access and basic domain knowledge. Once the attack images are crafted, attribute inference requires only a single model forward pass of the generated samples.

2 Background and Related Work

Most related to our work are attribute inference attacks (AIAs) [6], which aim to infer sensitive attribute values of an incomplete data record in the context of tabular data. Common AIAs make strong assumptions regarding the adversary's knowledge that is generally hard to gain under realistic assumptions, e.g., the adversary knows the marginal prior of sample attributes [6, 5, 22] or the model's confusion matrix on its training data [5, 15]. AIAs are typically limited to tabular data and are not applicable to image classification since the variation of single image attributes, e.g., changing the hair color in a facial image, is not trivially possible. This fact also makes it impossible to directly compare our ***C**lass **A**ttribute **I**nference **A**ttack* (CAIA) with common AIAs and corresponding defenses.

Another class of attacks, so-called model inversion attacks (MIAs), try to fill this gap for image classification. We note that the notion of MIAs is not consistent in the literature, and the term is sometimes also used for AIAs. Generally, given a classification model, an adversary attempts to create synthetic input samples that either reconstruct samples from the model's training data [5, 24, 3, 12] or craft synthetic samples that reflect the characteristics of a specific class [21, 19, 20]. Whereas most MIAs require access to samples from the target training distribution for training a custom generative adversarial network (GAN) [7], Struppek et al. [19] recently proposed Plug & Play (PPA) MIAs, which make the attacks agnostic to the target model, increase their flexibility, and enhance their inference speed by utilizing pre-trained GANs. We will use PPA as a baseline during our evaluation.

Novelty of CAIA. In contrast to AIAs, we move the scope of inference attacks from the sample level to a class level in the vision domain, with the goal of inferring sensitive information about the distinct classes. To achieve this, we use recent advancements in text-to-image synthesis to manipulate single attributes of input images, resulting in consistent images that differ only in the targeted attribute. Our approach is more effective than MIAs as it requires only black-box access to the target model and no extensive knowledge of the training data distribution. Furthermore, the inference step is done in seconds, since CAIA does not require any further sample optimization after constructing the initial attack dataset. This makes CAIA more flexible, target-independent and efficient than previous attacks.

3 Class Attributes Inference Attacks

We now introduce the novel ***C**lass **A**ttribute **I**nference **A**ttack* (CAIA), which consists of two steps. First, the adversary generates a set of attack samples by creating different versions of images through a generative approach that alters the sensitive attribute values. In the second step, these samples are used to infer the sensitive attribute values for the identities in a face recognition model. The underlying attack assumption is that image classifiers assign higher prediction scores to samples that share the sensitive attribute value with the training samples of a class. Before delving into the details, we outline our general threat model.

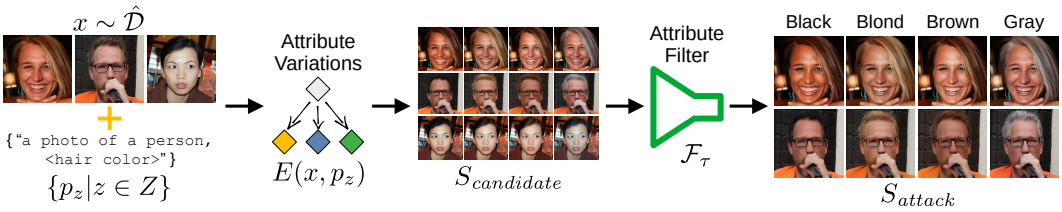


Figure 2: Overview of our attack dataset crafting process for the sensitive attribute *hair color*, which has four possible values. Real images are used to generate image variations by modifying characteristics associated with the sensitive attribute. The resulting candidate images are then filtered to ensure that each sample accurately reflects the intended attribute values. The final output of this process is the set of attack samples.

Adversary's Goal. Let $\mathcal{M}\colon X \to \mathbb{R}^{|Y|}$ denote the trained target classifier, which takes input images $x \in X$ and computes prediction scores for each class label $y \in Y$. The model's training data S_{train} consisted of labeled data samples $(x, y) \sim \mathcal{D}$. The underlying attack assumption is that samples of a certain class share a constant sensitive attribute value $z \in Z$, which is not part of the class label but is implicitly encoded in the image features. For example, a face recognition model is a classifier trained to predict the identity y of each facial image x. A sensitive attribute z in this context might be the gender appearance or hair color of a specific identity. The attack goal is to infer the value of this sensitive attribute for each individual class. Fig. 1 demonstrates the general setting, in which the victim trains the target model on labeled images of various identities. In the depicted case, class y corresponds to the identity of the actress *Gillian Anderson*. The adversary tries to predict sensitive attributes about the class y without any further information such as the name or training samples available. Multiple classes can share the same attribute, e.g., having the same hair color.

Adversary's Capabilities. The adversary has only black-box access to the trained target model $\mathcal{M}$, i.e., the adversary can query the target model and observe its output logits. Furthermore, the adversary knows the domain of the target model's training data, e.g., facial images, but not the exact training data distribution or labels. Instead, the adversary can sample images from a data distribution $\hat{\mathcal{D}}$ from the same domain. Importantly, the images available to the adversary contain no identity labels and have no overlapping with the target models training data. For the sensitive attributes, the adversary defines an individual set of possible values to infer. We emphasize that CAIA is model- and task-agnostic and does not require white-box access to the target model. Information about the prior distribution of the sensitive attributes is neither available nor required.

3.1 Crafting Attack Samples

We utilize recent advances in text-to-image synthesis to enable meaningful image manipulations. Text-to-image synthesis systems like Stable Diffusion [18] are capable of generating high-quality images following a user-provided text description p. The recently proposed *Null-text Inversion* [16] encodes real images into the domain of diffusion models and enables text-based editing while keeping the overall image composition and content fixed. In combination with *Prompt-to-Prompt* [9], it allows a user to instruct image edits $x_{edit} = E(x, p)$ on images x conditioned on a description p. We apply *Null-text Inversion* to generate variations of existing images by changing only the sensitive attribute values while aiming to leave other image aspects unchanged.

Fig. 2 illustrates the crafting process for the attack samples. Formally, the adversary has access to a data distribution $\hat{\mathcal{D}}$ from which to sample images x. Note that the attack does not require the attack distribution $\hat{\mathcal{D}}$ to be the same as the training data distribution $\mathcal{D}$ but only that both data distributions are from the same domain, e.g., facial images. As we will show in our experimental evaluation, even if the style, size, and quality of images between both distributions vary significantly, the attack is still highly successful. The adversary defines the target attribute with a set of k possible attribute values $Z = \{z_1, \ldots, z_k\}$ and corresponding edit prompts p_z that describe the general domain and explicitly state an attribute value $z \in Z$. For example, $p_z =$ "A photo of a person, $\langle$gender$\rangle$", where $\langle$gender$\rangle$ is replaced by $z \in \{\text{female appearance}, \text{male appearance}\}$. The candidate dataset $S_{candidate}$ is then constructed as

$$S_{candidate} = \{E(x, p_z) | z \in Z, x \sim \hat{\mathcal{D}}\}, \tag{1}$$

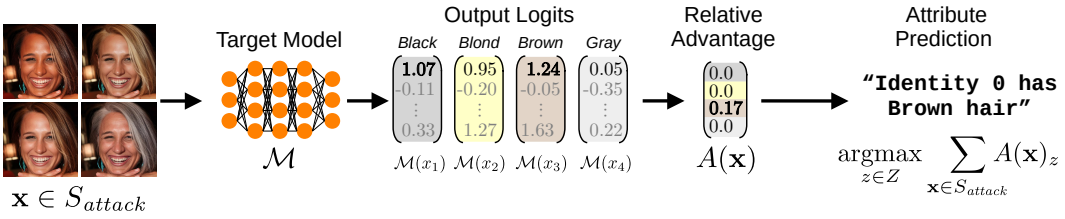


Figure 3: Overview of the class attribute inference step. Each image tuple from the attack set is fed sequentially into the target model to compute the logits for the target class. The relative advantage is then computed by subtracting the second-highest logit value from the maximum value, and this difference is added to a running sum for each sensitive attribute value. The final prediction for the sensitive attribute is the value with the highest relative advantage computed across all attack samples.

consisting of image tuples $\mathbf{x} = (x_1, \ldots, x_k)$, each containing k images with different sensitive attribute values. However, the attribute manipulation might not always succeed in changing an image attribute to the desired value. We, therefore, employ a filtering approach, i.e., filtering out all sample tuples $\mathbf{x}$ that are not correctly depicting the various attribute values. For this, we use a trained attribute classifier $\mathcal{F}_\tau \colon X \to Z$ to create a subset

$$S_{attack} = \{\mathbf{x} \in S_{candidate} | \mathcal{F}_\tau(x_z) = z, \forall z \in Z\} \tag{2}$$

of sample tuples $\mathbf{x} = (x_1, \ldots, x_k)$. The attribute classifier computes for each input image a probability score that the attribute value $z \in \mathcal{Z}$ is present in the image. We also add a threshold τ on the softmax scores of the attribute classifier and classify only predictions with a softmax score $\geq \tau$ as correct. This removes images for which the attribute classifier has only low confidence in its prediction. For the example of hair color, each resulting tuple $\mathbf{x} \in S_{attack}$ consists of four facial images $(x_1, \ldots, x_4)$ that only differ in the depicted hair color of the person. This use case is also depicted in the attack dataset crafting process in Fig. 2.

3.2 Revealing Sensitive Attributes

After crafting the attack samples, we can now begin inferring the sensitive class attributes learned by the target classifier. We recall that the adversary has no specific information about the distinct classes and particularly no access to the training data samples. Fig. 3 illustrates the basic concept of the inference process. In the depicted case, the adversary tries to infer the sensitive attribute *hair color* of the first identity, which corresponds to the first class of the face recognition model. While our filtering approach ensures that the variations of an image depict the different values of the sensitive attribute, other confounding behavior might still remain unintentionally. For example, changing a person's hair color to gray might also influence the depicted age. To mitigate such influences, we predict the sensitive attribute with a variety of different attack samples to, in turn, reduce the influence of these confounding factors over a larger number of samples.

Be $\mathcal{M}(x)_y \colon X \to \mathbb{R}$ the pre-softmax logits computed by the target model $\mathcal{M}$ on input image x for class y. To infer the sensitive attribute value z of class y, we query the target model consecutively with multiple sample tuples $\mathbf{x} \in S_{attack}$. We then compute for each tuple $\mathbf{x}$ the relative advantage $A(\mathbf{x}) \in \mathbb{R}^{|Z|}$. For each $x_z \in \mathbf{x}$, the relative advantage component $A(\mathbf{x})_z$ is defined by

$$A(\mathbf{x})_z = \max\left(0,\, \mathcal{M}(x_z)_y - \max_{\tilde{x} \in \mathbf{x}, \tilde{x} \neq x_z} \mathcal{M}(\tilde{x})_y\right). \tag{3}$$

The relative advantage computes the difference between the highest and the second-highest logit values and assigns this difference to the attribute sample $x_z \in \mathbf{x}$ with the highest logit. For all other samples $\tilde{x} \in \mathbf{x}$ with $\tilde{x} \neq x_z$, the relative advantage is set to zero. Fig. 3 illustrates the relative advantage computation for a single $\mathbf{x}$. In the depicted case, the sample depicting *brown hair color* achieves the highest logit value and its relative advantage of $A(\mathbf{x})_{brown} = 0.17$ describes the difference to the second highest logit value assigned to the sample with the attribute *black hair color*. The relative advantage of the other three attribute values is set to zero. The final attribute prediction $\hat{z}$ is then done by taking the attribute with the highest relative advantage summed up over all attack samples. We emphasize that only a single forward pass on all attack samples is sufficient to compute the relative advantage for all target classes, which makes the inferring step very fast.

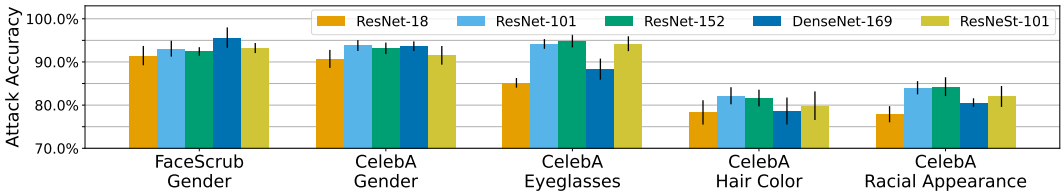
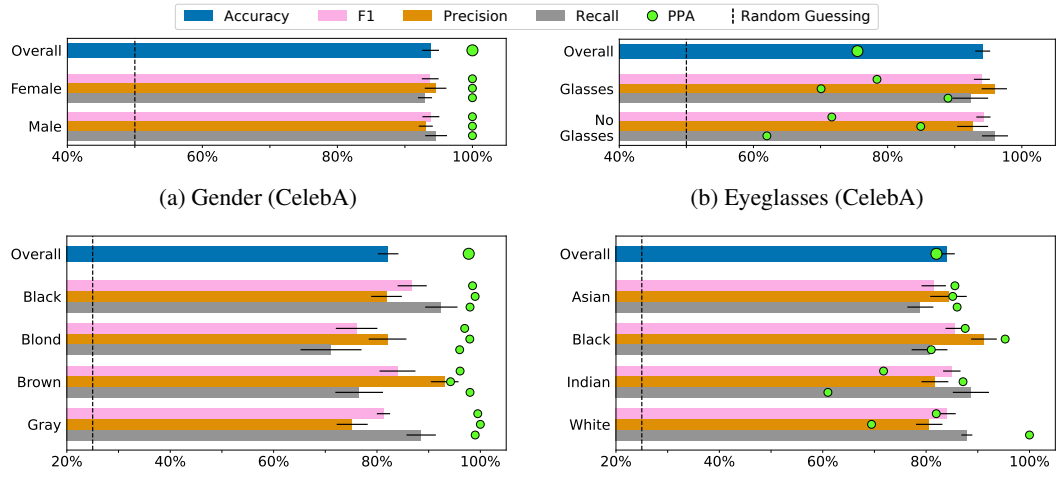


Figure 4: Attack accuracy for different target model architectures and CelebA attribute datasets. Results are averaged over three models and three attack datasets. The attacks are comparably successful on the different models.



(a) Gender (CelebA)

(b) Eyeglasses (CelebA)

(c) Hair Color (CelebA)

(d) Racial Appearance (CelebA)

Figure 5: Evaluation results for CAIA performed on ResNet-101 CelebA models to infer four different target attributes. The black horizontal lines denote the standard deviation over nine runs. We further state random guessing (dashed line) and Plug and Play Attacks (PPA, green dots) for comparison. While CAIA outperforms random guessing by a large margin, it extracts information on racial appearance and if someone is wearing eyeglasses even more reliably than the white-box PPA attack.

4 Experimental Evaluation

Target model training. We trained our target face recognition systems on the CelebA facial attributes dataset [14]. We selected the following sensitive attributes: *gender appearance* $= \{\text{female}, \text{male}\}$, *eyeglasses* $= \{\text{no eyeglasses}, \text{eyeglasses}\}$, *hair color* $= \{\text{black}, \text{blond}, \text{brown}, \text{gray}\}$, *racial appearance* = {Asian, Black, Indian, White}. Since the provided attributes and labels are often inconsistent for samples of one identity, we created custom subsets for each sensitive attribute group by selecting an equal number of identities for each attribute value and removed samples with inconsistent labels. Additional target models were trained on the FaceScrub [17] facial image dataset, which contains images of 530 identities with equal gender split.

We trained ResNets [8], DenseNets [10] and ResNeSts [23] as target models to predict a person's identity. We emphasize that the attribute labels were not part of the training process. Since not every attribute is present with every identity, we selected the 100 CelebA identities for each attribute value of *hair color*, *eyeglasses*, and *racial appearance*, respectively, with the most training samples available. For *gender appearance*, we selected 250 identities per attribute value. We split all datasets into 90% for training and 10% for testing the models' performance.

Attack Datasets. To craft the attack datasets, we used the Flickr-Faces-HQ (FFHQ) [13] dataset of high-resolution facial images. We generated and filtered images with attribute manipulations to collect 300 attack image tuples for each attribute group.

Metrics. We computed the precision, recall, and F1 score for each attribute value, together with the overall prediction accuracy for all attributes. All experimental results are averaged over three independently trained models and three disjoint subsets of the attack datasets.

Results. The attack accuracy for different models and target attributes in Fig. 4 demonstrates that CAIA performed comparably well on different architectures and predicted the sensitive attribute values correctly in over 90% of the cases for the attributes *gender* and *eyeglasses* and about 80% for the *hair color* and *racial appearance*. Only the attack results of ResNet-18 stand out and are a few percentage points lower than those of the other architectures, which we attribute to the small number of model parameters (only about a quarter of ResNet-101). Still, all attacks reliably inferred the sensitive attributes in most cases.

Next, we investigate the attribute leakage more closely. Therefore, we performed a more detailed analysis of the attribute leakage of ResNet-101 models, for which the results are depicted in Fig. 5. For all four attributes, CAIA significantly beats the random guessing baseline by a large margin. Whereas *gender* and *eyeglasses* were predicted correctly in about 94% of the cases, *racial appearance* could be inferred correctly in 84%. The attack accuracy for *hair color* was also about 82% on average, but the attack success varied substantially between the different attribute values. Further analyzing the individual predictions for the hair color, blond hair seems to be the hardest value to predict and is frequently confused with gray hair, which is not unexpected since hair colors have different shades, of which blond might be the broadest one. Another reason for the confusion in these attributes is that the CelebA dataset also contains numerous training images of poor quality and sometimes disturbing lighting, which could interfere with the ground-truth identity attributes.

We also compared CAIA to PPA [19] as a state-of-the-art white-box model inversion method that reconstructs characteristic class inputs. The attributes are predicted by applying our filter models on the generated attack result images. Whereas PPA beats CAIA for *gender* and *hair color* prediction, both approaches achieve similar results for the *racial appearance*. However, for inferring whether an identity is wearing *eyeglasses*, PPA fell significantly behind CAIA.

5 Discussion and Conclusion

Our experiments demonstrate that classifiers indeed leak sensitive class information, with significant implications for the secure application of machine learning models. CAIA infers sensitive information with high accuracy, comparable to white-box MIAs. An open question is how to limit the information about sensitive attributes in a model's output. Since attributes like hair color or eyeglasses inherently incorporate high predictive power, restricting their influence on a model's prediction probably leads to a substantial loss in model utility. However, we believe that encouraging a model to produce logits in a more narrow range already limits the relative logit differences between the presence of different attribute values, and thereby reduces the success of CAIA.

Whereas CAIA offers reliable attribute inference capabilities, it still faces some challenges and limitations. Accurately evaluating class information leakage requires high-quality data and consistent labeling, which heavily relies on the underlying dataset. We note that the sample quality of CelebA and FaceScrub images varies widely in terms of resolution, sharpness, and coloring, and the labeling is not always consistent within an identity or attribute class, with also falsely labeled samples contained in the datasets. For instance, the boundaries between gray and blond hair are not well-defined, which can lead to reduced precision for these attributes. However, CAIA is generally successful in inferring sensitive attributes in most cases. The attack metrics might even underestimate its effectiveness since false attribute predictions might still provide some identity information, e.g., black hair tone predictions make blond rather unlikely.

Expanding on the mentioned challenges, we expect that with the upcoming developments in text-guided image manipulation, we propose to enhance CAIA and extend its use to continuous features like age or more detailed skin tone grading. Moreover, in its current implementation, CAIA infers each attribute independently. However, in the real world, different attributes often correlate with each other. To this end, CAIA could leverage already inferred attributes as a prior for inferring additional attributes. We also envision the use of CAIA beyond privacy analyses. For example, it is exciting to explore it in the context of explainable AI to determine the importance of features by analyzing prediction scores for different feature characteristics.

In conclusion, our research provides novel insights into the privacy of image classifiers and shows that models leak more sensitive information than previously assumed. It combines both areas of research, which have largely been studied separately. We hope our work motivates future security research and defense endeavors in building secure and private models.

References

[1] Prabal Datta Barua, Nadia Fareeda Muhammad Gowdh, Kartini Rahmat, Norlisah Ramli, Wei Lin Ng, Wai Yee Chan, Mutlu Kuluozturk, Sengul Dogan, Mehmet Baygin, Orhan Yaman, Turker Tuncer, Tao Wen, Kang Hao Cheong, and U. Rajendra Acharya. Automatic covid-19 detection using exemplar hybrid deep features with x-ray images. *International Journal of Environmental Research and Public Health*, 18(15), 2021.

[2] Ulrich Baumann, Yuan-Yao Huang, Claudius Gläser, Michael Herman, Holger Banzhaf, and J. Marius Zöllner. Classifying road intersections using transfer-learning on a deep neural network. In *International Conference on Intelligent Transportation Systems (ITSC)*, pages 683–690, 2018.

[3] Si Chen, Mostafa Kahla, Ruoxi Jia, and Guo-Jun Qi. Knowledge-Enriched Distributional Model Inversion Attacks. In *International Conference on Computer Vision (ICCV)*, pages 16178–16187, 2021.

[4] Andre Esteva, Kat Chou, Serena Yeung, Nikhil Naik, Ali Madani, Ali Mottaghi, Yun Liu, Eric Topol, Jeff Dean, and Richard Socher. Deep learning-enabled medical computer vision. *npj Digital Medicine*, 2021.

[5] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. In *Conference on Computer and Communications Security (CCS)*, pages 1322–1333, 2015.

[6] Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon M. Lin, David Page, and Thomas Ristenpart. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In *USENIX Security Symposium*, pages 17–32, 2014.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2014.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[9] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint*, arXiv:2208.01626, 2022.

[10] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely Connected Convolutional Networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017.

[11] Ahmad Ibrahim, Hoda K. Mohamed, Ali Maher, and Baochang Zhang. A survey on human cancer categorization based on deep learning. *Frontiers in Artificial Intelligence*, 5, 2022.

[12] Mostafa Kahla, Si Chen, Hoang Anh Just, and Ruoxi Jia. Label-only model inversion attacks via boundary repulsion. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15025–15033, 2022.

[13] Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4401–4410, 2019.

[14] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep Learning Face Attributes in the Wild. In *International Conference on Computer Vision (ICCV)*, 2015.

[15] Shagufta Mehnaz, Sayanton V. Dibbo, Ehsanul Kabir, Ninghui Li, and Elisa Bertino. Are your sensitive attributes private? novel model inversion attribute inference attacks on classification models. In *USENIX Security Symposium*, pages 4579–4596, 2022.

[16] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. *arXiv preprint*, arXiv:2211.09794, 2022.

[17] Hongwei Ng and Stefan Winkler. A data-driven approach to cleaning large face datasets. In *IEEE International Conference on Image Processing (ICIP)*, pages 343–347, 2014.

[18] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022.

[19] Lukas Struppek, Dominik Hintersdorf, Antonio De Almeida Correia, Antonia Adler, and Kristian Kersting. Plug & play attacks: Towards robust and flexible model inversion attacks. In *International Conference on Machine Learning (ICML)*, volume 162, pages 20522–20545, 2022.

[20] Lukas Struppek, Dominik Hintersdorf, and Kristian Kersting. Be careful what you smooth for: Label smoothing can be a privacy shield but also a catalyst for model inversion attacks. In *International Conference on Learning Representations (ICLR)*, 2024.

[21] Kuan-Chieh Wang, Yan Fu, Ke Liand Ashish Khisti, Richard Zemel, and Alireza Makhzani. Variational Model Inversion Attacks. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[22] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *Computer Security Foundations Symposium (CSF)*, pages 268–282, 2018.

[23] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R. Manmatha, Mu Li, and Alexander Smola. Resnest: Split-attention networks. *arXiv preprint*, arXiv:2004.08955, 2020.

[24] Yuheng Zhang, Ruoxi Jia, Hengzhi Pei, Wenxiao Wang, Bo Li, and Dawn Song. The Secret Revealer: Generative Model-Inversion Attacks Against Deep Neural Networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 250–258, 2020.