IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 2, 2022, accepted February 28, 2022, date of publication March 8, 2022, date of current version March 18, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3157589

Privacy-Preserving Case-Based Explanations: Enabling Visual Interpretability by Protecting Privacy

HELENA MONTENEGRO[1,2], (Student Member, IEEE), WILSON SILVA[1,2], (Student Member, IEEE), ALEX GAUDIO[1,2,3], MATT FREDRIKSON[4], (Member, IEEE), ASIM SMAILAGIC[3,4], (Fellow, IEEE), AND JAIME S. CARDOSO[1,2], (Senior Member, IEEE)

[1]Faculty of Engineering, University of Porto, 4200-465 Porto, Portugal
[2]INESC TEC, 4200-465 Porto, Portugal
[3]Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[4]School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA

Corresponding author: Helena Montenegro (up201604184@edu.fe.up.pt)

This work was supported in part by the Project TAMI—Transparent Artificial Medical Intelligence through ERDF—European Regional Fund through the North Portugal Regional Operational Program—NORTE 2020 under Grant NORTE-01-0247-FEDER-045905, in part by the Portuguese Foundation for Science and Technology—FCT under Carnegie Mellon University (CMU)—Portugal International Partnership, and in part by FCT within Ph.D. under Grant SFRH/BD/139468/2018.

ABSTRACT Deep Learning achieves state-of-the-art results in many domains, yet its black-box nature limits its application to real-world contexts. An intuitive way to improve the interpretability of Deep Learning models is by explaining their decisions with similar cases. However, case-based explanations cannot be used in contexts where the data exposes personal identity, as they may compromise the privacy of individuals. In this work, we identify the main limitations and challenges in the anonymization of case-based explanations of image data through a survey on case-based interpretability and image anonymization methods. We empirically analyze the anonymization methods in regards to their capacity to remove personally identifiable information while preserving relevant semantic properties of the data. Through this analysis, we conclude that most privacy-preserving methods are not sufficiently good to be applied to case-based explanations. To promote research on this topic, we formalize the privacy protection of visual case-based explanations as a multi-objective problem to preserve privacy, intelligibility, and relevant explanatory evidence regarding a predictive task. We empirically verify the potential of interpretability saliency maps as qualitative evaluation tools for anonymization. Finally, we identify and propose new lines of research to guide future work in the generation of privacy-preserving case-based explanations.

INDEX TERMS Case-based interpretability, privacy-preserving machine learning, deep learning, computer vision.

I. INTRODUCTION

Deep Learning has led to significant advances in image analysis and has become state-of-the-art in most computer vision tasks. Some works even claim that the developed models can outperform human experts [1]. However, the lack of interpretability in Deep Learning makes it difficult to trust and use deep networks in several real-world applications, especially when wrong decisions have significant consequences [2] (e.g., in medicine and forensics).

The associate editor coordinating the review of this manuscript and approving it for publication was Charalambos Poullis.

In recent years, the research community has begun to recognize the importance of explainability for "black-box" machine learning models. When it comes to computer vision tasks, methods capable of generating visual explanations are of particular interest. We highlight two types of visual explanations: saliency maps and case-based explanations. Saliency maps show the regions of the images that are relevant to a model's decisions. Case-based explanations, or explanations-by-example, are data samples with similar task-related features as the observation under analysis.

Although commonly applied in domains with sensitive visual data, such as medical diagnosis, case-based

explanation methods may compromise the privacy of individuals, with this privacy leakage being of particular concern when data is shared with potentially unauthorized personnel (e.g., medical students, interns, patients, and family members). Privacy is less of an issue in saliency map methods, where, by design, the only sensitive information they reveal is the input image under consideration.

In image retrieval, case-based explainability is commonly used in scenarios such as medical image diagnosis to obtain examples of similar disease-matching images that can be compared to a case under analysis and provide additional insights to explain and support a diagnosis [3]–[5]. The retrieval process begins with a user entering an image into the retrieval system. Then, the retrieval system ranks the examples in its database according to a semantic similarity measure and presents the most similar examples to the user. Retrieving an image with sensitive identity information from a private dataset may violate the privacy of the individual present in the image. To address this issue, the case-based explanation must go through an anonymization process to wash the identity out of the image before presenting it to the consumer of the explanation, as illustrated in Figure 1. The greatest challenge in creating the washer model is to ensure that no identity is leaked in the privatized version of the explanation while preserving the explanatory evidence and realism.

The application of privacy-preserving methods to case-based explanations has rarely been addressed in the scientific literature. Only one privacy-preserving work [6] considers the characteristics of visual case-based explanations. As a guide for future research, we survey and analyze case-based interpretability methods, reflecting on their privacy needs, and privacy-preserving methods, highlighting their limitations when applied to case-based explanations. Furthermore, we formalize the generation of Privacy-Preserving Case-Based Explanations as a multi-objective problem to minimize identity information, while preserving realism and explanatory evidence in the images. Finally, we propose future research directions to guide future work in this under-explored field of research.

II. BACKGROUND ON DEEP GENERATIVE MODELS

In this section, we provide background on deep generative models, as they are used in several case-based interpretability and privacy-preserving methods.

Generative Models learn the probability distribution of a training dataset and can use it to generate new data samples. In the context of privacy-preserving methods, these models are applied to generate privatized images. Some case-based interpretability methods also incorporate generative models in the generation of the explanations. The most relevant generative models for this survey are Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs).

GANs [7] consist of two adversarial networks, a generator, and a discriminator, which compete against each other in a minimax game. The generator is responsible for generating new data samples, while the discriminator distinguishes between real and generated instances. The adversarial training promotes the generation of realistic images, as the generator tries to trick the discriminator into misclassifying the synthetic samples as real. The objective in Equation 1 describes the minimax game. The generator G minimizes the objective, while the discriminator D maximizes it, x represents real samples, and z represents the input to the GAN, often random noise. GANs are challenging to train, and the generator often undergoes mode collapse, a phenomenon describing an overall lack of diversity in the output.

$$\min_G \max_D E_{x\sim p(x)}\left[\log D(x)\right] + E_{z\sim p(z)}\left[\log(1 - D(G(z)))\right] \tag{1}$$

Privacy-preserving methods use conditional GANs (cGANs), a variation on the GAN restricted by predefined conditions. In specific, privacy-preserving GANs are conditioned by the input image, as they must preserve certain features during the anonymization.

VAEs [8] learn an approximation of the data distribution using two networks: an encoder and a decoder. The encoder maps samples $x \sim p_{\text{data}}(x)$ in the original data space into a latent space with a simpler data distribution (usually a Gaussian distribution). The decoder maps samples $z \sim p(z)$ from the latent space into the original data space. This network allows the generation of new images by sampling from $p(z)$ and converting the samples to the original data space through the decoder. The loss function used to train a VAE, represented in Equation 2, contains two terms: a reconstruction term and a regularization term. The reconstruction loss term approximates a reconstructed image obtained through the VAE to its original version and can be represented by loss functions like cross-entropy or mean squared error. The regularization loss term uses Kullback-Leibler (KL) Divergence to reduce the distance between the encoder's distribution $q_\theta(z \mid x)$ and the original data distribution $p(z \mid x)$.

$$L = -E_{x,z}[\log p(x \mid z) + D_{KL}(q_\theta(z \mid x)||p(z \mid x))] \tag{2}$$

In addition to the two Deep Generative Models presented, there are other models in the literature that have the potential to generate high-quality images, as is the case of Autoregressive Models and Normalizing Flows [9]. Research on these two models has been growing in recent years due to their ability to explicitly model the data distribution as a tractable distribution without the need for approximations. These models may also be relevant in the future development of privacy-preserving models. Nonetheless, they are not used in any of the works analyzed in this survey.

III. CASE-BASED INTERPRETABILITY

Case-based interpretability stands out among the various explainability strategies for its ability to generate intuitive and easy-to-understand explanations based on similar examples [10]–[12]. Case-based explanations are examples that resemble the image under analysis. The consumer of the

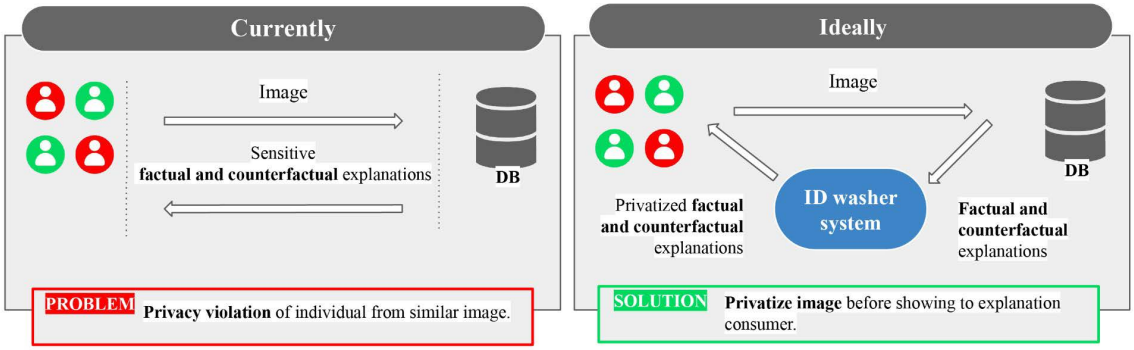


FIGURE 1. Diagram exemplifying the explanatory retrieval process. Consumers illustrated in red represent individuals who do not possess authorized access to the raw data (identity information) in the database. Consumers illustrated in green can access the raw data.

explanation can gain a deeper understanding of a model's decisions by comparing the explanation and the original image in the context of the model's prediction.

Explanations can be retrieved from the database used by the machine learning model or generated based on the data. Case-based interpretability methods can produce various types of explanations by example:

- **Similar examples**: the explanation is the most similar case from the training data whose prediction matches the case under analysis. Methods that retrieve this type of explanation define the similarity measure used to compare the data samples.
- **Typical examples**: the explanation is the case that best represents the prediction made for the case under analysis. These examples are often found in models that implement prototype learning.
- **Counterfactual examples**: the explanation is the most similar case whose prediction differs from the case under analysis. This type of example can also be generated based on the case under analysis, providing the alterations that the sample must suffer so that its prediction changes.
- **Semi-factual examples**: the explanation is a case that shares the same prediction as the case under analysis but closer to a decision boundary. These examples help convince the explanation consumer by showing that even if the data sample had changes commonly associated with another prediction, it would still be classified as is. These types of explanations can either be retrieved from the data or generated based on the original image.

The retrieval of case-based explanations in image data remains a challenge, due to the difficulty in defining a reliable metric for evaluating the similarity between two images with respect to a given task. This section introduces the current state-of-the-art case-based interpretability methods in traditional machine learning and deep learning. We differentiate the existing techniques in intrinsic and *post hoc* approaches. Intrinsic approaches involve the development of inherently interpretable models with case-based reasoning where the decisions are based on examples from the training data. *Post hoc* methods require the development of explanatory models to retrieve explanations for the base models' decisions.

A. TRADITIONAL MACHINE LEARNING METHODS

In traditional machine learning methods for case-based interpretability, the feature extraction process is separated from the decision process. To apply these methods to image data, we need to extract a vector of features, which will be used as the input to these models.

1) INTRINSIC METHODS

Intrinsically interpretable methods capable of retrieving case-based explanations typically fall under two types: distance-based methods or prototype-based methods. In distance-based methods, the data samples are compared according to a distance measure, and the closest ones can be retrieved as similar explanations. In prototype-based methods, the data is organized into clusters, and the explanations are prototypes from the cluster that best represents the original instance. In prototype-based reasoning, each prototype represents a cluster of data with certain characteristics, and the set of all prototypes should be representative of the whole training set.

The most well-known case-based method is the K-Nearest Neighbors (KNN) algorithm. It is a distance-based method that classifies an observation according to the K nearest training samples. The algorithm calculates the distance between the observation and the training samples according to a distance metric. Then, it classifies the observation with the majority of its neighbors' labels. This method can produce two types of explanations: similar examples corresponding to the neighbors that share the same label as the predicted one and counterfactual examples corresponding to the neighbors whose labels do not match the final prediction.

One example of a prototype-based method is the Bayesian Case Model (BCM) [13], which organizes the data in clusters represented by a prototype and a subspace of the features characteristic of each cluster. For classification, a new observation is mapped into a cluster and classified according to

its prototype, which is then used to explain the model's decision. As such, this model can produce typical examples as explanations.

2) POST HOC METHODS

Regarding *post hoc* interpretability techniques, traditional machine learning models without case-based reasoning can be used as a distance metric to retrieve similar examples [14]. For example, we can use a decision tree model, a rule-based model, to evaluate the similarity between two data samples. A decision tree takes the form of a tree where each node represents a feature in the data, and the edges are rules that apply to the features. To obtain a similar example, we can retrieve a training sample sorted into the same decision node as the observation in the tree. The explanation, in this case, would be an observation from the training data that conforms to the same rules as the observation under analysis and that shares the same class as the decision node. Furthermore, instances sorted into the same decision node but with a different class can be retrieved as counterfactual examples.

Post hoc methods can also be built upon interpretable models with case-based reasoning to improve the quality of the explanations. Nugent *et al.* [15] proposed a framework built upon the KNN algorithm called Explanation Oriented Retrieval (EOR). EOR aims to retrieve more convincing explanations by applying an explanation utility measure to re-order the nearest neighbors obtained using KNN. The neighbor with the highest explanatory value is the one retrieved as an explanation. In this case, the explanations are considered semi-factual examples, as they share the same classification as the original data instance but are closer to the classification task's decision boundary than the nearest neighbor.

B. DEEP LEARNING METHODS

Deep Learning methods can learn to extract features from the data, optimizing the feature extraction process according to a predictive task.

1) INTRINSIC METHODS

As in traditional machine learning, intrinsic methods can be prototype-based or distance-based. Existing prototype-based methods differ in the types of prototypes that are created in the model. In Deep Learning, prototypes can be training data instances, synthetic data generated based on the training data, and even parts of images.

The Explainable Deep Neural Network (xDNN) model [16] and its successor, Deep Machine Reasoning (DMR) [17], are examples of prototype-based methods. These methods define prototypes as local peaks in the data density and classify an observation according to the prototype that best represents it. The xDNN model extracts features from data into a latent space where it calculates each instance's probability distribution. The prototypes for each class are selected from the data instances with higher density. On inference, an instance is classified according to the prototype that best represents it. DMR proposes improvements to the xDNN network to deal with data imbalance. After the prototype selection process, DMR augments the data through linear interpolations between perturbed data points around a prototype. The inference process in the DMR network is also slightly different, as it includes a decision tree to compare the two most representative prototypes regarding minimum error during training. The instance is classified according to the prototype with the lowest minimum error, which can be used as an explanation by typical example.

The Prototype Classifier developed by Li *et al.* [18] is a prototype-based method where prototypes are generated instances that are similar or even identical to training samples. The model learns prototypes that best represent the training data and uses them to perform the classification task. It possesses an autoencoder, whose encoder extracts features from the data into a latent space, and whose decoder learns to map features from the latent space into the original data space. Following the autoencoder, the model possesses a prototype classifier, with a prototype layer that learns prototypes based on the training data's latent representations, and a decision layer that classifies a sample based on the prototypes. The prototypes learned by the network can be visualized as explanations through the decoder. Additionally, we can use the autoencoder's latent space to calculate the distance between a sample's representation and the prototypes to obtain the most similar prototype and use it as an explanation by typical example for the model's decision.

One prototype-based method where the prototypes are parts of images is the Prototypical Part Network (ProtoPNet) [19]. During training, the network creates a latent space where image patches relevant to an image's classification are represented in clusters. The clusters contain semantically similar patches whose images belong to the same class. On inference, the network finds prototypes similar to parts of an image and combines the respective similarity scores to make a prediction. The prototypes are provided as typical examples.

One example of a distance-based method is the Deep k-Nearest Neighbors (DkNN) [20]. This method computes an instance's neighbors at each layer in the model. The labels of the neighbors are then used to make a prediction. By using the neighbors at each layer, the model guarantees that the prediction is consistent across the whole model, enhancing robustness. In this method, it is possible to retrieve an instance's neighbors as similar examples to explain the prediction.

2) POST HOC METHODS

Regarding *post hoc* interpretability approaches, the methods can use an interpretable surrogate model to retrieve explanations from the base model or use the "black box" model to measure similarity between data instances and retrieve the most similar ones. One example of a method that can easily be used as a similarity measure for explanation retrieval is Concept Whitening [21]. Concept Whitening organizes a classification network's latent space according to pre-defined

high-level concepts. The network is trained with two sets of labels, one for the instances' classification task and the other with the concepts associated with the data. We can use this network's latent space to measure the distance between the new data instance and the training data and obtain similar examples to offer as explanations to an explanation consumer. Additionally, the distance between the data can be measured according to a specific concept or set of relevant concepts for the classification task.

Regarding *post hoc* methods specifically developed for image retrieval, one method is Interpretability-guided Content-Based Image Retrieval (IG-CBIR) [4]. In this work, the authors propose a new approach to improve the explanatory retrieval process by enhancing the semantic similarity measure between images using interpretability saliency maps as an attention mechanism to focus on image regions related to the classification task.

One type of interpretable surrogate model that can be used to retrieve explanations is unsupervised clustering [22]. This method retrieves layer activations at each layer in the base model and encodes them into a latent space where it is possible to measure the distance between the activations. In this latent space, the data is organized into clusters using Euclidean distance. During inference, at each layer from the base model, the observation is mapped into a cluster and associated with the respective centroid. Finally, the model assigns to the observation the weighted average of its centroids' labels. The centroids can be used as explanations by typical examples.

Another interpretable surrogate model that can be used in the explanation retrieval process is KNN, as suggested in the Twin Systems framework [23]. This framework consists of extracting features from the data using the base classification model and applying the KNN algorithm over these features to retrieve similar examples as explanations. The framework explores several ways to extract features from visual data, including perturbation-based methods, sensitivity analysis methods, and interpretability saliency maps.

Finally, counterfactual explanations are usually generated based on the original observation. The generation process implies making the least possible changes to an image to change its classification. Additionally, for the explanation to be plausible, a human must be able to detect the differences between the original observation and the counterfactual, as argued by Kenny and Keane [24]. The authors proposed a method to generate counterfactual explanations, called PlausIble Exceptionality-based Contrastive Explanations (PIECE) [24]. This method first identifies the target counterfactual class. To generate counterfactuals, the method finds features in the original image whose probability of occurring in the target counterfactual class is low. Then, using a Generative Adversarial Network (GAN), the method modifies these features into their expected values in the target class. This method can also be used to generate semi-factual explanations by stopping the generative model's training before the class of the synthetic images changes to the counterfactual class.

C. DISCUSSION

Deep Learning, in comparison with traditional Machine Learning, holds the advantage for image analysis tasks, as it learns to automatically extract features from the data according to the target task.

In intrinsic methods, the retrieved explanations are directly incorporated into the decision-making process and thus constitute an accurate representation of the model's reasoning. On the other hand, explanations obtained through *post hoc* methods are often criticized for not representing the model's real reasoning [18]. Nonetheless, without the restriction of interpretability, models may achieve better results, highlighting the usefulness of *post hoc* techniques. As such, when considering which method to implement, it is important to decide whether the priority is the model's interpretability or its performance. The ideal model is an intrinsically interpretable model capable of achieving high-quality results.

As for the types of explanations produced by the models, a factual explanation by a similar or typical example, by itself, is not sufficient to understand a model's predictions, as it only allows to introspect the characteristics associated with one class. Counterfactual explanations aid interpretability by highlighting features that are usually seen in the remaining classes, clarifying the decision boundary between two different classes. When providing explanations, it is relevant to consider mechanisms to obtain both factual and counterfactual examples in the proposed interpretability methods.

When case-based explanations expose the identity of an individual whose images are not accessible by the consumer of the explanation, the explanations pose a significant privacy threat. This issue is relevant in explanations retrieved from a database, exposing the identity of a data subject. In counterfactual and semi-factual explanations generated by making alterations to the original image, privacy is of no concern as the explanation consumers already have access to the original image.

The most critical aspects of case-based explanations, which need to be considered during anonymization, are intelligibility, explanatory evidence, and privacy. The explanations must be intelligible so that a human can understand them. They must not leak the identity of a subject to be applied in real-world scenarios, guaranteeing the fundamental human right to privacy. Finally, the explanations must contain relevant characteristics that allow an explanation consumer to compare them to the case being analyzed and obtain a deeper understanding of a model's decisions.

IV. PRIVACY-PRESERVING METHODS FOR VISUAL DATA

Privacy-preserving approaches are essential in case-based interpretability to apply the respective visual explanations in the real world. Given an image showing a subject's identity, the privatization goal is to prevent the subject's recognition

by a human or an identity recognition network while preserving semantically relevant features that allow the use of the resulting image as a meaningful explanation. In visual data, semantically significant features are often tangled with features that portray identity. The greatest hurdle in the generation of privacy-preserving images is to manage the trade-off between privacy, explanatory evidence, and intelligibility.

In this section, we discuss privacy-preserving methods considering their application to case-based explanations. The methods are analyzed in regards to their capacity to preserve intelligibility, privacy, and explanatory evidence. Methods such as encryption, whose results are unintelligible, will be ignored.

We discriminate the current privacy-preserving methods into two groups: traditional and deep learning methods. Traditional methods are applied over the whole input, requiring an additional step to identify the parts of the images to be privatized. In contrast, deep learning methods are capable of identifying the parts that leak identity and anonymizing them. Furthermore, deep learning methods can also identify relevant explanatory features in images.

A. TRADITIONAL METHODS

Traditional methods privatize the whole input image. They do not have the capacity to evaluate which parts of the images transmit identity or even explanatory evidence. As such, they may lead to the unnecessary loss of semantically significant features that do not expose identity.

1) FILTER-BASED METHODS

Filter-based methods apply filters such as pixelation or blur to the data. Pixelation consists of dividing an image into a grid and assigning to each grid cell the average value of the pixels inside it. With blurring, the image's pixels are altered according to a kernel applied over the region surrounding each pixel. These methods suffer the most from a trade-off between privacy and explanatory evidence since both identity features and explanatory features get equally distorted [25]. Blurring images can preserve privacy, but it leads to a significant loss of intelligibility. Montenegro *et al.* [25] support this claim using eye images from the Warsaw-BioBase-Disease-Iris v2.1 [26], [27], which were classified according to the presence or absence of glaucoma. We present an example in Figure 2 from a blurred dataset where a multi-class identity recognition network was capable of recognizing the original identity with the relatively high accuracy of 23.24%, in a dataset with images from 115 different subjects, where random guessing would lead to an accuracy of ≈0.88%. Blurring did not preserve privacy to a satisfying degree, and the image quality was significantly damaged. Detailed results regarding blurring (and other privacy-preserving methods) are available in Table 1.

2) K-SAME-BASED METHODS

K-Same [28] is an algorithm for privatization originally proposed to de-identify face images. The algorithm finds clusters

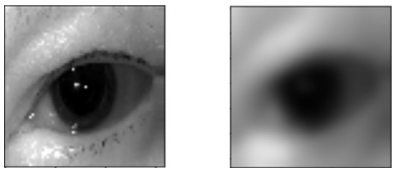


FIGURE 2. Example of results of blurring images to privatize them. The images represent the original image and its privatized version (taken from the work of Montenegro *et al.* [25]), respectively.

of K most similar images. The images are then replaced by the pixel-wise cluster mean. In the original paper, image similarity is measured by the pixel-wise distance between the images. As this method does not consider the preservation of semantic features that are relevant for a certain task, Gross *et al.* [29] proposed K-Same-Select. This method separates the data into subsets according to a utility function and then applies K-Same to each subset. When there is only one instance per person in the dataset, these methods ensure K-anonymity, *i.e.,* an identity is recognized with probability at most $\frac{1}{K}$. When the same person appears in multiple samples in the dataset, the respective images may be organized into the same cluster, thus violating K-anonymity. Averaging different images sacrifices higher detail features from the original image that are relevant for the explanation [25]. Figure 3 shows an example image obtained with K-Same-Select. This method can generate intelligible images for low values of K. Although there is also a visible privacy-intelligibility trade-off using this method, it is not as significant as in the filter-based methods, since the images look more intelligible and guarantee a higher degree of privacy, with the lower identity recognition accuracy values of 2.94% and 1.47% for $K = 6$ and $K = 9$, respectively. Furthermore, the accuracy of the identity recognition network at identifying any identity used in the privatized images is 14.41% for both values of K, which is better than the accuracy obtained using blurring (23.24%) [25]. Further results are available in Table 1.

FIGURE 3. Example of results of using the K-Same-Select method. The images represent the original image and its privatized versions with K = 6 and K = 9 (taken from the work of Montenegro *et al.* [25]).

3) MODEL-BASED METHODS

Model-based methods replace image parts that transmit identity by a model. One example is the face-swapping method developed by Bitouk *et al.* [30]. This method estimates the pose of a face in an image and replaces it with a similar model face from a public library. In this method, the task used to obtain explanations could be used to evaluate the semantic similarity between the replacement models and the original

image, ensuring the preservation of semantic features. The biggest drawback in this method is the need to have public models of the image parts we want to privatize. Furthermore, the explanatory features tangled with the identity-related features may be lost, as they are located in image regions replaced by the model.

B. DEEP LEARNING METHODS

Deep learning privatization approaches generally use a deep generative network to produce the privatized images. The privatization process is guided by an identity recognition network whose loss is backpropagated into the generative network to remove the identity from the generated image. Regarding the preservation of explanatory evidence, some models preserve semantic features relevant to a task. These task-dependent models use the task-related model to guide the generative process by backpropagating its loss. Task-independent models focus on preserving general image features to preserve the image's utility.

One of the task-dependent methods proposed in the literature is the Privacy-Preserving Representation Learning Variational Generative Adversarial Network (PPRL-VGAN) [31]. This model performs privatization through identity replacement. It comprises a GAN with a conditional VAE as the generator and a multi-task discriminator. The VAE receives as a condition the target identity, which will be used to replace the identity in the input image. The multi-task discriminator possesses a real/fake classifier to guarantee realism in the privatized images, a multi-class identity recognition network to aid the identity replacement process, and a task-related classifier to preserve the utility of the privatized images. As highlighted in the work of Montenegro *et al.* [25], this method threatens the privacy of the subject used as a replacement, as revealed by the high identity recognition accuracy obtained when trying to recognize the identity used as a replacement in the privatized images (Table 1). This network could only be applied to replace the identity in the images if there was a predefined model that does not expose anyone's identity.

Regarding the preservation of explanatory evidence, this method only guarantees the preservation of the class of the original image and not of its exact task-related semantic features. To show an example evidencing this claim, we apply the PPRL-VGAN model to privatize an image using its own identity as the replacement identity. We conduct this experiment with a facial expression recognition dataset used in the original work: FERG dataset [32], composed of images from 6 different identities and seven different facial expressions. The example shown in Figure 4 expresses the facial expression: joy. To guarantee the preservation of semantic features, we expected to obtain an image very similar to the original one, presenting the exact same expression-related features, with an open mouth showing teeth and slightly closed eyes. Instead, we obtain a privatized image representing the same facial expression as the original one but containing different semantic features: closed mouth and more open eyes.

FIGURE 4. Example of results obtained from applying the PPRL-VGAN network to replace the identity in the original image by itself. The first image represents the original image and the second one represents its synthetic version.

Furthermore, the PPRL-VGAN model was only validated with datasets for facial expression recognition (its target task), where each subject contains images for all the different classes. However, when it comes to contexts where each subject only possesses images from one task-related class, the model fails to perform the feature disentanglement process. This claim was empirically demonstrated by Montenegro *et al.* [25], who applied the model to medical and biometric data for glaucoma detection and verified that the glaucoma recognition accuracy significantly drops when using replacement identities whose pathology differs from the original image (accuracy of 65.06%) as opposed to replacement identities that share the same pathology as the original image (accuracy of 85.56%). The range of application of the model is further diminished by the use of a multi-class identity recognition model, which is unfeasible to train for datasets with a reduced number of images per subject, often seen in real contexts like in the medical scene.

One privacy-preserving approach which addresses the weaknesses of the PPRL-VGAN network is introduced in the work of Montenegro *et al.* [6], where the authors propose two privacy-preserving models, one using a multi-class identity recognition network, which we will call PP-MIR in this review, and the other using a Siamese identity recognition network, which will be called PP-SIR. As shown in Figure 5, both models contain a generative module, composed of a WGAN-GP network [33], responsible for the generation of intelligible images. Both models also explicitly preserve explanatory evidence by using interpretability saliency maps to reconstruct the relevant explanatory features in the privatized images. The two networks differ in regards to the privacy mechanism. PP-MIR promotes privacy for the entire dataset by approximating the multi-class identity recognition to random guessing, promoting a uniform identity distribution as the privatized image's prediction. The PP-SIR model was proposed to widen the range of application of the privacy-preserving model in the medical scene through a Siamese identity recognition network, which computes the distance between two images in regards to identity features. The PP-SIR network privatizes an image by increasing its identity-related distance to its privatized version.

TABLE 1. Results of some traditional and deep learning privacy-preserving methods taken from the work of Montenegro *et al.* [25]. The table highlights in bold the best results obtained for each metric in each method.

Experiment	Dataset	Identity Recognition Accuracy	Replacement Identity Recognition Accuracy	Glaucoma Recognition Accuracy	Glaucoma Recognition F1-Score
Baseline	Original test set	90.00%	-	93.24%	87.83%
PPRL-VGAN	Privatized set w/ random identities	**0.50%**	74.68%	79.18%	62.22%
	Privatized set w/ identities w/ the same pathologies	1.76%	78.35%	**86.56%**	76.01%
	Privatized set w/ identities w/ different pathologies	0.71%	60.26%	65.06%	48.30%
	Averaged privatized set	2.56%	**14.35%**	86.24%	**78.80%**
Blurring	Privatized set with kernel size 3	69.41%	-	**93.24%**	**87.57%**
	Privatized set with kernel size 9	31.76%	-	88.82%	77.11%
	Privatized set with kernel size 15	**23.24%**	-	81.47%	55.32%
K-Same-Select	Privatized set with 3 identities	7.06%	22.94%	**82.35%**	**61.54%**
	Privatized set with 6 identities	2.94%	**14.41%**	81.76%	53.73%
	Privatized set with 9 identities	**1.47%**	**14.41%**	78.53%	42.52%

Furthermore, the network guarantees privacy for the whole dataset by increasing the identity-related distance between a privatized image and an image from each subject in the dataset.

The most significant weakness in these models when considering their application to the anonymization of case-based explanations obtained from intrinsic interpretability methods is that they use a *post hoc* interpretability method to preserve explanatory evidence. The use of *post hoc* methods, which are criticized for not reflecting a model's real reasoning, clashes with the intrinsic methods' goal of producing accurate explanations of a model's predictions, as noted by the original authors [6]. One more weakness is that, although PP-SIR was proposed to apply to data with very few images per identity, it has not yet been validated in that same scenario, as the dataset in which the authors validate their approach contains enough data to train a multi-class identity recognition network. Finally, despite privatized images obtained through these methods being intelligible, their quality should be improved, especially in the PP-MIR model. An example of images from these methods is shown in Figure 6.

Other types of privacy-preserving methods in the literature are independent of a classification task, aiming to preserve general features to guarantee an image's utility. Since these methods do not explicitly preserve task-related features, they may fail to guarantee that the privatized image contains the relevant explanatory evidence needed for an explanation consumer to understand the explanation. Some task-independent methods disentangle identity features from the remaining image features. These methods can obtain a vector of identity features that can be modified to privatize the image.

One example of such a method is CLEANIR [34]. CLEANIR is a Variational Autoencoder (VAE) applied to de-identify face images. The encoder is trained to explicitly disentangle identity features from the image's utility features, producing two vectors in its latent space. The decoder maps these vectors into an image in the original data space. During training, the network is trained with a reconstruction loss, forcing the decoder to learn how to obtain the original image based on its latent vectors. The encoder learns to disentangle identity features from the remaining ones through an embedding loss which approximates the latent vector of identity features to identity embeddings obtained from applying a pre-trained facial embedding extractor to the original image. On inference, the network alters the latent vector of identity features, resulting in a privatized image. One strength of this method is that its privatization process is independent of the dataset, guaranteeing privacy for the subjects in the training set. However, this method does not guarantee that the latent vector of remaining features does not contain information that leaks identity. For instance, if the facial embeddings capture most or all the information needed to reconstruct the image, then this utility-related latent vector could be correlated or even equal to the identity-related latent vector.

Replacing and restoring variational autoencoders (R^2VAE) [35] is another method that disentangles identity features from the remaining ones. The network comprises a GAN with a VAE as the generator. The VAE contains two encoders, where one extracts identity-related features, and the other one extracts features independent from identity. The decoder maps the latent representations obtained with the encoders into an image in the original data space. The network also possesses a discriminator to aid the generation of realistic images and an identity recognition network to aid feature disentanglement. During training, the generator receives a pair of images (A, B) and outputs an image with identity-related features from the input image A and identity-independent features from B. The network is trained to output a realistic image through a discriminator, to preserve the identity features of the input image A through the identity recognition network, and to preserve features unrelated to identity through a reconstruction loss between B and the output image. Furthermore, to aid feature disentanglement, the output image is then given to the generator along with the image B, obtaining the original image B, which contains identity features from B and general features from the output image. On inference, the network uses the encoder that extracts identity-independent features from the original image, to preserve general features, and uses an identity vector obtained from averaging several identity vectors from

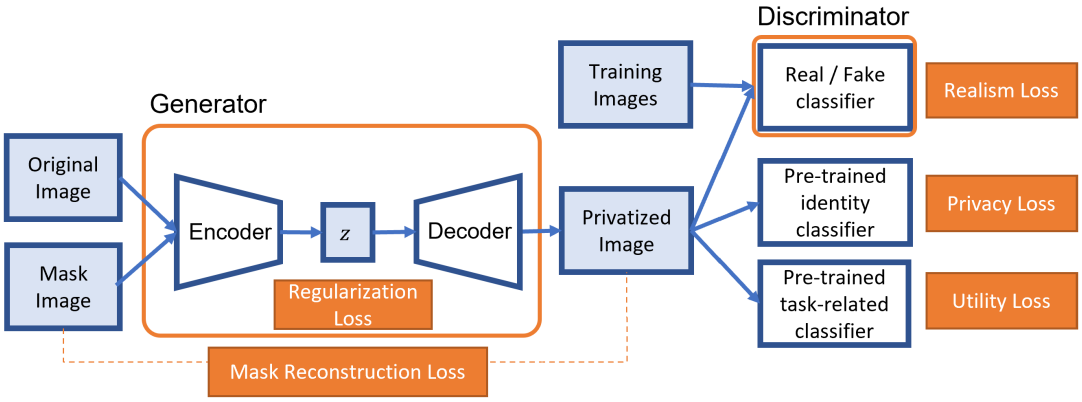


FIGURE 5. Architecture of the PP-SIR and PP-MIR models [6], highlighting the loss terms used during the models' training. The two models differ in the architecture of the identity classifier used.

FIGURE 6. Example of results obtained from the PP-SIR and PP-MIR models, taken from the work of Montenegro *et al.* [6]. The figures represent the original image and its privatized versions with the PP-MIR model and the PP-SIR model, respectively.

the dataset, to obtain a privatized image. One weakness of the method is that it uses identities from the dataset to privatize the images, threatening the privacy of the training data's subjects. However, this problem is easily fixed by applying an operation over the identity features obtained with the VAE to privatize them, like in CLEANIR.

Methods that do not explicitly perform feature disentanglement use a Siamese identity recognition network to ensure that the privatized image does not look like the original one in regards to identity. There are two examples of such networks in the literature: Privacy-Protective GAN (PP-GAN) [36] and Siamese Generative Adversarial Privatizer (SGAP) [37]. Both networks contain a GAN with a UNET generator and a Siamese identity recognition network which outputs the identity-related distance between two images. In the loss function, these methods use the Structural Similarity Index Measure (SSIM) to ensure the preservation of general features and a contrastive loss to increase the identity-related distance between the original image and its privatized version. Regarding the networks' results, while PP-GAN results in high-quality privatized images, SGAP severely lacks quality, putting the utility of the resulting images at risk. One problem common to both networks is that they only use the Siamese network to guarantee that the resulting image does not expose the identity from the original image. They do not make any guarantees regarding the privacy of the subjects from the training data. As a result, the GAN could learn to generate images that are very similar to the training data, exposing the identity of subjects from the data.

To exemplify the issues that disqualify the PP-GAN and SGAP models as candidates to preserve privacy for case-based explanations, we performed an experiment with a privacy-preserving model that uses SSIM loss to guarantee the preservation of utility features and a Siamese network to preserve privacy by comparing the original and privatized images in regards to identity. In specific, we altered the previously introduced PP-SIR model to only guarantee privacy for the original subject and replaced the loss functions that guarantee the preservation of task-related features by SSIM. The privacy-preserving model that results from applying the mentioned alterations to PP-SIR is a GAN that generates a privatized image based on the input image I from the original data space's distribution p_d. The GAN is composed of a discriminator D which is trained using Wasserstein loss and gradient penalty [33]. The generator G, whose loss function is shown in Equation 3, is a VAE trained to maximize the identity-related Euclidean distance ED between an original image and its privatized version using a contrastive loss and to preserve utility features by using SSIM to maximize the similarity between the original and privatized images. In this equation, λ_x represents the parameters used to assign a degree of relevance to each loss term x.

$$\begin{aligned}\mathcal{L}_G = E_{(I)\sim p_d(I)}[&-\lambda_1 D(G(I)) + \lambda_2[\max(0, m - ED(I, G(I)))]^2 \\ &+\lambda_3 \frac{1 - \mathcal{L}_{SSIM}(I, G(I))}{2} \\ &+\lambda_4 KL(q(f(I) \mid I) || p(f(I)))]\end{aligned} \tag{3}$$

The similarity between two images x and y according to SSIM takes into account the images' structure, luminance, and contrast [38]. It is calculated according to Equation 4, where μ represents an image's mean intensity, σ represents the standard deviation used to estimate contrast, and C_1 and C_2 are constants to avoid instability.

$$\mathcal{L}_{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1) + (2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{4}$$

This experiment uses the previously introduced dataset Warsaw-BioBase-Disease-Iris v2.1 [26], [27], which is also used by the authors of the PP-SIR model. To confirm the hypothesis that simply distancing the privatized image from the original one in regards to identity is not enough to guarantee privacy for the entire dataset, we verified that some images generated by the privacy-preserving framework are similar to images from different identities available in the database, as exemplified in Figure 7. In the case of the PP-GAN model, while its results show that the network preserves the general facial structure, it seems to alter features like the eyes, nose, and mouth, for example. As some of these features, like the eyes, can also portray identity, there is a need to verify that these features do not resemble features from images in the database. The solution to this problem introduced in the PP-SIR model is to increase the identity-related distance between the privatized images and images from each subject in the dataset during training [6].

FIGURE 7. Example of results obtained from applying the PP-SIR network to only increase the distance between the privatized image and its original version. The first image represents the original image and the second one represents its privatized version. The third image is a data instance from the subject that was identified by a multi-class identity recognition network when analyzing the privatized image.

Regarding the preservation of explanatory evidence, we verified that images generated with the model using SSIM as the loss function to preserve general utility features possess a very low glaucoma recognition accuracy of 64.71% and an F1-score of 28.57%. These results prove that SSIM does not guarantee the preservation of the relevant features needed for a specific task. Therefore, the task-independent methods PP-GAN and SGAP do not fulfill two of the fundamental requirements of privacy-preserving case-based explanations: privacy for all subjects and preservation of explanatory evidence.

In general, the task-independent methods were validated on datasets for face recognition and evaluated on their capacity to preserve general features relative to age, ethnicity, face pose, among others. The only method that was validated using a different biometric dataset (fingerprints) was SGAP. Nonetheless, they do not guarantee the preservation of explanatory features according to a specific task. As such, these methods cannot be applied to privatize case-based explanations.

C. DISCUSSION

Comparing traditional methods with Deep Learning methods, the latter holds the advantage of being able to find identity-related features and even disentangle them from the remaining features, allowing to preserve features that are independent of identity.

Deep Learning methods are organized into task-dependent and task-independent methods. The limitation of models independent of a data mining task is that they might discard semantic features that are tangled with identity features. As they do not consider the preservation of relevant explanatory features needed to achieve a certain task, these methods are not appropriate to privatize case-based explanations. Regarding task-dependent methods, only one method [6] has been proposed specifically to privatize case-based explanations, guaranteeing the preservation of explanatory evidence. Nonetheless, the method uses *post hoc* interpretability methods to preserve explanatory evidence, which may be incompatible with explanations obtained through intrinsically interpretable approaches. Furthermore, the obtained results lack image quality.

We compare the existing privacy-preserving methods in Table 2. This table considers that methods that directly use data instances from the training data in the privatization process do not guarantee privacy for all data subjects, even if these methods guarantee K-Anonymity.

In the deep learning privacy-preserving approaches, two types of identity recognition networks were used to guide the privatization process: multi-class identity recognition and Siamese identity recognition. The use of a multi-class identity recognition model limits the range of application of the privacy-preserving model, as it is unfeasible to train this type of network in domains where the data contains few images per subject, frequently seen in the medical scene. A Siamese identity recognition network is easier to train in this context.

The most significant challenge in the privatization of case-based explanations is to manage the trade-off between privacy, explanatory evidence, and intelligibility, as improving one of these dimensions usually leads to worsening the others [6].

In this section, we conclude that research regarding privacy-preserving models for case-based interpretability is lacking, as only one approach considers the requirements of case-based explanations in the anonymization process.

V. PRIVACY-PRESERVING CASE-BASED EXPLANATIONS

Having analyzed the existing literature on case-based interpretability and privacy-preserving methodologies for visual data, we can now introduce and reflect on the novel paradigm of Privacy-Preserving Case-Based Explanations. In this section, we formalize the objective of privacy-preserving methods for case-based explanations and discuss its use in the evaluation of privacy-preserving methodologies.

Given an input image to be explained, the multi-task objective of privacy-preserving case-based explanations considers three dimensions: privacy, realism, and explanatory evidence. The resulting explanation, a privatized image, should simultaneously conceal sensitive information, remain as realistic as possible to enable intelligibility, and preserve information relevant to the primary task. The privatized image explanation

TABLE 2. Comparison between privacy-preserving methods.

Privacy-preserving model	Preserves original image's class	Preserves original image's exact semantic features	Guarantees privacy for all data subjects	Generates high-quality images	Applicable to data with few images per subject
Blurring			×		×
K-Same [28]				×	×
K-Same-Select [29]	×			×	×
Face Swapping [30]			×	×	×
CLEANIR [34]			×	×	
R^2VAE [35]				×	
PP-GAN [36]				×	×
SGAP [37]					×
PPRL-VGAN [31]	×			×	
PP-MIR [6]	×	×	×		
PP-SIR [6]	×	×	×	×	×

encodes features relevant to each of the three dimensions, and we suggest a saliency method to understand how the explanation represents the three dimensions. For the realism and explanatory evidence dimensions, feature saliency can be measured via the respective classifiers that discriminate between realism and the predictive tasks. We next consider features of the privacy dimension.

The concept of privacy is intimately linked to the concept of identity. All privacy-preserving networks in deep learning literature use identity recognition networks to privatize the input. Thus, we clarify the relationship between privacy and identity features. Privacy is the non-disclosure of an identity of a real person. Identity features are utilized by an identity recognition network to identify an image as belonging to a particular class (e.g., a person). To maximize privacy in an image, we need to guarantee that all its features do not leak any identity information available in the training data. Privacy-preserving features can be partitioned into two subsets: identity-independent features and identity-related features that do not leak a real person's identity. Identity-independent features are not used in the identity recognition process and, therefore, do not leak identity. Promoting that all image features become identity-independent is sufficient to preserve privacy. However, ensuring that all features are identity-independent becomes unfeasible when we consider that the privatized image must look real. In a realistic-looking image, there would inevitably be identity features that an identity recognition network could extract to attempt to recognize a subject. In this case, the goal of privacy preservation is to synthesize a set of identity-related features sufficiently different from any identity available in the training data to guarantee that no identity belonging to a real person is leaked. In summary, the maximization of privacy should result in a privatized image where the features would either be unrelated to the identity recognition task or relevant for the task but sufficiently different from existing identities as not to leak any subject's identity.

To formalize the objective of privacy-preserving case-based explanations, we consider the existence of three loss functions that evaluate privacy $\mathcal{L}_p$, realism $\mathcal{L}_r$ and preservation of explanatory evidence $\mathcal{L}_d$ in a privatized image. The loss functions can be defined as functions of corresponding oracle recognition networks ($D_{id}(x)$, $D_r(x)$ and $D_d(x)$) applied to an image I with respect to the three tasks: identity, realism, and detection. Considering a generative model G that, from an image I, outputs its privatized version $G(I)$, the objective comprises the optimization of the generative model's parameters in regards to the minimization of the three losses, as specified in Equation 5. In this equation, λ_i is a non-negative parameter that controls the relative degree of importance assigned to each task i.

$$\min_G[\lambda_p\mathcal{L}_p(G(I)) + \lambda_r\mathcal{L}_r(G(I)) + \lambda_d\mathcal{L}_d(G(I))] \tag{5}$$

Each loss function will be application-dependent. For example, in the work of Montenegro *et al.* [6], two different privatization methods were considered, and consequently, two different privatization loss functions were used. The PP-SIR model defines the privacy loss as the maximization of the identity-related Euclidean distance between a privatized image and the source image and between the privatized image and images from N identities in the training data, as shown in Equation 6. The PP-MIR model promotes privacy by approximating a uniform distribution U over the identities in the training data, using a multi-class recognition network D_{id}, as expressed in the privacy loss function in Equation 7.

$$\mathcal{L}_p = E_{(I)\sim p_d(I)}[\lambda_1[\max(0, m - ED(I, G(I)))]^2 + \lambda_2\sum_{i=0}^{N}\frac{[\max(0, m - ED(G(I), I_N))]^2}{N}] \tag{6}$$

$$\mathcal{L}_p = E_{(I)\sim p_d(I)}[-D_{id}(G(I))\log(U)] \tag{7}$$

A. QUALITATIVE EVALUATION OF PRIVACY-PRESERVING CASE-BASED EXPLANATIONS USING SALIENCY MAPS

The evaluation of privacy-preserving case-based explanations is a complex problem that requires examining and comparing the original explanations and their anonymized versions. In this section, we investigate the potential of using interpretability saliency maps as a qualitative measure to verify the preservation of privacy, realism, and explanatory evidence in the anonymized explanations.

The multi-task nature of the objective in Equation 5 exposes the interplay between identity, realism and detection. Privatizing images with the generative model $G(I)$ should not remove discriminative information or make images seem less realistic, nor should the generator introduce realistic or identity-preserving features that contribute to the correct discriminative prediction for the wrong reason. We analyze these underlying phenomena through a qualitative visual analysis of a set of saliency attribution maps. Our analysis below shows how the generative model correctly hides identifying features of an individual's eye while preserving realism.

Saliency attribution maps of a predictive model's input image identify which pixels contribute most to a model's prediction. We visualize nine saliency maps in Figure 8, where Figure 8a does not use the privatization generator $G(I)$ and where Figures 8b and 8c do. Within each subfigure, we compute three saliency maps, one for each of the three discriminator tasks required in Equation 5: detection, identity and realism. The top row is the saliency map, with positive (yellow) and negative (blue) values. The bottom row overlays the magnitude of the saliency on the input image. We adopt the SmoothGrad saliency method [39]. We define the saliency method below, where I is an input image, $I' = G(I)$ is a privatized image output by the generative model, $\mathcal{N}(0, \sigma)$ is Gaussian noise, $D_x : \mathcal{I} \rightarrow \mathbb{R}$ is one of the three discriminators in the PPCE objective $D_x \in \{D_d, D_p, D_{id}\}$, and $s_x(I)$ visualized saliency attribution maps corresponding to $x \in \{d, p, \text{id}\}$. Note that the discriminator input may be either an image I or a privatized image I'. We use G and D_x from the PP-SIR model and the PP-MIR model in Figure 8b and Figure 8c, respectively. For our visuals, we choose $N = 50$ and, following the standard SmoothGrad procedure, we clip extreme gradient values using percentile thresholds of 1% and 99%. Equations 8 and 9 describe the method applied to obtain the saliency maps for the original images and the anonymized images, respectively.

$$2s_x(I) = \frac{1}{N}\sum_{i=1}^{N}(I \mathrm{d} D_x(I + \mathcal{N}(0, 1))I) \tag{8}$$

$$s_x(I) = \frac{1}{N}\sum_{i=1}^{N}(I \mathrm{d} D_x(G(I))I) \tag{9}$$

The visualized saliency attribution maps in Figure 8 qualitatively show how privatization works. Compare the most salient identity features in the non-privatized model of Figure 8a to the corresponding identity maps in the privatized models of Figures 8b and 8c. We observe the privatized models obscure the identity features, presenting a more diffuse, randomized identity saliency map. In all three cases, the realism saliency maps tend to emphasize areas unrelated to glaucoma detection, and we observe a stronger emphasis on the skin, eyebrow, and right edge of the image. Skin features should be a salient component of a realistic image, and our observations confirm this intuition. The detection saliency maps vary between the three models, though they emphasize the pixels corresponding to the cornea and tear duct. We observe that the privatization models force the glaucoma detection model to work harder.

The qualitative saliency-based analysis of the privatization model provides a useful framework to validate and confirm the effects of privatization. We observe that privatization works. It obfuscates the identity features, preserves realism features and finds detection features that fit the privacy-preserving and realism constraints.

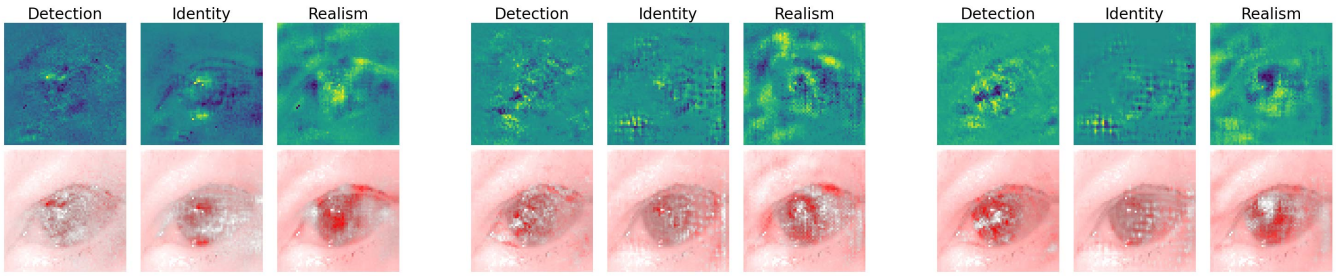
VI. DISCUSSION AND FUTURE WORK

Literature on privacy-preserving explanation methodologies is lacking. Our survey of privacy-preserving methodologies identified only one work [6] capable of privatizing images for case-based explanations (utilizing the PP-SIR and PP-MIR models). Nonetheless, there are still issues to address in the design of an ID washer model that can be applied to a wider range of image retrieval systems.

The literature lacks a case-based interpretability method that considers the goals of privacy when retrieving explanations from a database. For instance, consider two images from a database (A, B) and a third image C under analysis, where a case-based interpretability method considers A to be the best explanation for C's prediction. In this scenario, the privatized version of B may be a better explanation than the privatized version of A, depending on the privacy-preserving model's capacity to preserve explanatory evidence while removing identity. It is relevant to consider privacy during image retrieval rather than in the *post hoc* manner it has been considered so far.

With this reflection, we can establish a taxonomy to categorize current and future privacy-preserving models for case-based explanations: *post hoc* methods and intrinsic methods. *Post hoc* privacy-preserving methods are applied to privatize an explanation after it has been obtained, independently from the image retrieval process. Intrinsic methods integrate privacy directly in the image retrieval process. Currently, there are no intrinsic privacy-preserving methods for case-based interpretability. By considering privacy during the image retrieval, intrinsic approaches would increase the explanatory value of the privatized explanation given to the explanation consumers. On the other hand, *post hoc* approaches hold the advantage when it comes to their range of application, as these can be applied to existing case-based interpretability methods as they are.

Moreover, it is relevant not only to consider privacy in the image retrieval process but also to consider interpretability in the privatization process. For instance, the privacy-preserving methods can be used to generate explanations. Montenegro *et al.* [6] apply their privacy-preserving models to the generation of privacy-preserving counterfactual explanations. The authors add a module to their network that generates a counterfactual explanation similar to the privatized factual explanation. The privacy-preserving models could be extended not only to produce counterfactual explanations but also semi-factual explanations, highlighting the changes that



(a) Without Privatization (b) With Privatization (PP-SIR) (c) With Privatization (PP-MIR)

FIGURE 8. Saliency-based validation of the PPCE Framework. Example saliency maps describing the image pixels relevant to identity preservation, glaucoma detection and realism objectives. The top row shows the saliency maps with positive (yellow) and negative (blue) values, while the bottom row shows the absolute value of the maps, where it is overlayed on the input image and also used as an alpha mask.

would approximate the relevant features in a factual explanation to a particular decision boundary.

Another future line of research in Privacy-Preserving Case-based Explanations is the integration of causality [40] into the anonymization process. The preservation of explanatory evidence is one of the greatest challenges during privatization, as it is difficult to pinpoint which image features constitute explanatory evidence. Establishing a causal relationship between the features preserved from the original image and the prediction made by the explanatory detection task could enable a more faithful preservation of explanatory evidence in the images. Additionally, the generation of counterfactual or semi-factual explanations could also gain from ensuring that modified features are causally related to the explanatory detection task.

VII. CONCLUSION

Case-based explanations are intuitive, easy-to-understand, and versatile tools to enable the visual interpretability of Deep Learning models. Nonetheless, by design, the state-of-the-art case-based interpretability methods do not preserve privacy and are therefore not applicable to the case-based explanation of sensitive data.

As an initial step towards enabling case-based explanations in domains with sensitive data, we survey case-based interpretability and privacy-preserving techniques. We empirically evaluate the reviewed privacy-preserving methods considering their application in the domain of case-based explanations. The literature review on privacy-preserving methods shows that most techniques do not guarantee the simultaneous preservation of privacy, intelligibility, and explanatory evidence of the explanations, rendering their anonymized versions useless. To address the lack of research in Privacy-Preserving Case-Based Explanations, we formalize this novel paradigm as a multi-objective problem to preserve privacy, intelligibility, and explanatory evidence in images. We propose interpretability saliency maps as a qualitative measure to evaluate the quality of the anonymization. Our experiments show that a saliency-based analysis offers valuable insights to assess the effects of anonymization. Finally, we propose the development of intrinsically privacy-preserving methods and the integration of causality into the privacy-preserving procedure as future research directions in the novel field of Privacy-Preserving Case-Based Explanations.

To conclude, this work contributes towards the development of more rigorous privacy-preserving methodologies capable of anonymizing case-based explanations without compromising their explanatory value. The development of adequate anonymization techniques is imperative to enable the use of case-based explanations in real-world contexts that deal with sensitive data, like in medicine.

REFERENCES

[1] J. Egger, C. Gsaxner, A. Pepe, and J. Li, "Medical deep learning—A systematic meta-review," 2020, *arXiv:2010.14881*.

[2] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," 2017, *arXiv:1702.08608*.

[3] Z. Camlica, H. R. Tizhoosh, and F. Khalvati, "Autoencoding the retrieval relevance of medical images," in *Proc. Int. Conf. Image Process. Theory, Tools Appl. (IPTA)*, Nov. 2015, pp. 550–555.

[4] W. Silva, A. Poellinger, J. S. Cardoso, and M. Reyes, "Interpretability-guided content-based medical image retrieval," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, 2020, pp. 305–314.

[5] K. Karthik and S. S. Kamath, "A deep neural network model for content-based medical image retrieval with multi-view classification," *Vis. Comput.*, vol. 37, no. 7, pp. 1837–1850, Jul. 2021.

[6] H. Montenegro, W. Silva, and J. S. Cardoso, "Privacy-preserving generative adversarial network for case-based explainability in medical image analysis," *IEEE Access*, vol. 9, pp. 148037–148047, 2021.

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Red Hook, NY, USA: Curran Associates, 2014, pp. 2672–2680. [Online]. Available: http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf

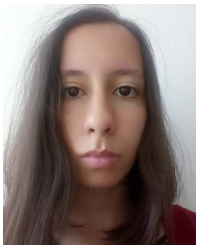
[8] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. 2nd Int. Conf. Learn. Represent. (ICLR)*, 2014, pp. 1–14.

[9] D. J. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 37, 2015, pp. 1530–1538.

[10] E. M. Kenny, C. Ford, M. Quinn, and M. T. Keane, "Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in XAI user studies," *Artif. Intell.*, vol. 294, May 2021, Art. no. 103459.

[11] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery," *Queue*, vol. 16, no. 3, pp. 31–57, Jun. 2018.

[12] D. B. Leake, "CBR in context: The present and future," in *Case-Based Reasoning: Experiences, Lessons, and Future Directions*. Cambridge, MA, USA: MIT Press, 1996, pp. 3–30.
[13] B. Kim, C. Rudin, and J. Shah, "The Bayesian case model: A generative approach for case-based reasoning and prototype classification," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst.*, vol. 2. Cambridge, MA, USA: MIT Press, 2014, pp. 1952–1960.
[14] R. Caruana, H. Kangarloo, J. Dionisio, U. Sinha, and D. Johnson, "Case-based explanation of non-case-based learning methods," in *Proc. AMIA Symp.*, Feb. 1999, pp. 212–215.
[15] C. Nugent, D. Doyle, and P. Cunningham, "Gaining insight through case-based explanation," *J. Intell. Inf. Syst.*, vol. 32, no. 3, pp. 267–295, Jun. 2009.
[16] P. Angelov and E. Soares, "Towards explainable deep neural networks (xDNN)," *Neural Netw.*, vol. 130, pp. 185–194, Oct. 2020.
[17] P. Angelov and E. Soares, "Towards deep machine reasoning: A prototype-based deep neural network with decision tree inference," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2020, pp. 2092–2099.
[18] O. Li, H. Liu, C. Chen, and C. Rudin, "Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions," in *Proc. AAAI*, 2018, pp. 3530–3537.
[19] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su, "This looks like that: Deep learning for interpretable image recognition," in *Proc. Neural Inf. Process. Syst. (NeurIPS)*, vol. 32, 2019, pp. 8930–8941.
[20] N. Papernot and P. D. McDaniel, "Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning," 2018, *arXiv:1803.04765*.
[21] Z. Chen, Y. Bei, and C. Rudin, "Concept whitening for interpretable image recognition," *Nature Mach. Intell.*, vol. 2, no. 12, pp. 772–782, Dec. 2020.
[22] S. Arik and Y. Liu, "Explaining deep neural networks using unsupervised clustering," in *Proc. Workshop Hum. Interpretability Mach. Learn.*, 2020, pp. 377–389.
[23] E. M. Kenny and M. T. Keane, "Twin-systems to explain artificial neural networks using case-based reasoning: Comparative tests of feature-weighting methods in ANN-CBR twins for XAI," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 2708–2715.
[24] E. M. Kenny and M. T. Keane, "On generating plausible counterfactual and semi-factual explanations for deep learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 13, 2021, pp. 11575–11585.
[25] H. Montenegro, W. Silva, and J. S. Cardoso, "Towards privacy-preserving explanations in medical image analysis," presented at the 1st Workshop Interpretable Mach. Learn. Healthcare (IMLH) ICML Conf., 2021. [Online]. Available: https://www.cse.cuhk.edu.hk/~qdou/public/IMLH2021_files/36_CameraReady_Towards_Privacy-preserving_Explanations_in_Medical_Image_Analysis.pdf
[26] M. Trokielewicz, A. Czajka, and P. Maciejewicz, "Assessment of iris recognition reliability for eyes affected by ocular pathologies," in *Proc. IEEE 7th Int. Conf. Biometrics Theory, Appl. Syst. (BTAS)*, Sep. 2015, pp. 1–6.
[27] M. Trokielewicz, A. Czajka, and P. Maciejewicz, "Implications of ocular pathologies for iris recognition reliability," *Image Vis. Comput.*, vol. 58, pp. 158–167, Feb. 2017.
[28] E. M. Newton, L. Sweeney, and B. Malin, "Preserving privacy by de-identifying face images," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 2, pp. 232–243, Feb. 2005.
[29] R. Gross, E. Airoldi, B. Malin, and L. Sweeney, "Integrating utility into face de-identification," in *Privacy Enhancing Technologies*. Berlin, Germany: Springer, 2006, pp. 227–242.
[30] D. Bitouk, N. Kumar, S. Dhillon, P. Belhumeur, and S. K. Nayar, "Face swapping: Automatically replacing faces in photographs," *ACM Trans. Graph.*, vol. 27, no. 3, pp. 1–8, Aug. 2008.
[31] J. Chen, J. Konrad, and P. Ishwar, "VGAN-based image representation learning for privacy-preserving facial expression recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 1570–1579.
[32] D. Aneja, A. Colburn, G. Faigin, L. Shapiro, and B. Mones, "Modeling stylized character expressions via deep learning," in *Proc. Asian Conf. Comput. Vis.* Taipei, Taiwan: Springer, 2016, pp. 136–153.
[33] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of Wasserstein GANs," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, Red Hook, NY, USA: Red Hook, NY, USA: Curran Associates, 2017, pp. 5769–5779. Inc.
[34] D. Cho, J. H. Lee, and I. H. Suh, "CLEANIR: Controllable attribute-preserving natural identity remover," *Appl. Sci.*, vol. 10, no. 3, p. 1120, 2020.
[35] M. Gong, J. Liu, H. Li, Y. Xie, and Z. Tang, "Disentangled representation learning for multiple attributes preserving face deidentification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 1, pp. 244–256, Jan. 2022.
[36] Y. Wu, F. Yang, Y. Xu, and H. Ling, "Privacy-protective-GAN for privacy preserving face de-identification," *J. Comput. Sci. Technol.*, vol. 34, no. 1, pp. 47–60, 2019.
[37] W. Oleszkiewicz, T. Wlodarczyk, K. Piczak, T. Trzcinski, P. Kairouz, and R. Rajagopal, "Siamese generative adversarial privatizer for biometric data," in *Computer Vision—ACCV*. Perth, WA, Australia: Springer, Apr. 2018, pp. 427–497.
[38] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[39] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "SmoothGrad: Removing noise by adding noise," 2017, *arXiv:1706.03825*.
[40] B. Schölkopf, "Causality for machine learning," 2019, *arXiv:1911.10500*.



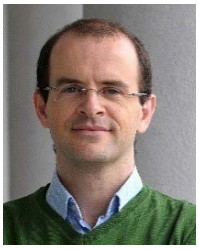
HELENA MONTENEGRO (Student Member, IEEE) was born in Vila Nova de Famalicão, Portugal, in 1998. She received the M.Sc. degree in informatics and computing engineering from the Faculty of Engineering, University of Porto, in 2021, where she is currently pursuing the Ph.D. degree in informatics engineering with the Faculty of Engineering. She is also a Research Assistant at INESC TEC. Her research interests include machine learning and computer vision, with a special focus on privacy-preserving methods for visual data and interpretability.

WILSON SILVA (Student Member, IEEE) received the M.Sc. degree in electrical and computer engineering from the Faculty of Engineering, University of Porto (FEUP), in 2016, where he is currently pursuing the Ph.D. degree in electrical and computer engineering with the Faculty of Engineering. He is also a Research Assistant at INESC TEC, where he is associated with the Visual Computing & Machine Intelligence Group (VCMI) and Breast Research Group. His research interests include machine learning and computer vision, with a particular focus on explainable artificial intelligence and medical image analysis.

ALEX GAUDIO received the master's degree from Carnegie Mellon University and the B.A. degree in music with the Bard College. He is currently pursuing the dual Ph.D. degree with Carnegie Mellon University and the Faculty of Engineering, University of Porto. He is also a Research Assistant at INESC TEC. He co-founded the non-profit NYC Makerspace, offering advanced education and technology to underserved populations in New York City. He was previously a Data Scientist and an Engineer at NYC for seven years. His research interests include explainable machine learning, computer vision, and medical image analysis.

MATT FREDRIKSON (Member, IEEE) received the Ph.D. degree in computer science from the University of Wisconsin-Madison, in 2015. He is currently an Assistant Professor with the School of Computer Science. His research interests include to enable technologies based on fair and trustworthy AI, with a particular focus on verifying privacy and safety properties of machine-learning components. His research has received several awards, including best paper awards at the Usenix Security Symposium and the IEEE Symposium on Foundations of Computer Security. He has attracted more than 5000 citations.

ASIM SMAILAGIC (Fellow, IEEE) has been a Research Professor with the Department of Electrical and Computer Engineering, College of Engineering, Carnegie Mellon University (CMU), where he has been a Faculty Member, since 1992. He is currently the Director of the Laboratory for Interactive and Wearable Computer Systems. He is also the Leader of the Virtual Coaches Research Thrust, NSF Engineering Research Center on Quality of Life Technology, CMU, combining machine learning, sensors, and image analysis. He was a recipient of the Fulbright Postdoctoral Award in Computer Science, CMU, in 1988, the 2000 Allen Newell Award for Research Excellence from CMU School of Computer Science, the 2003 Carnegie Science Center Award for Excellence in Information Technology, the 2003 Steve Fenves Systems Research Award from the CMU College of Engineering, and other prestigious awards. He was three times Program Chairperson of the Quality of Life Technology Symposium, sponsored by the NSF and CMU. He was a Program Chairperson of the IEEE conferences over ten times and the Chair of the IEEE Technical Committee on Wearable Information Systems. He was a Co-Editor, an Associate Editor, and a Guest Editor in leading technical journals, such as the IEEE TRANSACTIONS ON MOBILE COMPUTING, the IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION (VLSI) SYSTEMS, the IEEE TRANSACTIONS ON COMPUTERS, the *Journal on VLSI Signal Processing*, and the *International Journal on Pervasive Computing*.



JAIME S. CARDOSO (Senior Member, IEEE) worked as the President of the Portuguese Association for Pattern Recognition (APRP), affiliated in the IAPR, from 2012 to 2015. He is currently a Full Professor with the Faculty of Engineering, University of Porto (FEUP). His research can be summed up in three major topics, such as computer vision, machine learning, and decision support systems. Image and video processing focuses on medicine and biometrics. The work on machine learning cares mostly with the adaptation of learning to the challenging conditions presented by visual data, with a focus on deep learning and explainable machine learning. The particular emphasis of the work in decision support systems goes to medical applications, always anchored on the automatic analysis of visual data. He has coauthored more than 300 articles and more than 90 of which in international journals, which attracted more than 6500 citations, according to google scholar.