Artificial Intelligence Review manuscript No.
(will be inserted by the editor)

Visual Privacy Attacks and Defenses in Deep Learning: a Survey

Guangsheng Zhang · Bo Liu · Tianqing Zhu* · Andi Zhou · Wanlei Zhou

Received: date / Accepted: date

Abstract The concerns on visual privacy have been increasingly raised along with the dramatic growth in image and video capture and sharing. Meanwhile, with the recent breakthrough in deep learning technologies, visual data can now be easily gathered and processed to infer sensitive information. Therefore, visual privacy in the context of deep learning is now an important and challenging topic. However, there has been no systematic study on this topic to date. In this survey, we discuss algorithms of visual privacy attacks and the corresponding defense mechanisms in deep learning. We analyze the privacy issues in both visual data and visual deep learning systems. We show that deep learning can be used as a powerful privacy attack tool as well as preservation techniques with great potential. We also point out the possible direction and suggestions for future work. By thoroughly investigating the relationship of visual privacy and deep learning, this article sheds insights on incorporating privacy requirements in the deep learning era.

Keywords Visual privacy · Attack and defense · Deep learning · Privacy preservation

1 Introduction

With the widespread use of smartphones and other mobile applications in modern society, people incline to share high-quality images and videos (visual data) on social network platforms, such as Facebook, Instagram, Snapchat, and TikTok. These shared images and videos can easily reveal the owner's personal information, such as location, identity, relationships with other people, and various personal information types. Besides, the adoption of deep learning technologies has allowed attackers to retrieve such private information easily. Privacy preservation of these data has become a critical issue with the booming of deep learning.

Tianqing Zhu is the corresponding author.
Guangsheng Zhang, Bo Liu, Tianqing Zhu, Andi Zhou are all from
Centre for Cyber Security and Privacy, School of Computer Science, University of Technology Sydney, Sydney, Australia.
Wanlei Zhou is with
City University of Macau, Macao.
E-mail: Guangsheng.Zhang@student.uts.edu.au, {Bo.liu, Tianqing.Zhu, Andi.Zhou}@uts.edu.au, wlzhou@cityu.edu.mo

Privacy preservation in images and videos, or as we call it, visual privacy, has become an important topic in recent years for several reasons. First, the online sharing of photos and videos has grown tremendously in popularity, and photography is integral to many mobile applications. As a result, these shared images and videos require privacy preservation (Tonge et al. 2018). Second, there are more surveillance cameras or other image capturing devices in the world than ever before. Privacy-preserving techniques are required for the collection, process, and storage of these images and videos (Jordon et al. 2019). Third, with the booming of deep learning technologies in vision-related tasks, visual data and neural network models are more vulnerable now than transaction data (Orekondy et al. 2019).

We can identify multiple scenarios that require visual privacy preservation. Here we share some examples. The first scenario is online sharing. Image and video sharing is very common in social networks. Sharing photos and videos online is becoming more common, while many attackers are eager to collect these photos and videos published in social networks. As a result, these photos and videos need to be protected in this scenario (Yu et al. 2018). The second scenario is public databases and datasets. With big data and IoT technology developing, many public databases and datasets are collected by governments and enterprises. Visual privacy is essential when these databases are shared or published (Yu et al. 2019). The third scenario is machine learning and deep learning applications. Many images and videos are used in the training and inference phase of deep learning algorithms, which may meet privacy violations (Oh et al. 2018). An illustration of the above three scenarios is presented in Figure 1.

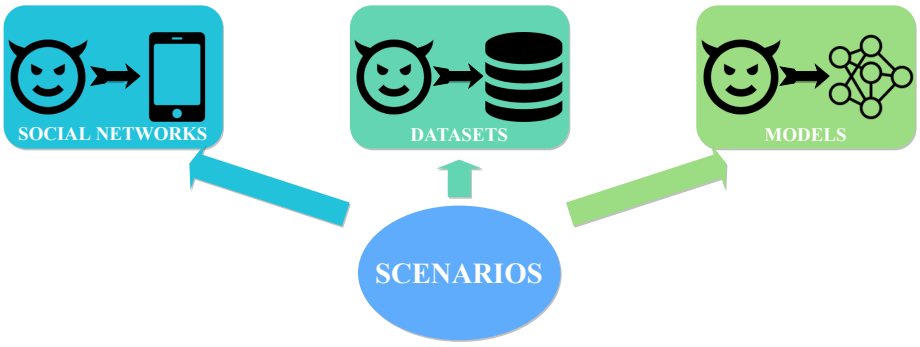

Fig. 1 Visual privacy scenarios.

From the above three scenarios, we can see the importance of privacy preservation in images and videos. Meanwhile, there are some common challenges when preserving visual privacy. The challenges are listed as follows:

- How to define visual privacy. The definition of privacy is not clear across diverse researches. Researchers may have a different research focus, leading to a difference in privacy definition. The definition of visual privacy may need a context analysis.
- How to detect visual privacy. Not all the images and videos contain private content. The detection of private content in images and videos is an issue.
- How to preserve visual privacy. For different types of privacy and in different contexts, many different algorithms can be adopted to protect privacy in images and videos.
- Deep learning is a double-edged sword. Deep learning can be used for both attack methods and defense mechanisms. Whenever an attack method is proposed, there will be a

solution for defending this attack next. Then another corresponding attack method will appear. Deep learning improves the performance of defense mechanisms, but at the same time, more attack methods are proposed.

Visual privacy covers a wide range of topics in the area of computer vision and deep learning. Although it is a very promising topic attracting increasing attention, there has been no survey on visual privacy in the context of deep learning up to date. Existing surveys mainly focused on a part of the research field of visual privacy.

Mireshghallah et al. (2020) discussed privacy in deep learning systems. Liu et al. (2021a) provided a comprehensive review on privacy with deep learning and the outlook in this research field. Some surveys covered only on privacy-preserving algorithms in deep learning systems (Tanuwidjaja et al. 2019; Boulemtafes et al. 2020) or on machine learning as a service (MLaaS) (Tanuwidjaja et al. 2020). Zhang et al. (2021b) discussed both attacks and defenses in cloud-based deep learning. Ha et al. (2020); He et al. (2020a); Liu et al. (2021d) have discussed not only on privacy, but also some methods on security in deep learning systems. However, none of these surveys discuss privacy from a visual perspective. Some other surveys have discussed privacy in the view of social networks (Liu et al. 2021c), database systems (Samaraweera and Chang 2021), or internet of things (IoT) (Amiri-Zarandi et al. 2020). However, these surveys did not include deep learning related technologies. Hu et al. (2021); Enthoven and Al-Ars (2020) discussed privacy attack on membership inference and model inversion, but they did not focus on vision-related tasks. The comparison of surveys on privacy topics is illustrated in Table 1.

There are some other surveys concerning specific technologies related to visual privacy. Differential privacy, one of the leading privacy preservation mechanism, has been widely discussed in Gong et al. (2020); Zhu et al. (2020a). However, differential privacy with the vision-related task is not its primary focus. The survey on homomorphic encryption (Acar et al. 2018) only discussed different encryption schemes without providing the algorithms in a visual perspective. There has been a growing trend of reviewing literature on adversarial examples (Yuan et al. 2019; Serban et al. 2020; Xu et al. 2020; Zhang and Li 2020). However, none of these surveys discuss using adversarial examples as a privacy preservation mechanism. Yinka-Banjo and Ugot (2020); Pavan Kumar and Jayagopal (2021); Wang et al. (2021d) have presented surveys on generative adversarial networks (GAN), but privacy with GAN is not its main topic. Federated learning is another hot research field (Yang et al. 2019a; Li et al. 2020; Lim et al. 2020; Kairouz et al. 2021; Lo et al. 2021; Abdulrahman et al. 2021; Jere et al. 2021; Zhang et al. 2021a). However, these surveys do not discuss privacy issues in federated learning from a visual perspective. The concepts of these technologies are introduced in Section 2 and the methods adopting these technologies are discussed in the following sections. The comparison of surveys on these technologies is illustrated in Table 2.

After reviewing the literature from high-quality journals and conferences, we categorize some deep learning algorithms and technologies into attack methods and corresponding defense mechanisms concerning visual privacy. The contributions of this survey are as follows:

- Instead of focusing on parts of privacy issues or a single kind of technology related to visual privacy, we provide a comprehensive study covering all aspects of visual privacy in deep learning in this survey.
- The definition of visual privacy and related concepts are discussed in this survey. The visual privacy issues are categorized into visual data privacy and visual deep learning system privacy.
- We summarize the attack methods for visual privacy in deep learning. We first highlight some hot-topic attacks methods (membership inference attacks, model inversion attacks,

Table 1 Comparison of surveys on privacy topics

Literature	Visual Privacy Coverage	Topic
Tanuwidjaja et al. (2019)	Yes	Privacy preservation in deep learning
Amiri-Zarandi et al. (2020)	No	Privacy in internet of things
Boulemtafes et al. (2020)	Yes	Privacy preservation in deep learning
Enthoven and Al-Ars (2020)	Yes	Privacy in model inversion attacks
Ha et al. (2020)	Yes	Privacy and security in deep learning
He et al. (2020a)	Yes	Privacy and security in deep learning
Mireshghallah et al. (2020)	Yes	Privacy in deep learning
Tanuwidjaja et al. (2020)	Yes	Privacy preservation on MLaaS
Hu et al. (2021)	Yes	Privacy in membership inference attacks
Liu et al. (2021c)	Yes	Privacy in social networks
Liu et al. (2021d)	Yes	Privacy and security in deep learning
Liu et al. (2021a)	Yes	Privacy in deep learning and its outlook
Samaraweera and Chang (2021)	No	Privacy in database systems
Zhang et al. (2021b)	Yes	Privacy attacks and defenses in cloud-based deep learning

Table 2 Comparison of surveys on privacy technologies

Literature	Privacy Coverage	Technology
Acar et al. (2018)	Yes	Homomorphic encryption
Yang et al. (2019a)	Yes	Federated learning
Yuan et al. (2019)	No	Adversarial examples
Gong et al. (2020)	Yes	Differential privacy
Li et al. (2020)	Yes	Federated learning
Lim et al. (2020)	Yes	Federated learning
Serban et al. (2020)	No	Adversarial examples
Xu et al. (2020)	No	Adversarial examples
Yinka-Banjo and Ugot (2020)	No	Generative adversarial networks
Zhang and Li (2020)	No	Adversarial examples
Zhu et al. (2020a)	Yes	Differential privacy
Abdulrahman et al. (2021)	Yes	Federated learning
Jere et al. (2021)	Yes	Federated learning
Kairouz et al. (2021)	Yes	Federated learning
Lo et al. (2021)	Yes	Federated learning
Pavan Kumar and Jayagopal (2021)	No	Generative adversarial networks
Wang et al. (2021d)	No	Generative adversarial networks
Zhang et al. (2021a)	Yes	Federated learning

and model extraction attacks) and then briefly introduce several other attack methods, along with in-depth comparisons and discussions.

– We review the defense mechanisms corresponding to the attack methods for visual privacy in deep learning. State-of-the-art literature regarding defense mechanisms is introduced in this part and compared, including generative adversarial networks, adversarial examples, differential privacy, homomorphic encryption, secure multi-party computation, and some other mechanisms.
– We provide some insights for future research directions on visual privacy.

All of the existing surveys only focused on parts of visual privacy. In contrast, our survey gives a comprehensive review of visual privacy and the corresponding attack and defense methods in deep learning.

The remaining part of the paper proceeds as follows. Preliminaries are given in Section 2, where notations, privacy definition, the concept of visual privacy and deep learning, and the commonly-used datasets are introduced. We present attack methods and correspond-

ing defense mechanisms in Section 3 and Section 4 respectively. At last, future directions and the conclusion are addressed in Section 5 and Section 6.

2 Preliminaries

This section provides a definition of privacy and explains the fundamentals of deep learning and the relationship between visual privacy and deep learning, visual datasets mentioned in this survey, and the taxonomy of the methods covered in this survey. Notations and explanations are given in Table 3 for the convenience of readers.

Table 3 Notations.

Notation	Explanation
D	Dataset
n	Size of dataset
X	A set of input data (e.g. training set)
x	An instance of input data
y	The ground truth for the input instance x
F	The deep learning model
θ	The parameters of the deep learning model

2.1 Privacy Definition

There are very few articles that give an exact definition of privacy. Usually, privacy means something special or sensitive to an individual person. In the EU's General Data Privacy Regulation (GDPR) (European Parliament 2016), the definition of personal data is given below:

> 'Personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.

GDPR is also intended to regulate how companies collect or process personal data. For example, the regulation addresses how companies should collect or store users' data, how data consent should be gained from users, and how personal data should be processed or applied in products.

Images and videos are two visual data formats that usually contain personal data, including faces, personal belongings, and location data. In this survey, aligning with the GDPR, visual privacy covers the data that can identify a person in images and videos. The privacy can be the objects in images/videos or even the whole images/videos. As deep learning models are trained by image/video datasets, these visual datasets and models can also be considered as privacy.

2.2 Deep Learning

Deep learning is a sub-type of machine learning method derived from deep neural networks. To discover the intricate structures in high-dimensional data, many research fields, such as computer vision (He et al. 2016), natural language processing (Collobert et al. 2011), voice recognition (Mikolov et al. 2011), and medical research (Leung et al. 2014), have adopted deep learning technologies to gain state-of-the-art results (LeCun et al. 2015).

A deep learning model consists of neural network layers as basic blocks. With the input data processed through, the parameters of the model are trained by optimizing the output. The goal is to find the relation between the input data and the prediction. The relation is considered as a function: $F : X \rightarrow Y$. There are two phases to discover the relationship, the training phase and the inference phase.

Training phase. The goal of the training phase is to find the relation F by minimizing the objective function L, usually called the loss function or the cost function. The loss function L can be described as:

$$L(F(X;\theta),Y) = \frac{1}{N}\sum_{i} L(F(x_i;\theta),y_i), \tag{1}$$

where x is an instance of dataset X; y represents the model's prediction.

In order to optimize the lost function, the stochastic gradient descent (SGD) (Bottou 1998) and back propagation () are applied to compute the gradients and update the parameters θ with a learning rate.

Inference phase. The inference phase is also called the testing phase or the deploying phase, where the model F is used to predict unseen data. The deep learning model takes input data from the same distribution and outputs prediction.

There are two types of deep learning models based on the requirement of dataset labels. The *supervised learning* model is trained with a labeled dataset. In some applications, it is impossible for researchers to obtain enough samples and labels. The *unsupervised / self-supervised learning* model is trained by the training dataset without human-annotated labels. The labels are obtained via a semi-automatic process.

Convolutional neural networks (CNNs) are widely used in deep learning, which can extract representations from input data (LeCun et al. 2010). A typical CNN structure generally consists of multiple convolution layers, activation layers, pooling layers, and some fully connected layers at the end. Batch normalization (Ioffe and Szegedy 2015) and dropout (Hinton et al. 2012) are also incorporated for performance optimization. With good feature extraction and discrimination ability, CNNs are often applied in various computer vision applications.

2.3 Visual Privacy and Deep Learning

In this survey, deep learning is considered as a tool: it can be used either for adversaries to attack visual privacy or for defenders to resist privacy violations. In the following subsections, we introduce key algorithms and technologies for privacy preservation, system settings for deep learning, and visual privacy attack and defense settings.

2.3.1 Key Algorithms and Technologies for Privacy Preservation

Generative Adversarial Networks (GANs). A standard generative adversarial network consists of two components: a generator F_G and a discriminator F_D (Goodfellow et al. 2014).

The generator fools the discriminator, and the discriminator distinguishes between real images and generated images. Given a random image z and its distribution $z \sim p_z$, F_G calculates the distribution p_G as $F_G(z) \sim p_G$. Given a real image x and its distribution $x \sim p_r$, the aim of a GAN is to learn p_G that matches p_r. The loss function for GAN is defined as:

$$\min_G \max_D \mathbb{E}_{x \sim p_r} \log[F_D(x)] + \mathbb{E}_{z \sim p_z} \log[1 - F_D(F_G(z))]. \tag{2}$$

Adversarial Examples. Adversarial examples (Szegedy et al. 2014) are attracting growing interest in the deep learning community. Given a trained model F and an input data x, the task generates an adversarial example x' which satisfies this:

$$\begin{aligned} &\min_{x'} ||x' - x|| \\ &s.t.\ F(x') = l', F(x) = l, l \neq l', x' \in [0,1], \end{aligned} \tag{3}$$

where l and l' mean the output label of x and x' respectively, and $||\cdot||$ denotes the distance between two data sample.

Differential Privacy. The algorithms of differential privacy involve adding noises to the data or the deep learning model until the attacker cannot distinguish an individual data sample from the dataset or cannot recover certain private information from the model. Differential privacy is a kind of mechanism to guarantee the privacy preservation of a specific dataset. For any dataset D, there exists a neighboring dataset D' with a randomized mechanism M, which gives (ϵ, δ)-differential privacy (Dwork et al. 2006) for every set of outputs Ω, if M satisfies:

$$Pr[M(D) \in \Omega] \leq e^{\epsilon} \cdot Pr[M(D') \in \Omega] + \delta \tag{4}$$

where ϵ is the privacy budget parameter which decides the privacy level, and δ loosens the bound of the error.

Homomorphic Encryption. Homomorphic encryption is an encryption method for those who would like to make a computation of the encryption of input data without decrypting it and without having the private key Rivest and Dertouzos (1978); Gentry (2009). The concept could date back to Rivest and Dertouzos (1978) and the first fully homomorphic encryption (FHE) was proposed by Gentry (2009), which described a mechanism of any computation on encrypted data without having the private key. The definition of the encryption function Enc can be described as:

$$Enc(a) * Enc(b) = Enc(a * b) \tag{5}$$

where $Enc : x \rightarrow y$ is a homomorphic encryption scheme that maps data x to encrypted data y. The operation $*$ represents the encryption process.

Secure Multi-Party Computation. Secure multi-party computation (SMC) offers a way of enabling multiple parties (who do not trust each other) to compute together without revealing their input data. Each party knows nothing else but its own input and output data. The concept of secure computation was first proposed by Yao (1986), and secure multi-party computation was proposed by Goldreich et al. (1987). This mechanism gives a solution of protecting each party's data privacy when computing with other non-trusted parties.

2.3.2 System Settings in Visual Deep Learning Systems

The system settings in visual deep learning system can be divided into three different subgroups:

Machine learning as a service (MLaaS). Machine learning as a service (MLaaS) is a cloud-based system generally built by a large company. This paradigm is a cloud infrastructure for customers to upload datasets and machine learning models. Then the server offers prediction results to the customers. The privacy risks in this system setting are: 1) Data and model queries are easily acquired by service providers; 2) Outside attackers can also obtain model queries by service APIs.

Centralized systems. A centralized system is a system where all the components are located on the same computer. This is a default system setting for deep learning models.

Distributed systems. A distributed system is a system where all the components are located on different computers but connected via networks. Usually, there are multiple clients and one server. The working clients store the data, and the system server allocates the data and trains the model more efficiently based on a specific schedule.

A federated learning system, one of the popular distributed system settings, is built to protect the data owner's data privacy. Federated learning is a framework where a centralized model is trained using decentralized data. Federated learning was first proposed by Google (Konečný et al. 2016, 2017; McMahan et al. 2017). Here we describe the definition of federated learning. Suppose there are n data owners $\{O_1, ..., O_n\}$ whose goal is to train a deep learning model with their datasets $\{D_1, ..., D_n\}$. In a centralized system, the total dataset D_{cen} is formed by all the datasets put together, where $D_{cen} = D_1 \cup ... \cup D_n$. The model F_{cen} is trained by D_{cen}. In a federated learning system, the model F_{fed} is trained in a way where all the data owners do not expose their own datasets. Besides, the performance of F_{fed} (e.g. accuracy in image classification), denoted as $Perf_{fed}$, should be very close to the performance of F_{cen}, $Perf_{cen}$. The federated learning algorithm has δ-accuracy loss, if

$$|Perf_{fed} - Perf_{cen}| < \delta \tag{6}$$

where δ is a non-negative real number (Yang et al. 2019a).

2.3.3 Visual Privacy Attack and Defense Settings

Type of Attackers and Defense Mechanisms The attackers in visual privacy can be either machines or human.

- *Machines.* Well-designed deep learning algorithms can be considered as attackers aiming to breach privacy.
- *Human.* Human attackers can retrieve private information by inspecting leaked private data or making queries of private data.

To defend against these attackers, deep learning based defense mechanisms can be used to preserve privacy.

Privacy Preservation Targets The targets for privacy preservation can be divided into two different groups: visual data (image and video data samples) and visual deep learning systems (deep learning systems focused on visual data).

- *Visual data privacy.* Privacy in visual data can be easily breached in social networks if not protected. This category can be divided into two different subgroups:
  Object-level privacy. Some object-level contents in images and videos are considered as privacy preservation targets. For example, in an online photo-sharing application, people's faces are private contents. In some other applications, the person's belongings are private contents.
  File-level privacy. In some applications, considering the content within images and videos is insufficient. The whole image or video should be considered private. In this way, the private data in data is unreadable to human attackers.
- *Visual deep learning system privacy.* That is, deep learning systems focused on visual data. In order to attack the targeted system to breach privacy, some attackers aim to attack deep learning models or a whole deep learning system. In these circumstances, deep learning systems should be considered as private data to be protected, including the deep learning model and the whole training/inference phase.
  Dataset-level privacy. The dataset is essential in a deep learning task. People who develop deep learning models can see the data directly in the dataset. If the data in the dataset is lost or stolen, this is an obvious privacy leak. The attacker can also extract private information by inferring the deep learning models. In these scenarios, dataset-level privacy should be protected while maintaining the dataset's usability.
  Model-level privacy. Deep learning models are files saved after the training phase that contain model parameters and model structures. They can be used in the inference phase, and in many scenarios, they will be published on the internet. Without proper preservation, the attackers can find ways to extract sensitive information from these deep learning models.

Attacker Knowledge in Visual Deep Learning Systems Based on the accessability of deep learning models, the attacker knowledge can be divided into two subgroups:

- *Black-box settings.* In a black-box setting, an attacker has no access to the model parameters and can only make a series of prediction queries to the model and receive corresponding output results.
- *White-box settings.* In a white-box setting, an attacker is allowed to download the model. The attacker has total knowledge of the model, including model structures, model parameters, and the training dataset.

Different settings on the attacker knowledge lead to different attack methods, which are introduced in Section 3.

2.4 Visual Datasets

Table 4 is a comparison of all the datasets evaluated in the literature which is referenced in this survey. We provide the following features of the datasets: source, dataset size, resolution, contents, and task. The source denotes whether it is an image or video dataset. The dataset size means the size of the dataset samples, categorized as small (less than 10,000), medium (10,100 100,000), or large (more than 100,000). If it is a video dataset, the dataset size is

categorized as small (less than 10 video sequences) or large (more than 10 video sequences). The resolution indicates the image or video frame resolution, which is divided into small or large (smaller or larger than 512×512 pixels). The contents and task describe the general or specific contents contained in the dataset and which task the dataset is for.

Table 4 Comparison of visual datasets

Name	Source	Dataset Size	Resolution	Task
3DPW (von Marcard et al. 2018)	Video	Large	Small	Pose reconstruction
10MonkeyS (Kaggle 2021a)	Image	Small	Small	Image classification
AT&T Faces (Samaria and Harter 1994)	Image	Small	Small	Face recogniztion
BDD100K (Yu et al. 2020)	Image	Small	Large	Semantic segmentation
Caltech-256 (Griffin et al. 2007)	Image	Medium	Small	Image classification
CelebA (Liu et al. 2015)	Image	Large	Small	Face recogniztion
CelebA-HQ (Karras et al. 2018)	Image	Medium	Large	Face recogniztion
ChestX-ray (Wang et al. 2017)	Image	Large	Large	Image classification
CIFAR (Krizhevsky 2009)	Image	Medium	Small	Image classification
Cityscapes (Cordts et al. 2016)	Image	Small	Large	Semantic segmentation
COCO (Chen et al. 2015)	Image	Large	Large	Object detection
CUBS-200-2011 (Wah et al. 2011)	Image	Medium	Small	Image classification
DALY (Weinzaepfel et al. 2017)	Video	Large	Large	Action detection
DR (Kaggle 2021b)	Image	Medium	Small	Image classification
FaceScrub (Ng and Winkler 2014)	Image	Medium	Small	Face recogniztion
FMNIST (Xiao et al. 2017)	Image	Medium	Small	Image classification
FEMNIST (Caldas et al. 2019)	Image	Large	Small	Image classification
FERG (Aneja et al. 2017)	Image	Medium	Small	Expression recognition
FFHQ (Karras et al. 2019)	Image	Medium	Large	Face recogniztion
Flickr Logos (Kalantidis et al. 2011)	Image	Small	Small	Image classification
Flowers-17 (Nilsback and Zisserman 2006)	Image	Small	Small	Image classification
Gisette (Guyon et al. 2004)	Image	Medium	Small	Image classification
GTSRB (Stallkamp et al. 2012)	Image	Medium	Small	Traffic sign recognition
HMDB51 (Kuehne et al. 2011)	Video	Large	Large	Action recognition
ImageNet (Russakovsky et al. 2015)	Image	Large	Large	Image classification
Indoor Scenes (Quattoni and Torralba 2009)	Image	Medium	Small	Image classification
INRIA (Dalal and Triggs 2005)	Image	Small	Small	Pedestrian detection
JHMDB (Jhuang et al. 2013)	Video	Large	Large	Action detection
LFW (Huang et al. 2008)	Image	Medium	Small	Face recogniztion
Mapillary (Neuhold et al. 2017)	Image	Medium	Large	Semantic segmentation
Market1501 (Zheng et al. 2015)	Image	Medium	Large	Person Re-id
MNIST (LeCun and Cortes 2005)	Image	Medium	Small	Image classification
MUG (Aifanti et al. 2010)	Video	Large	Small	Expression recognition
MS-Celeb-1M (Guo et al. 2016)	Image	Large	Small	Face recogniztion
Pascal VOC (Everingham et al. 2012)	Image	Large	Large	Object detection
PETS2010 (Ferryman and Ellis 2010)	Video	Large	Large	Pedestrian tracking
PicAlert (Zerr et al. 2012)	Image	Medium	Small	Privacy research
PIPA (Zhang et al. 2015)	Image	Medium	Large	Face recogniztion
PubFig (Kumar et al. 2009)	Image	Medium	Small	Face recogniztion
RAVDESS (Livingstone and Russo 2018)	Video	Large	Large	Emotion recognition
SBU kinect (Yun et al. 2012)	Video	Large	Small	Action recognition
STL-10 (Coates et al. 2011)	Image	Medium	Small	Image classification
SVHN (Netzer et al. 2011)	Image	Medium	Small	Image classification
SynthText (Gupta et al. 2016)	Image	Large	Small	Text detection
UCF101 (Soomro et al. 2012)	Video	Large	Small	Action recognition
UTKFace (Zhang et al. 2017)	Image	Medium	Small	Face recogniztion
Venice (Leal-Taixé et al. 2015)	Video	Large	Large	Pedestrian tracking
VGGFace2 (Cao et al. 2018)	Image	Large	Small	Face recogniztion
VISPR (Orekondy et al. 2017)	Image	Medium	Small	Privacy research
WebFace (Yi et al. 2014)	Image	Large	Small	Face recogniztion
Yale Face B (Lee et al. 2005)	Image	Small	Small	Face recogniztion
YourAlert (Spyromitros-Xioufis et al. 2016)	Image	Small	Small	Privacy research

2.5 Taxonomy

A taxonomy of different tasks and algorithms in visual privacy is detailed in Figure 2. The algorithms in this survey can be categorized into attack methods and corresponding defense mechanisms. The algorithms are then divided into visual data privacy and visual deep learning system privacy. Every algorithm in the taxonomy is discussed in the following two sections.

Fig. 2 A taxonomy of visual privacy and deep learning.

3 Attack Methods

We roughly divide the attack methods into two groups: 1) methods with private visual data as the target, and 2) methods with private visual deep learning systems as the target.

Visual data can be attacked by deep learning-assisted adversaries and defended through corresponding methods. As adversaries can easily attack unprotected visual data by general deep learning technologies, few research papers discuss these algorithms from the perspective of adversaries.

Visual deep learning system privacy focuses on the privacy aspect of deep learning systems using visual data as training datasets. In a deep learning task, both datasets and models are essential. Many deep learning models involve processing large amounts of datasets in which the data may contain sensitive or private information. Deep learning models can also contain private information. Attackers adopt various methods to extract private information from visual datasets, while defenders develop corresponding methods to defend against these attacks.

In the following subsections, we first discuss the privacy attack methods on visual data, followed by privacy attack methods on visual deep learning systems.

3.1 Privacy Attack on Visual Data

Sharing images and videos in online social networks has become a part of everybody's daily activity. With cameras on cellphones capturing various scenes of people's daily life, posting on social networks can be completed anywhere and anytime. This causes severe privacy leakage, as all the sensitive information on an individual is exposed on the internet that the attacker can possess.

With the advent of the deep learning era, sharing data between organizations has become frequent nowadays. Many public institutions also share datasets for research and information publishing (i.e. the American data portal (AmericanDataPortal 2021) with 209,765 datasets). The drawback in this scenario is the privacy risk of individual data contained in these datasets. The attacker can obtain these unprotected data for malicious usage.

The above two scenarios reveal the threat of privacy on visual data. Visual data can be attacked by humans or some deep learning methods. As visual data in these scenarios is public to everyone, human attackers can directly browse and search private information in these images and videos. When the attackers have deep learning model-based attack tools, they can detect, extract, or retrieve sensitive information in visual data. The attacker can use recognition models to find people in the photos, or detection models to detect texts and extract sensitive information in the photos, or recognition models to recognize some well-known buildings to locate where a person is. This is how a privacy attack is carried out with deep learning.

Visual data privacy can be divided into two groups: 1) object-level privacy, referring to the specific visual content in an image; 2) file-level privacy, which means that we consider the entire image as private information. In order to protect privacy in visual data, some simple defense mechanisms consist of blurring, cropping, or pixelation. Oh et al. (2016) proved that these simple obfuscation mechanisms were not practical to current recognition systems. When the faces in images were obfuscated, the attacker could still recognize the person's identity using body features.

There is more literature on how to defend visual data privacy, which is presented in Section 4.1. Next, we introduce various privacy attack methods on visual deep learning systems.

3.2 Privacy Attack on Visual Deep Learning Systems

Deep learning algorithms have achieved remarkable success in various computer vision applications. Researchers are eager to publish their work along with source codes and trained deep learning models. These deep learning models are vulnerable to various attacks.

Some companies and organizations have developed machine learning as a service (denoted as MLaaS) to provide deep learning models queries to the customers. The attack can still be launched by making model queries even if the models are hidden behind the service. In a federated learning setting, the clients only need to send parameters to the server to protect local sensitive data. However, some attackers can reconstruct the local sensitive data by making analyses on the server.

The above examples show how visual deep learning system privacy can be attacked. Visual deep learning system privacy can be divided into two groups: 1) dataset-level privacy, meaning the private information in the training dataset; 2) model-level privacy, meaning the parameters, hyperparameters, or the structure of the model. Both the dataset and the model can be attacked to extract sensitive information. In the following subsections, we introduce the attack methods for extracting dataset-level and model-level privacy.

3.2.1 Membership Inference Attacks

With sensitive information in the training dataset, the deep learning models are vulnerable to various attacks. Membership inference attacks aim to find whether a specific data sample has participated in the training phase of the model. Although the attacker does not have direct access to the training dataset or the trained model, they can still make queries and use the trained model for prediction, which means the model's prediction is applied to check whether a specific data sample is in the training dataset.

Shokri et al. (2017) proposed the first paper on membership inference attacks. The targeted deep learning model is called the victim model. The attacker had black-box access to the victim model, so they could only obtain confidence scores after queries. In their method, generating multiple shadow models were required to imitate the behavior of the victim model. Using these shadow models, an attack model was trained to classify between member samples and non-member samples. A high confidence score would be received in the victim model's prediction if a data sample was in the victim model's training dataset. The attack model could get a strong membership signal based on the prediction.

Research Focus A large group of the research focused on membership inference attacks under three aspects: the use of shadow models, different prediction outputs, and applications with different domains.

Shadow models are essential in membership inference attacks, as they are used to imitate the behavior of the victim model. Shokri et al. (2017) applied multiple shadow models and multiple corresponding shadow datasets. Salem et al. (2019) proved that only one shadow model was enough for the attack.

Not all membership inference frameworks leverage confidence scores of the victim model's prediction to launch the attacks. Some methods can launch the attacks leveraging other prediction output information. By making only one query to the victim model, Yeom et al. (2018) leveraged the victim model's training loss to launch the membership inference attack. Sablayrolles et al. (2019) showed that the optimal membership inference depended on the loss function, not the prediction scores. Choquette-Choo et al. (2021) and Li and

Zhang (2021) both proposed attacks with only access to the victim model's prediction labels. Instead of using confident scores of the victim model's output, they applied various data augmentations to the original images and obtained corresponding prediction labels as membership signals.

Early papers on membership inference attacks were demonstrated in classification models. Some papers proved the existence of membership inference attacks in other domains. Hayes et al. (2019) and Chen et al. (2020b) presented membership inference attacks against GANs. Similar to classification models, the generators in a GAN setting also tend to implicate privacy risks from the victim model's training dataset. There are more papers extending the scenarios to attacks against other deep learning models, including semantic segmentation He et al. (2020b) and object detection Park and Kang (2020). Contrastive learning is a kind of self-supervised learning which aims to learn representations by maximizing agreement between differently augmented views of the same data sample (Chen et al. 2020c). These informative representations learned by contrastive models may leak privacy risks (He and Zhang 2021). Melis et al. (2019); Nasr et al. (2019); Truex et al. (2019) studied membership inference attacks under a federated system setting. Salem et al. (2020) proposed the attacks for the updated dataset in online learning. Song and Raghunathan (2020); Zou et al. (2020) presented the attacks against a transfer learning paradigm. Song et al. (2019b) demonstrated that models for defense against adversarial examples might be vulnerable for membership inference attacks.

Discussion There are three main reasons for the existence of membership inference attacks: the overfitting feature of the deep learning models, the data distribution between member and non-member samples, and the influence of data augmentations.

A deep learning model usually obtains better accuracy results on the training dataset than on the testing dataset. This overfitting feature is why membership inference attacks work. Yeom et al. (2018) demonstrated that overfitting was sufficient for an attacker to perform an attack. Salem et al. (2019) discovered that a more overfitted deep learning model would be more vulnerable to membership inference attacks. Chen et al. (2020b) demonstrated that overfitting also extended the attacks on generative adversarial networks. Leino and Fredrikson (2020) studied that when a victim model had the overfitting issue, the membership signal was leaked by the victim model's personal use of features. He and Zhang (2021) argued that contrastive models are less vulnerable to membership inference attacks than supervised models because contrastive models are less overfitted.

The second aspect is the data distribution. Song et al. (2019a) showed that there would be a significant divergence between the loss distributions over member and non-member samples for the attack model. Leveraging this divergence, the attacker could successfully train an attack model and launch membership inference attacks.

The third aspect is data augmentations. Applying data augmentations to the victim model's training dataset leads to a dataset with enhanced membership signals. Choquette-Choo et al. (2021) demonstrated that the label-only attacks can be launched with data augmentations. Kaya and Dumitras (2021); Yu et al. (2021a) proved that data augmentations could bring a more generalized attack model.

The first membership inference attack (Shokri et al. 2017) is a black-box attack, followed by several papers. The intuitive idea is that white-box attacks are more straightforward and can extract more information. However, Sablayrolles et al. (2019) demonstrated that white-box settings for the attacks did not contribute more membership signals than black-box settings.

There have been a few papers discussing the theory behind membership inference attacks. Saeidian et al. (2021) provided a quantitative analysis of membership inference attacks via information leakage. The information leaking on the dataset's data sample was measured using a conditional form of maximal leakage (Issa et al. 2020).

3.2.2 Model Inversion Attacks

Another hot topic in the research field is the model inversion attacks, also called attribute inference attacks in some literature. These attacks target obtaining information on training data from queries. These attacks could reconstruct possible data samples in the training data.

The work of model inversion attacks starts by Fredrikson et al. (2014, 2015). Fredrikson et al. (2015) proposed to infer training classes in a white-box setting as an optimization-based problem. In their paper, model inversion attacks were used in a face recognition task to reconstruct recognizable faces. Their method managed to find the optimal data for a given class.

Research Focus There are two different scenarios in model inversion attacks: data reconstruction (the attacker can reconstruct some data sample in the training dataset) and training class reconstruction (the attacker can reconstruct some data that belong to a training class). In a distributed system, sharing model gradients could leak private information from the model's training dataset. The attacker can pretend to be one of the clients or can intercept information from the clients to the server in the federated learning system. The federated server can update gradients using distributed SGD (Lin et al. 2018) or update parameters using federated averaging (McMahan et al. 2017).

Yang et al. (2019b) proposed a model inversion attack in a training-based approach. The inversion model could be trained in a black-box setting. The method in this paper did not require access to the original training data using background knowledge. Zhu et al. (2019) discovered the possibility to obtain private data from publicly shared gradients. However, their method had the problem of model convergence. Zhao et al. (2020a) proposed an improved method called iDLG and can extract the victim model's ground truth labels. These papers aimed to reconstruct training classes.

Geiping et al. (2020) demonstrated an algorithm to reconstruct the victim model's data samples in federated learning. Wei et al. (2020) provided an evaluation framework for client data leakage in federated learning. These two papers presented data reconstruction methods.

Model inversion attacks can also be applied in applications other than image classification, such as generative adversarial networks (GANs). Hitaj et al. (2017) demonstrated a GAN-based attack in collaborative learning. Zhang et al. (2020b) proposed a generative model inversion attack, where a distributional prior by GANs was learned to guide the inversion process.

Discussion Deep learning models are more vulnerable when they are overfitted, causing the success of model inversion attacks, which is similar to membership inference attacks (Yeom et al. 2018). Moreover, the overfitting feature of deep learning models is not the only reason that leads to privacy risk: the victim model's output can also leak training data information. This factor then inspires a series of model inversion attacks to reconstruct training data by model gradients in federated learning.

The strategies of launching model inversion attacks can be summarized as optimization. Geiping et al. (2020) reconstructed training data using model gradients. They trained the

attack model by minimizing the reconstruction error between the victim model gradients and the attack model gradients to regenerate data samples that were likely to be from the victim model's training data. In this way, the attack model could inverse the victim model to reconstruct data samples.

3.2.3 Model Extraction Attacks

Developing a deep learning model requires data acquisition and annotation, model training, and deploying. Service providers tend to provide deep learning models in machine learning as a service (MLaaS) with all these costs. However, the popularity of MLaaS makes deep learning models a potential privacy risk due to the unlimited model queries. Attackers develop multiple malicious algorithms to steal well-designed deep learning models. Model extraction attacks, also called model stealing attacks or model reverse-engineering, aim to extract private information from deep learning models in these scenarios. Deep learning models can be accessed via MLaaS APIs by these attacks. Model extraction attacks were first proposed by Tramèr et al. (2016), who showed how to extract model parameters from MLaaS prediction results. Their simple equation-solving attacks used non-adaptive, random queries to solve the model parameters. They evaluated the attacks on logistic regressions, decision trees, and simple neural networks but did not scale to deep learning models. More papers on attacks for deep learning models are reviewed in the following paragraphs.

Research Focus There are several dimensions of model extraction attacks here, including extracting model parameters, extracting hyperparameters, and stealing model functionality.

Extracting model parameters is the most common one in model extraction attacks. Several papers discussed this topic after the work of Tramèr et al. (2016). In Milli et al. (2019), the algorithm learned a two-layer ReLU network to reconstruct models from querying the gradients. Jagielski et al. (2020) developed a learning-based attack by exploring the objective of accuracy and fidelity. Pal et al. (2020) presented ActiveThief, a framework to extract neural network models with fewer data samples.

Extracting hyperparameters is another research focus in this research field. Oh et al. (2018) proposed a method to expose the model's internal information from querying and a method to predict attributes related to the model's hyperparameters and architecture. Wang and Gong (2018) gave a systematic study of hyperparameter stealing attacks in some machine learning algorithms.

Stealing functionality or model copying forms the last research focus. Correia-Silva et al. (2018, 2021) proposed CopycatCNN, an approach to create a copycat of CNN model using random unlabeled images. Orekondy et al. (2019) studied model functionality stealing, a method for the attacker to transfer the model's functionality via black-box access. Barbalau et al. (2020) presented a teacher-student framework to distill the target model (teacher model). They generated data samples for the student model by an evolutionary strategy.

Discussion Parameters, hyperparameters, and structures of the deep learning model are critical because they contribute to the model's performance. Model extraction attacks aim to either steal all the parameters from the victim model to form a copied model for malicious usage or break the existing victim model to lower its performance. The main objectives of model extraction attacks are breaking model accuracy and fidelity (Jagielski et al. 2020). Accuracy means the model prediction performance in the inference phase. Fidelity measures the difference between the victim model and the copied model. A model extraction attack

usually achieves high fidelity while it results in low prediction accuracy or copies the victim model with high prediction accuracy but suffers low fidelity. The balance between accuracy and fidelity is crucial for model extraction attacks.

3.2.4 Other Attack Methods

In addition to the several mainstream attack methods mentioned in the above subsections, there are some other attack settings in the research field.

Property Inference Attacks Property inference attacks also reconstruct sensitive information from the training data. Unlike membership inference attacks and model inversion attacks, which focus on specific data samples in the training dataset, property inference attacks aim to infer some fundamental statistical properties in the training dataset. Some papers refer to these attacks as attribute inference attacks or feature inference attacks (He and Zhang 2021).

Property inference attacks have been first introduced by Ateniese et al. (2015). The attack model was a classifier to determine whether the target model had a specific property. Similar to membership inference attacks, shadow model techniques were also applied in this attack. However, their work only discussed property inference attacks on traditional machine learning models.

Ganju et al. (2018) have discussed the effectiveness of property inference attacks on more complex machine learning models, such as fully connected neural networks. Melis et al. (2019) have extended property inference attacks to the domain of collaborative learning. He and Zhang (2021) demonstrated that contrastive models were more vulnerable to property inference attacks than supervised models because representations generated by contrastive models contained richer and more information for the attacks to exploit.

Model Memorization Attacks Song et al. (2017) proposed model memorization attacks in a MLaaS setting. A malicious service provider supplies model training codes to a data holder without access to the training phase. However, they can access the training model, in which sensitive training data can be extracted for other malicious use, resulting in privacy leakage.

Privacy Violation in Data Aggregation In a deep learning life cycle, data aggregation means data collection and annotation by researchers, data scientists, and data engineers, which is crucial to develop a powerful deep learning model. These researchers, data scientists, and data engineers can be called modelers (people who work with deep learning models). During model aggregation and model training, direct data inspection is inevitable by modelers, which should be considered privacy leakage. The modeler may not intentionally breach dataset privacy, but it is still similar to a more typical attack in this case. Thus, we still need to find a way to prevent visual privacy violations in data aggregation.

3.3 Discussion on Attack Methods

Since there is no specific attack method for visual data privacy, we mainly introduced different attack methods for visual deep learning system privacy. We have reviewed three mainstream attack methods (membership inference attacks, model inversion attacks, and model extraction attacks) along with some other attack methods (property inference attacks, model memorization attacks, and privacy violation in data aggregation). We discuss the difference among these attack methods in this subsection.

Table 5 Literature of attack methods for visual deep learning system privacy

Attack [1]	Literature	Attack Target	System Setting	Attacker Knowledge	Datasets
Minf	Shokri et al. (2017)	Dataset	MLaaS	Black-box	CIFAR, MNIST
	Yeom et al. (2018)	Dataset	Centralized	Black-box	CIFAR, MNIST
	Hayes et al. (2019)	Dataset	Centralized	Both	CIFAR, LFW, DR
	Melis et al. (2019)	Dataset	Distributed	Black-box	LFW, FaceScrub, PIPA
	Nasr et al. (2019)	Dataset	Distributed	White-box	CIFAR
	Sablayrolles et al. (2019)	Dataset	Centralized	Both	CIFAR, ImageNet
	Salem et al. (2019)	Dataset	MLaaS	Black-box	CIFAR, MNIST, LFW
	Song et al. (2019b)	Dataset	Centralized	White-box	CIFAR, FMNIST, Yale Face
	Truex et al. (2019)	Dataset	MLaaS	Black-box	CIFAR, MNIST
	Chen et al. (2020b)	Dataset	Centralized	Both	CelebA
	He et al. (2020b)	Dataset	Centralized	Black-box	Cityscapes, BDD100K, Mapillary
	Park and Kang (2020)	Dataset	Centralized	Black-box	Pascal VOC, INRIA, SynthText
	Salem et al. (2020)	Dataset	Distributed	Black-box	CIFAR, MNIST
	Zou et al. (2020)	Dataset	Centralized	Black-box	Caltech, CIFAR, Flowers, PubFig
	Choquette-Choo et al. (2021)	Dataset	Centralized	Black-box	CIFAR, MNIST
	He and Zhang (2021)	Dataset	Centralized	Black-box	CIFAR, STL-10, UTKFace, CelebA
	Li and Zhang (2021)	Dataset	Centralized	Black-box	CIFAR, GTSRB
MInv	Fredrikson et al. (2015)	Dataset	Centralized	Both	AT&T Faces
	Hitaj et al. (2017)	Dataset	Distributed	White-box	MNIST, AT&T Faces
	Yeom et al. (2018)	Dataset	Centralized	Black-box	CIFAR, MNIST
	Yang et al. (2019b)	Dataset	MLaaS	Black-box	CIFAR, MNIST, FaceScrub, CelebA
	Zhu et al. (2019)	Dataset	Distributed	White-box	CIFAR, MNIST, SVHN, LFW
	Geiping et al. (2020)	Dataset	Distributed	White-box	CIFAR, ImageNet
	Wei et al. (2020)	Dataset	Distributed	White-box	CIFAR, MNIST, LFW
	Zhang et al. (2020b)	Dataset	Centralized	White-box	MNIST, ChestX-ray, CelebA
	Zhao et al. (2020a)	Dataset	Distributed	White-box	CIFAR, MNIST, LFW
MExt	Correia-Silva et al. (2018)	Model	MLaaS	Black-box	ImageNet, COCO
	Oh et al. (2018)	Model	Centralized	Black-box	MNIST, ImageNet
	Jagielski et al. (2020)	Model	Centralized	Black-box	CIFAR, MNIST
	Milli et al. (2019)	Model	Centralized	Black-box	CIFAR, MNIST
	Orekondy et al. (2019)	Model	Centralized	Black-box	Caltech256, CUBS200, IScenes, DR
	Barbalau et al. (2020)	Model	Centralized	Black-box	CIFAR, FMNIST, 10MonkeyS CelebA-HQ, ImageNet
	Pal et al. (2020)	Model	Centralized	Black-box	CIFAR, MNIST, GTSRB, ImageNet
	Correia-Silva et al. (2021)	Model	MLaaS	Black-box	ImageNet, COCO, Pascal VOC
Other	Song et al. (2017)	Dataset	Centralized	Both	CIFAR, LFW, FaceScrub
	Ganju et al. (2018)	Dataset	MLaaS	White-box	MNIST, CelebA
	Melis et al. (2019)	Dataset	Distributed	Black-box	LFW, FaceScrub, PIPA
	He and Zhang (2021)	Dataset	Centralized	Black-box	CIFAR, STL-10, UTKFace, CelebA

[1] MInf: Membership inference attacks; MInv: Model inversion attacks; MExt: Model extraction attacks.

A summary of attack methods in visual deep learning system privacy is given in Table 5. This table explains several aspects of these papers: attack targets, system settings, attacker knowledge, and datasets. The papers are then grouped via the type of attacks.

For the aspect of the attack target, the only one to attack deep learning models is model extraction attacks, including stealing parameters, hyperparameters, and structures of the

model. The objective of all the other attack methods is the model's training datasets, as shown in the table.

In terms of system settings, most papers focused their work on a centralized system, while some other papers addressed their issue in a distributed system or MLaaS. Membership inference attacks and model extraction attacks were more likely to be launched in a centralized system or MLaaS. Many papers studied model inversion attacks in federated learning, which belonged to the distributed systems.

For the aspect of the attack knowledge, these attack methods were developed in black-box / white-box settings. Most membership inference attacks and model extraction attacks happened in a black-box setting, while model inversion attacks in a white-box setting. This is because the membership inference attacks and model extraction attacks were usually launched according to the victim model's prediction output. Hence, most of the attackers had black-box access to the victim model. As gradient sharing is utilized for model inversion attacks to reconstruct private information in federated learning, these papers have been classified as white-box attacks, as shown in the table.

In terms of evaluating the attack methods in visual datasets, CIFAR and MNIST in image classification were mainly the first choices when evaluating the attacks. However, there has been a growing trend of applying different attacks in larger datasets and other tasks, such as ImageNet in image classification, COCO in object detection, or Mapillary in semantic segmentation.

There are only a few research papers on model memorization attacks and property inference attacks. We can conclude that the mainstream attack methods are membership inference attacks, model inversion attacks, and model extraction attacks.

Next, we discuss the differences among these attack methods. Membership inference attacks, model inversion attacks, and property inference attacks aim to discover some information in the training data. The difference is that membership inference attacks aim to find whether some specific data samples are in the training data. Model inversion attacks focus on reconstructing some possible data samples in the training data. Property inference attacks try to find statistical features in the training data.

Attackers in visual deep learning systems can be both human or machine. Human attacks tend to happen when the training dataset is available to the attacker, violating data aggregation. Machine-based attacks include other attack methods listed above.

In the next section, defense mechanisms against attacks on visual data and visual deep learning systems are introduced.

4 Defense Mechanisms

In this section, we follow the same structure as Section 3. We discuss defense mechanisms when visual data is the privacy target, followed by defense mechanisms when visual deep learning systems are the privacy target. We do not introduce defense mechanisms specifically against each attack method. Instead, we introduce these defense mechanisms based on algorithms and technologies and discuss how they can be used against these attacks.

As discussed in Section 2.3.3, we categorize visual data privacy as object-level privacy and file-level privacy. Object-level privacy means a specific part of the visual content is private, such as faces, personal belongings, and classified texts. File-level privacy means the whole visual data entry (file) is private. Researchers can protect visual privacy by removing, replacing, or perturbing private objects or whole images and videos.

We categorize visual deep learning system privacy as dataset-level privacy and model-level privacy. Dataset-level privacy means the target model's training data contains private information. Model-level privacy means the deep learning model parameters, hyperparameters or architectures are private. Perturbation-based or encryption-based mechanisms can protect datasets and models.

4.1 Defending Visual Data Privacy

To defend against visual data attacks, we introduce several defense mechanisms divided into five categories in this subsection. We first discuss how to detect privacy in visual data. Then we review research papers regarding three mainstream mechanisms, including GAN-based privacy preservation, and adversarial example-based privacy preservation, differential privacy-based privacy preservation. At last, we introduce other mechanisms which do not belong to these mainstream mechanisms.

4.1.1 Privacy Detection

As a significant activity in social networks, image or video sharing occurs within a small group of friends and sometimes can be discovered by people outside the sharing circle. Some social network websites allow users to change their sharing privacy settings, but this usually applies to all the shared images. The research field for privacy detection aims to intelligently detect privacy in visual data for better photo sharing in social networks. Researchers adopted various methods to detect privacy using different terms to describe this task, such as privacy detection, classification, or prediction. Privacy detection can also be considered as the first step towards privacy preservation.

Research Focus The literature in this category can be split into several groups based on privacy detection levels: simple sharing, personalized sharing, and complex sharing.

The first privacy detection level is to classify images as public or private simply. Tonge and Caragea (2016) adopted deep neural networks to extract deep visual features and making tags for images. Based on the predicted tags, they used a support vector machine classifier to determine images in online social networks as public or private. A similar approach was proposed in Tran et al. (2016), which used hierarchical features (both object and convolutional features) to detect privacy in photos and achieved a better accuracy result. Tonge et al. (2018); Tonge (2018) presented image prediction based on user tags (tags made by users), object tags (tags indicating objects in images), and scene tags (tags showing current scene context in images). Tonge and Caragea (2019) demonstrated a multi-model fusion approach fusing these different tags to predict image privacy accurately. Their continued work in Tonge and Caragea (2020) compared different neural network architectures for privacy prediction.

The second privacy detection level is a personalized privacy detection, where the system determines the images as public or private with a user/group impact. Spyromitros-Xioufis et al. (2016) proposed a personalized image classification system with user-specific feedbacks. By asking users to provide some external public/private photos or exploiting user interaction logs, they achieved performance improvements on their personalized model. Zhong et al. (2017) developed a personalized framework with the concept of privacy groups, which gave privacy results by subsets of users. Orekondy et al. (2017) presented an approach

to provide user customizations according to their individual selections. Based on the user-specific privacy settings in social networks, the system provided appropriate privacy advice.

The third privacy detection level is a complex photo-sharing scheme defined as completely-share, partially-share, share-with-blurring, and not-share. Yu et al. (2017) developed iPrivacy, a framework for recognizing human objects, determining the privacy levels, recommending privacy settings, and blurring human faces. Their work continued in Yu et al. (2018), which provided a deep neural network using feature-based and object-based image sensitiveness representation and user-trustworthiness characterization. The new algorithm determined whether or not to allow users to see the images based on the image privacy setting recommendation.

Discussion In this subsection, we have reviewed different privacy detection mechanisms. Privacy detection has a significant impact on data sharing in online social networks. The detection level for an image-sharing application can be selected based on different requirements and scenarios. A simple application only needs a binary private detection system with public/private result (Tonge and Caragea 2020), while a complex application requirements multiple levels of privacy results (Yu et al. 2018).

The technologies of detecting privacy are mostly deep learning algorithms. Combining learning-based privacy tags with human-annotated tags, the privacy detection system can usually produce better detection results. This research field even takes a step towards privacy personalization. The privacy detection output can be given based on a person's requirements (Zhong et al. 2017).

In the following several subsections, we introduce defense mechanisms in visual data utilizing different technologies after successfully detecting visual data privacy.

4.1.2 Defense with Generative Adversarial Networks

Generative adversarial networks (GANs) were first proposed by Goodfellow et al. (2014). GANs have been widely applied in various domains, such as computer vision (Goodfellow et al. 2014), natural language processing (Jetchev et al. 2017), and semantic segmentation (Dong et al. 2017). With the advantage of efficiently generating data samples, GANs are often adopted for multiple applications, such as plausible image generation (Choe et al. 2017), image-to-image translation (Choi et al. 2018), and image super-resolution (Ledig et al. 2017).

Generating data samples also allows GANs to replace private objects in images with synthetic objects for visual privacy preservation. We review the research focus in different tasks based on GANs.

Research Focus In order to protect visual privacy, most of the methods using GANs have concentrated on the task of face recognition. The objective could be either cropped face images/videos or general images/videos with faces. A few other pieces of literature presented GAN-based privacy-preserving methods in the tasks of facial expression recognition, general object removal, and image storage. GAN-based methods applying to different tasks are presented in the following paragraphs.

As a popular research area in computer vision, face recognition always results in privacy leakage. Shared photos in social media can be attacked to extract a person's identity for malicious usage. In order to protect individual identities in images and videos, GANs are applied to generate fake faces to replace identities, which is often called face de-identification (Samarzija and Ribaric 2014).

The research on face de-identification targeting cropped face images focused on manipulating facial attributes in images. Facial attributes are about the size, color, or existence of facial features, or the expressions on faces, including hair color, big/small nose, beard/no-beard, young/old, smile/no-smile, and so on. Li and Lin (2019) proposed AnonymousNet, a face de-identification method based on both GAN and adversarial examples. They first applied facial attribute prediction to have an attribute distribution. Then an attribute selection and update with perturbation were performed to the attribute distribution. According to the attribute distribution, a GAN-based face-identification was applied, followed by adversarial perturbations added in the generated images. Wang et al. (2021a) presented InfoScrub, a GAN-based approach to manipulate privacy attributes while maintaining the whole appearance of the original facial images. They first performed attribute inversion (change/remove the facial attributes) to obfuscate face identity. Then they generate obfuscated faces by maximizing the uncertainty on the attribute presence. Chen et al. (2021) presented a face data privacy-preserving approach using GAN by generating fake images of faces. They extracted face attribute labels, generated face images, and transformed the fake features to match the original image features. This framework protected the original identity information while preserving the original facial attributes as much as possible. Li and Choi (2021) demonstrated a face image obfuscation approach by blurring the latent space of a generative model, resulting in altering the identity in the image while preserving high quality.

Some other literature also focused on face de-identification, but they applied the methods in general images instead of cropping face images. They aimed to achieve natural-looking synthetic faces without the original identities. In Sun et al. (2018a,b), they removed the individual identification information by GAN-based methods. Sun et al. (2018a) mainly extracted facial pose information from images. Then in Sun et al. (2018b), the output de-identified images were reconstructed with a parametric face generation network using GAN. The disadvantages of these two methods were low-resolution output, no face expression preserved, unnatural image output, and unsuitable for videos. Ren et al. (2018) demonstrated a GAN-based face anonymization for videos. Using face detection and modification methods, along with action detection to track face movement in videos, they protected face identities in visual data. In the work of Gafni et al. (2019), the method handled videos and generated convincing faces. This method could provide videos with natural-looking faces, which was similar to original video inputs. In order to look natural in video data, the faces in this method need to be changed as minimal as possible. By concatenating the representation layer of a face classifier to the latent space of an encoder-decoder, they were able to protect the face identity while preserving both the face pose and expression. Yu et al. (2021b) proposed a GAN-based differentially private image protection framework for street view scenes. Their framework followed three steps: detecting private contents in the image; using GAN to project real private objects into latent space and obtaining the corresponding latent vector; adding noises by differential privacy into the latent vector and generating synthetic contents to replace private objects. Their method demonstrated the effectiveness of replacing human faces and license plates for street view images.

Automatic facial expression recognition in human-machine interaction systems is a crucial module for smart cameras or smart IoT devices. Current recognition methods depend on high-resolution images without protecting visual privacy. Chen et al. (2018) proposed to replace the face identity in an image without degrading the system utility. Leveraging variational generative adversarial networks (VGANs), the framework presented a realistic version of the image with a different identity while recognizing facial expression at the same time.

Applying GANs for object removal was discussed in Shetty et al. (2018). This paper provided a solution to remove small private objects if they were contained in an image,

and removing them would not break the fidelity of the image. Experimenting on general scene images, they proposed a two-staged architecture: a mask generator and an image inpainter to remove objects, a discriminator to check the performance of the generated images. Similarly, Uittenbogaard et al. (2019) developed a framework that automatically removed and moving objects in street-view images using GANs.

The GAN-based image storage was presented in Nguyen et al. (2020), which demonstrated a different scenario where face images were collected in each client and then sent to the server for storage and processing. The images were compressed in the clients by an Autoencoder GAN-based dimension reduction method to achieve data privacy. The attacker who has access to the server can not visualize or reconstruct the visual data because of compression and dimension reduction.

Discussion Utilizing GANs in privacy preservation has been reviewed in this subsection. The feature of generating synthetic data samples makes GANs an excellent tool to replace private content in visual data. GANs have been used to perform face de-identification in both cropped images (Wang et al. 2021a) and in general images (Yu et al. 2021b). The latter is more realistic in real-world applications. There have been research papers focused on leveraging GANs to remove private information besides faces (Shetty et al. 2018). These research papers all demonstrated the approaches of defending visual data using GANs.

One of the problems of GANs is their difficulty in training. Achieving Nash equilibrium during training between the generator and the discriminator is a primary issue, as the generator sometimes fails to learn the data distribution well, resulting in lower performance of image generation (Li et al. 2017). This then leads to bad privacy preservation results, such as synthetic face images with unnatural looking (low utility performance) (Sun et al. 2018b) or failed de-identification (low privacy performance) (Ren et al. 2018). There will be a long way for privacy preservation in visual data based on GANs. In the following subsection, we review other algorithms as a defense mechanism.

4.1.3 Defense with Adversarial Examples

Adversarial examples, first proposed by Szegedy et al. (2014), are a kind of algorithm that generates small perturbations on the input to fool the neural network models. These small perturbations are hardly recognizable by human eyes but fatal to neural network models, leading to reduced performance, such as misclassification on an image classification task.

Since then, various adversarial example methods were developed (Goodfellow et al. 2015; Carlini and Wagner 2017; Moosavi-Dezfooli et al. 2017) and they were applied in multiple computer vision tasks, such as image classification (Li et al. 2019b), face recognition (Chhabra et al. 2018), and semantic segmentation (Poursaeed et al. 2018).

An essential feature of adversarial examples is transferability. Papernot et al. (2016) showed that adversarial examples were likely to transfer from one model to another, making it possible for attackers to create adversarial examples in a black-box setting by having the substitute model. Then attackers were able to deploy adversarial examples to fool target models. As previous work primarily focused on transferability using small datasets, Liu et al. (2017b) conducted the study under large models and large scale datasets. They showed that transferability came from the similarities of decision boundaries, and the targeted transfer was much more challenging than non-targeted.

Even though adversarial examples are considered as attacking methods in the view of security, at the same time, we can apply adversary examples to protect the private contents in visual data (Liu et al. 2019a) by causing private contents to be misclassified in an image

classification task. In this way, adversarial examples can be considered as defense methods to achieve privacy preservation.

Research Focus The research focus utilizing adversarial examples to defend visual data can be divided into privacy preservation in cropped face images, general images, and other applications.

Several recent papers have started to research privacy preservation with adversarial examples. The proliferation of face recognition systems nowadays makes severe privacy leakage in social networks. The attacker can extract people's identities easily with these unprotected images. The adversarial examples were proved to be possible to protect face privacy against state-of-the-art face recognition applications (Liu et al. 2019b). Feng (2020) demonstrated an adversarial perturbation-based approach as privacy defense against face recognition systems. Leveraging two adversarial perturbations (universal ensemble perturbations and k-randomized transparent image overlays that are semantic adversarial perturbations), they successfully preserved the image privacy against recognition models. However, being easily identified by human eyes, their generated adversarial examples were not usable in real applications. Shan et al. (2020) presented Fawkes, a system for adding imperceptible pixel-level changes to the users' photos against face recognition models. The perturbations were added without significantly distorting the original images. Their image clocking techniques were also robust to clock detection algorithms.

Some research did not satisfy at only making adversarial perturbations on cropped face images. Instead, they applied adversarial examples on general images. Liu et al. (2017c) applied adversarial examples on online photo sharing. They generated perturbations to the whole image to resist the automatic detection system. Liu et al. (2019a) employed a framework to protect image privacy using adversarial examples, and they also provided two new metrics to measure image privacy. Shen et al. (2019) addressed privacy concerns for shared photos in social media. Their method added adversarial noises to protect image privacy against deep learning models while making the noises human-imperceptible. Xue et al. (2020) extended the research to large street scene images. They proposed a framework including three steps: defining the private information, identifying private objects and their position, and protecting private objects with adversarial noises. Human faces and licence plates were considered private objects. The object detection models were adopted to find the private objects in the images. Then adversarial noises were only added on private objects in the images. Using their framework, users could protect their individual privacy, while at the same time, the semantic image information was protected.

There have been several papers concerning privacy preservation with adversarial examples in other applications. Sattar et al. (2020) proposed a defense mechanism against body shape extraction. In this human pose application, the task was to have a 3D body shape estimation by making 2D keypoint detection and mapping keypoints to 3D body models. Adversarial perturbations were added to the keypoint detection framework to obfuscate keypoints, resulting in protecting original human poses. Xiao et al. (2020) applied adversarial examples to the image retrieval scenario. The current image retrieval application adopted deep hashing to search images similar to the input query. In this way, the attacker could search the database for private information. Adversarial perturbations were added to the private images to evade the attack.

Discussion These research papers prove that adversarial examples can be used to protect visual privacy effectively. Since adversarial examples have the phenomenon of transferability, adversarial examples crafted based on one deep learning model can also be deployed to other

models (Papernot et al. 2016). In the online photo sharing scenario, adversarial examples can be used in online photos. When an attacker wants to retrieve these photos, adversarial examples can prevent the attacker from extracting sensitive information. Even if the attacker changes attack models, adversarial examples may still work because of transferability.

Many adversarial example-based mechanisms have been conducted in cropped face images for face de-identification (Shan et al. 2020). There are also multiple papers targeting general images using adversarial perturbations (Liu et al. 2019a). Adversarial examples are generally applied to the whole image, but some work conducted experiments of adding adversarial noises on only private objects Xue et al. (2020). Adversarial examples have been applicable for privacy preservation in visual data. In the next subsection, we review research papers regarding differential privacy for defending visual data.

4.1.4 Defense with Differential Privacy

Differential privacy is originally used to protect privacy in database applications (Dwork et al. 2006). It prevents the attacker from obtaining useful information by making queries to the database. However, images and videos are already exposed to the public in online sharing and publishing applications on social networks. Unlike traditional differential privacy, defending visual data using differential privacy needs to add noises directly to image or video data. Unlike defense with GANs or adversarial examples, Using differential privacy gives a provable privacy measurement to the visual data.

Research Focus Fan (2018, 2019) presented approaches of perturbing face image with differential privacy. They made image pixelization and image obfuscation by adding noises with a differential privacy guarantee. However, their methods led to generated images with low quality. Liu et al. (2021b) provided a DP-Image framework where they added noises to the image feature vector in the latent space to modify private information contained in the image and reconstructed the image from the perturbed vector to replace the original image. Being tested on the face dataset, this approach protected face identities while keeping the application utility. Similar work was illustrated in Li and Clifton (2021) to protect face images utilizing differential privacy with the Laplace mechanism. Wen et al. (2021) proposed IdentityDP, where differential privacy perturbation was directly added into the identity representation to ensure image privacy, while the attribute representation was unchanged to preserve visual similarity.

Discussion Applying differential privacy in visual data is different from that in general databases. Differential privacy used in general databases protects a single database record so that database queries do not reveal any record information. In contrast, applying differential privacy in visual data needs to prevent identity information leakage in any image while preserving image quality, as these images are for browsing in social networks (Liu et al. 2021b).

Defending visual privacy with differential privacy is still in its early stage. We discuss more differential privacy as a defense mechanism against visual deep learning systems in Section 4.2.1, where differential privacy can be used to protect dataset privacy or model privacy. In the following subsection, defense with other mechanisms for visual data is introduced.

4.1.5 Defense with Other Mechanisms

Apart from generative adversarial networks, adversarial examples, and differential privacy, a few other mechanisms are involved in defending visual data privacy. We briefly categorize them into learning-based, encryption-based, and policy-based mechanisms.

Learning-based Mechanisms Smart surveillance cameras and smart home cameras provide various services and assistance to users. When the cloud service providers store, organize, and share the uploaded visual data, the attackers may potentially steal or misuse this visual data and look for private information. Wu et al. (2018, 2020) proposed a privacy-preserving video action recognition system in the cloud service to address this issue. Leveraging an adversarial training and evaluation framework, they protected video privacy in this scenario. They also proposed restarting and ensembling strategies to enhance the anonymization ability to defend unseen attack models.

Encryption-based Mechanisms Encrypting the visual data before uploading to the cloud service can prevent privacy leakage to the attackers. However, simple encryption to the visual data makes them unable to browse or share by users. Tajik et al. (2019) presented a thumbnail-preserving encryption (TPE) method for balancing the image privacy and utility. In this method, the user encrypted the image by TPE and uploaded the ciphertext to the cloud service. The cloud service received the encrypted image and provided the thumbnail for browsing and sharing. In this way, the utility of the system was guaranteed without leaking the image privacy.

Policy-based Mechanisms Although photo sharing in online social networks has been a common activity in our daily life, not all people want to be viewed online or to be accidentally included in a photo due to privacy concerns. Li et al. (2019a) proposed a policy-based mechanism to protect privacy in online image sharing. Their framework first extracted factors when the photo was uploaded, followed by a privacy policy setting based on scenarios by the user's associated friends. At last, the people in the photos could be hidden away based on these policies.

4.2 Defending Visual Deep Learning System Privacy

The defense mechanisms for visual deep learning systems are discussed in this subsection, dealing with defense against many different attack methods. We do not introduce defense mechanisms corresponding to each attack method. Instead, defense mechanisms are discussed based on what technologies are adopted.

The literature in this field can be divided into four parts. The first part is differential privacy, a perturbation-based defense mechanism. Differential privacy can control how many noise values are added to the target. The second and the third parts are encryption-based defense mechanisms, including homomorphic encryption and secure multi-party computation. Homomorphic encryption guarantee the encryption of the whole computation process, while secure multi-party computation guarantee the encryption among multiple parties. The last part is other defense mechanisms that cannot be categorized into the above three.

4.2.1 Defense with Differential Privacy

In the field of visual deep learning system privacy, differential privacy can be used in preserving privacy in visual data and visual deep learning models. Similar to Section 4.1.4, differential privacy for system privacy is also different from that for database applications. Using differential privacy to add noises to deep learning models, the privacy in models and training datasets can be well protected.

Research Focus We discuss differential privacy as a defense mechanism in several topics: how to add differentially private noises, differential privacy in the teacher-student model framework, in generative models, in federated learning, in other mechanisms or assumptions.

Differentially private noises can be added in the model gradients or the loss functions in the deep learning model. Abadi et al. (2016) proposed differentially private stochastic gradient descent (DPSGD), which later became a standard for guaranteeing privacy in deep learning models. Utilizing gradient clipping and noise adding in SGD process, they preserved privacy at each iteration during training. They also presented moments accountant (MA) mechanism which computed the overall privacy cost during training. Gradient perturbations were also analyzed in Chen et al. (2020d). Zhao et al. (2020b) presented a collaborative learning system, where they used the functional mechanism to perturb the loss function of the model in the training phase to achieve differential privacy.

Some literature applied differential privacy in the teacher-student model framework. Papernot et al. (2017) proposed Private Aggregation of Teacher Ensembles (PATE), which provided a framework with multiple teacher models and a student model. The teacher models were trained on sensitive data, and the student model predicted the framework output by noising voting among all the teachers without access to any teacher data or parameters. The framework also provided a differential privacy guarantee in the student model. Their improved framework in Papernot et al. (2018) introduced a new noisy aggregation mechanism with a tighter differential privacy guarantee. Wang et al. (2019) presented a private model compression framework using differential privacy and knowledge distillation. The noises were added in the process of knowledge distillation to guarantee data privacy in student models. This paper provided a model compression solution on deep learning models on mobile devices due to the limited device capacity. Zhu et al. (2020b) proposed private-kNN, an algorithm with comparable or better accuracy than PATE. Based on k-nearest neighbor queries to the private dataset from a random subsample scheme, Their design prevented splitting the training dataset.

Generative models can solve the data scarcity issue by generating more samples from the same data distribution. The learned generative distribution density can easily reflect training samples due to the high model dimensions, leading to privacy leakage. As a result, differential privacy can be used to protect privacy during the training stage of GANs. Xie et al. (2018) proposed a differentially private GAN model (DPGAN), where they achieved differential privacy in GANs by adding well-designed perturbations to gradients during the model training phase. Jordon et al. (2019) presented PATE-GAN to protect private training data in GANs by applying PATE framework. Torkzadehmahani et al. (2019) introduced a Differentially Private Conditional GAN (DP-CGAN) framework to improve the model performance while guaranteeing training dataset privacy. Chen et al. (2020a) provided Gradient-sanitized Wasserstein Generative Adversarial Networks (GS-WGAN), which ensured private data a sanitized form with privacy guarantees. The generated samples were proved to be private with good quality.

Differential privacy in collaborative/federated learning also draws growing interest in the research community. Shokri and Shmatikov (2015) designed a privacy-preserving system for collaborative deep learning. The system allowed the clients to selectively shared their model gradients during training. This framework was called distributed selective stochastic gradient descent (DSSGD), which enabled the clients to benefit from other clients without sharing the training data. They also applied differential privacy into DSSGD to prevent privacy leakage. This work inspired later literature on privacy preservation for federated learning. Geyer et al. (2018) achieved client level differential privacy in federated learning, which applied a randomized mechanism during gradient aggregation to hide a client's contribution. Private data aggregation or release in a decentralized setting requires privacy preservation in GANs. Triastcyn and Faltings (2020) proposed FedGP, a federated data synthesis engine based on GANs. They demonstrated an approach for private data release to the federated setting using a weaker empirical measure of privacy. Augenstein et al. (2020) provided methods for training generative models by combining federated learning with differential privacy. With deep generative models synthesizing novel examples, federated learning training and evaluating against distributed data, and both federated learning and differential privacy affording user privacy preservations, this paper provided a well-established system.

Some research papers proposed other mechanisms or assumptions of differential privacy to defend deep learning systems. Phan et al. (2017b) provided differential privacy preservation of deep learning models with an adaptive Laplace mechanism. Phan et al. (2017a) adopted differential privacy in convolutional deep belief networks. Yu et al. (2019) employed concentrated differential privacy in model training to provide a sharper privacy analysis.

Discussion Being a standard for measuring how much privacy is protected, differential privacy has been widely adopted for visual deep learning models. There has been several places of adding differentially private noises: model gradients (Abadi et al. 2016) and model loss functions (Zhao et al. 2020b). PATE-based algorithms are applying different privacy in a teacher-student framework, but they can also be considered as adding noises in model prediction labels (Papernot et al. 2017). Applying differential privacy in GANs and distributed systems are also attracting research focus, providing a privacy guarantee in generative models (Xie et al. 2018) and federated learning (Augenstein et al. 2020).

Many research paper showed that differentially private deep learning models provided privacy protection against membership inference attacks, while offering reduced model utility at the same time (Rahman et al. 2018; Truex et al. 2019; Chen et al. 2020b; He et al. 2020b; Leino and Fredrikson 2020; Park and Kang 2020; Choquette-Choo et al. 2021). Applying differential privacy into model training also resulted in a much higher computation cost because of per-sample gradient modification (Chen et al. 2020b). Differential privacy was also effective against model inversion attacks (Hitaj et al. 2017; Aïvodji et al. 2019; Park et al. 2019; Zhang et al. 2020b), which could dramatically reduce the attack rate.

The main issue of leveraging differential privacy is the balance between model privacy and utility. With differentially private noises are added in the deep learning model, privacy in the model can be guaranteed. However, at the same time, the utility performance is significantly reduced. For example, the accuracy in a classification model drops, making the model fails to classify images. Finding a trade-off between model privacy and utility is a challenge when leveraging differential privacy (Choquette-Choo et al. 2021). In the following two subsections, we introduce two encryption-based defense mechanisms.

4.2.2 Defense with Homomorphic Encryption

As an encryption-based mechanism, homomorphic encryption allows the data owners to encrypt their data locally, and then encrypted data is sent to the server for neural network processing. The encrypted result is given back to the data owners to decrypt locally. Homomorphic encryption can be used as a defense mechanism for visual deep learning system privacy.

Gilad-Bachrach et al. (2016) proposed a method for converting existing neural networks to encrypted networks, called CryptoNets, which could be used to guarantee the neural network process in an encrypted form. The user could send their data in an encrypted form to the neural network, where the data remained confidential. The system could make encrypted predictions and return encrypted data. The user then decrypted the result locally. The whole inference phase of the neural network was under encryption, and the attackers could not gain any information of the data.

Research Focus Most of the papers applied homomorphic encryption in the inference phase, while some applied this mechanism in the training phase.

Gilad-Bachrach et al. (2016) first applied homomorphic encryption process to the inference phase of deep learning. The main issue of CryptoNets was ineffective for deeper neural networks. In order to receive a stable encryption result, Chabanne et al. (2017) applied homomorphic encryption in CNNs by replacing ReLU functions with polynomial approximation combined with batch normalization. Juvekar et al. (2018) proposed Gazelle, a system for the inference stage of deep learning, using a combination of homomorphic encryption and traditional two-party computation. Sanyal et al. (2018) presented TAPAS, a framework that solves the drawback of large amounts of time required to evaluate the large machine learning model when using homomorphic encryption. They found ways to speed up parallelize computation using encrypted data. Bian et al. (2020) presented ENSEI, a protocol to inference images securely using homomorphic frequency-domain convolution. Lou et al. (2020) proposed an automated layer-wise parameter selector to determine homomorphic encryption parameters. Lou and Jiang (2021) demonstrated a privacy-preserving mobile neural network structure using homomorphic encryption during the inference phase.

Unlike the above research, some papers focused on applying homomorphic encryption to the training phase of deep learning. In a server-client system setting, local sensitive data could be unwillingly revealed to the malicious server. Phong et al. (2018) demonstrated that gradients could be encrypted and stored on the server in the process of asynchronous SGD leveraging additively homomorphic encryption. Zhang et al. (2020a) presented BatchCrypt, a system solution for homomorphic encryption in federated learning. They encoded a batch of quantized gradients and encrypted it in one go instead of encrypting individual gradients with total precision.

Discussion We have reviewed the literature on defending visual deep learning system privacy using homomorphic encryption. Being an encryption-based mechanism, homomorphic encryption is applied both in the training phase (Zhang et al. 2020a) and in the inference phase (Gilad-Bachrach et al. 2016). Although encrypting techniques are generally discussed in cryptography, the aim of homomorphic encryption in this survey is to protect private information in deep learning models. Hence, homomorphic encryption can be considered as a privacy-preserving technique for visual deep learning systems.

The main concern of utilizing homomorphic encryption for defending visual deep learning system privacy is the communication and computation cost. The communication cost

means the expense of communication between the client and the server, while the computation cost means the time for encrypting the private data. Phong et al. (2018) tried to balance between the encryption performance and the cost of increased communication from the clients to the server. Sanyal et al. (2018) accelerated the process by sparsifying encrypted computations. The method of Lou et al. (2020) reduced inference latency with still accurate neural network inference. Zhang et al. (2020a) reduced both communication and computation cost. The communication, computation, and privacy should be well considered when utilizing homomorphic encryption. In the next subsection, we introduce another encryption-based defense mechanism.

4.2.3 Defense with Secure Multi-Party Computation

Secure multi-party computation is also an encryption-based mechanism to defend visual deep learning system privacy. When multiple parties, not all of which can be trusted, participate in a training or inference phase of deep learning, they can protect the privacy of the whole system utilizing secure multi-party computation. The data owners do not need to reveal the original input, and the server providers do not need to learn any information beyond the trained model.

Research Focus Similar to homomorphic encryption, secure multi-party computation can also be used either in the training phase or the inference phase. Mohassel and Zhang (2017) proposed SecureML, which preserved privacy in the training phase in a two-server model. The data owners distributed their private data among two non-colluding servers using secure two-party computation. Mohassel and Rindal (2018) presented ABY3, a secure framework for data owners to share sensitive data in a three-server model, which also provided a privacy guarantee in the training phase. Liu et al. (2017a) proposed a method to support privacy-preserving model predictions. Pentyala et al. (2021) utilized secure three-party computation in video classification applications without revealing videos and classification models.

Discussion We have reviewed the literature on secure multi-party computation, which can be used in both the training and inference phase. Secure multi-party computation provides a solution for encrypting data in distributed systems. The participants are varied in different research papers (Mohassel and Zhang 2017; Mohassel and Rindal 2018). This is still a growing research field, as not many related papers have been published. In the next subsection, we introduce some other mechanisms for visual deep learning systems.

4.2.4 Defense with Other Mechanisms

In this subsection, we review those defense mechanisms against visual deep learning system privacy which do not leverage differential privacy, homomorphic encryption, or secure multi-party computation. By taking the life cycle of the deep learning system into consideration, we roughly categorize the remaining defense mechanisms as a defense in the training phase and defense in the inference phase.

Defense in the Training Phase In this category, researchers came up with new training structures or adding additional training steps to defend against various attacks.

Salem et al. (2019) proposed dropout as a defense mechanism. As overfitting was one reason that membership inference attacks were effective, dropout, which prevented overfitting, could defend these attacks. Dropout was also evaluated against membership inference

attacks in multiple papers (Hayes et al. 2019; He et al. 2020b; Leino and Fredrikson 2020; Park and Kang 2020).

The regularization technique is another mechanism to avoid model overfitting. The L_2-norm regularization was often applied against membership inference attacks (Shokri et al. 2017). Membership inference attacks learn the data distributions between member samples and non-member samples. Nasr et al. (2018) proposed an adversarial regularization method, which was an optimization privacy-preserving mechanism, by making the victim model's prediction in the training data indistinguishable from other data samples of the same distribution. This approach mitigated membership inference attacks. The adversarial regularization was also evaluated in (Choquette-Choo et al. 2021)

Shejwalkar and Houmansadr (2021) presented a framework using distillation for membership privacy (DMP), a defense mechanism against membership inference attacks. They utilized knowledge distillation to protect visual deep learning system privacy. He and Zhang (2021) proposed the first privacy-preserving contrastive learning mechanism called Talos using adversarial training. They could successfully reduce the attack rate while maintaining the utility of contrastive models. Wang et al. (2021c) demonstrated an approach of defending membership inference attacks via model weight pruning.

Model inversion attacks in a federated learning setting require gradient sharing between the clients and the server. Approaches of processing model gradients could mitigate such attacks. Zhu et al. (2019) presented gradient compression to defend the attacks by pruning gradients with small magnitudes to zero. Wei et al. (2020) demonstrated two other approaches: they inserted Gaussian or Laplacian noises to the gradients or scheduled and controlled gradient sharing after multiple iterations in the training phase. Model inversion attacks can also be defended via perturbation of data representations. Sun et al. (2021) showed that data representation leakage from gradients was the cause of privacy leakage in federated learning. Hence, their defense by data representation perturbation resulted in data reconstruction failure.

Defense in the Inference Phase As the victim model's prediction results are the targets for some attacks methods, adding noises or perturbations in the prediction results during the inference phase can mitigate these attacks.

Shokri et al. (2017) proposed several mitigation strategies for defending membership inference attacks, including restricting confidence score vectors to top k classes, coarsening precision of the confidence score vectors, or increasing entropy of the confidence score vectors. In a general membership inference attack, confidence scores of a victim model's prediction are utilized to learn the difference between member samples and non-member samples. Jia et al. (2019) proposed MemGuard, a defense mechanism against membership inference attacks via adversarial examples. MemGuard added crafted noises to the confidence score vector in an adversarial-example way, which misled the attack model. MemGuard was also evaluated in Choquette-Choo et al. (2021) against their membership inference attacks. Yang et al. (2020) presented a purifier model to reshape confidence scores to remove redundant information. This process could effectively defend membership inference attacks and model inversion attacks. Wang et al. (2021b) proposed a Mutual Information Regularization based Defense (MID) mechanism, where they limited the information contained in the victim model's prediction, thereby preventing the attacker from inferring the model privacy.

Model extraction attacks can be mitigated by making prediction perturbations. Wang and Gong (2018) proposed to round model parameters as a defense against attacks. Lee et al. (2019) proposed to add smart noise in the output probability to make the attacks fail.

Orekondy et al. (2020) proposed to add bounded perturbations to the attacker's objective, which effectively defended various attack strategies. Kariyappa and Qureshi (2020) proposed adaptive misinformation to produce uncorrelated predictions against multiple model stealing attacks. By proposing approaches of detecting model stealing, we can also defend model extraction attacks in the early stage. Kesarwani et al. (2018) presented a cloud-based model extraction attack monitor and two novel metrics to provide the current knowledge learned by attackers. Juuti et al. (2019) proposed PRADA, a framework for detecting model extraction attacks.

4.3 Discussion on Defense Mechanisms

In this section, we have reviewed various defense mechanisms in visual data and in visual deep learning systems. The visual data privacy subsection begins with approaches of privacy detection, followed by three mainstream defense mechanisms, including GANs, adversarial examples, and differential privacy. At last, some other mechanisms are proposed. For visual deep learning system privacy, we have reviewed defense mechanisms, including differential privacy, homomorphic encryption, and secure multi-party computation, along with some other mechanisms.

Table 6 Literature of privacy detection in visual data

Literature	Datasets	Protection Target	Detection Level
Spyromitros-Xioufis et al. (2016)	PicAlert, YourAlert	Object	Personalized
Tonge and Caragea (2016)	PicAlert	File	Simple
Tran et al. (2016)	PicAlert, own dataset	File	Simple
Orekondy et al. (2017)	VISPR	File	Personalized
Yu et al. (2017)	Own dataset	Object	Complex
Zhong et al. (2017)	Own dataset	Object	Personalized
Tonge et al. (2018)	PicAlert	Object	Simple
Tonge (2018)	N/A	Object	Simple
Yu et al. (2018)	PicAlert, own dataset	Object	Complex
Tonge and Caragea (2019)	PicAlert	Object	Simple
Tonge and Caragea (2020)	Own dataset	Object	Simple

Table 6 compares different papers for privacy detection in visual data. Most papers conducted their experiments on PicAlert (a popular privacy research dataset) or constructed their own datasets. The protection target for privacy detection can be either object-level or file-level. The object-level target means whether to use object-level features to predict privacy. The file-level target means to predict privacy based on the whole image feature. The object-level takes the majority, as illustrated in this table. The detection level means the level of privacy detection output, including simple, personalized, and complex, which is aligned with the review in Section 4.1.1. Based on different scenarios and different settings, the current research varies in this classification.

The comparison of all the defense mechanisms in visual data is given in Table 7. The papers are grouped via different defense mechanisms, including generative adversarial networks, adversarial examples, differential privacy, and other mechanisms. These papers are compared according to several aspects: visual source, defense target, datasets, and descriptions.

Table 7 Literature of defense mechanisms in visual data

Defense [1]	Literature	Visual Source	Defense Target	Datasets	Description
GAN	Chen et al. (2018)	Image	Object	FERG, MUG	Preservation in facial expression recognition
	Ren et al. (2018)	Video	Object	DALY, JHMDB	General face de-identification
	Shetty et al. (2018)	Image	Object	COCO, Pascal VOC, Flickr Logos	Preservation in object removal
	Sun et al. (2018a)	Image	Object	PIPA	General face de-identification
	Sun et al. (2018b)	Image	Object	PIPA	General face de-identification
	Gafni et al. (2019)	Video	Object	CelebA, CelebA-HQ, LFW, PubFig	General face de-identification
	Li and Lin (2019)	Image	File	CelebA	Cropped face de-identification
	Uittenbogaard et al. (2019)	Image	Object	Own dataset	Preservation in object removal
	Nguyen et al. (2020)	Image	File	AT&T Faces, CelebA, Yale Face B	Preservation in image storage
	Chen et al. (2021)	Image	File	CelebA	Cropped face de-identification
	Li and Choi (2021)	Image	File	FFHQ, CelebA	Cropped face de-identification
	Wang et al. (2021a)	Image	File	CelebA	Cropped face de-identification
	Yu et al. (2021b)	Image	Object	Own dataset	General face de-identification
AE	Liu et al. (2017c)	Image	Object	Pascal VOC	Preservation in general images
	Liu et al. (2019a)	Image	File	MNIST, ImageNet	Preservation in general images
	Liu et al. (2019b)	Image	File	MS-Celeb-1M, LFW, VGGFace2, WebFace	Cropped face de-identification
	Shen et al. (2019)	Image	File	Own dataset	Preservation in general images
	Feng (2020)	Image	File	FaceScrub	Cropped face de-identification
	Sattar et al. (2020)	Image	File	3DPW	Preservation in body shape extraction
	Shan et al. (2020)	Image	File	VGGFace2, WebFace, PubFig, FaceScrub	Cropped face de-identification
	Xiao et al. (2020)	Image	File	CIFAR, FMNIST, ImageNet	Preservation in image retrieval
	Xue et al. (2020)	Image	Object	VISPR	Preservation in general images
DP	Fan (2018)	Image	File	PETS2010, Venice	Noise in pixels
	Fan (2019)	Image	File	PIPA	Noise in image transformation vectors
	Liu et al. (2021b)	Image	File	FFHQ	Noise in latent space
	Li and Clifton (2021)	Image	File	FFHQ	Noise in latent space
	Wen et al. (2021)	Image	File	CelebA, CelebA-HQ	Noise in latent space
Other	Wu et al. (2018)	Video	File	SBU kinect, UCF101, VISPR	Learning-based
	Wu et al. (2020)	Video	File	SBU kinect, UCF101, HMDB51, own datasest	Learning-based
	Tajik et al. (2019)	Image	File	Own dataset	Encryption-based
	Li et al. (2019a)	Image	File	N/A	Policy-based

[1] GAN: Generative adversarial network; AE: Adversarial example; DP: Differential privacy.

As a video clip is essentially a sequence of images, it is common sense that research on visual privacy should start from images. As a result, there are more papers concerning images as the visual source than papers concerning videos, which is aligned with our table. Privacy preservation in videos can be more difficult due to the complexity of video data, such as keeping the consistency in adjacent frames.

Table 8 Literature of defense mechanisms in visual deep learning systems

Defense [1]	Literature	Defense Target	System Setting	Datasets
DP	Shokri and Shmatikov (2015)	Dataset	Distributed	MNIST, SVHN
	Abadi et al. (2016)	Dataset	Centralized	CIFAR, MNIST
	Papernot et al. (2017)	Dataset	Centralized	MNIST, SVHN
	Phan et al. (2017a)	Dataset	Centralized	MNIST
	Phan et al. (2017b)	Dataset	Centralized	CIFAR, MNIST
	Geyer et al. (2018)	Dataset	Distributed	MNIST
	Papernot et al. (2018)	Dataset	Centralized	MNIST, SVHN
	Xie et al. (2018)	Dataset	Centralized	MNIST
	Torkzadehmahani et al. (2019)	Dataset	Centralized	MNIST
	Wang et al. (2019)	Dataset	Centralized	CIFAR, MNIST, SVHN
	Yu et al. (2019)	Dataset	Centralized	CIFAR, MNIST
	Augenstein et al. (2020)	Dataset	Distributed	Federated EMNIST
	Chen et al. (2020a)	Dataset	Both	MNIST, FMNIST
	Chen et al. (2020d)	Dataset	Centralized	CIFAR, MNIST
	Triastcyn and Faltings (2020)	Dataset	Distributed	MNIST, CelebA
	Zhao et al. (2020b)	Dataset	Distributed	MNIST, SVHN
	Zhu et al. (2020b)	Dataset	Centralized	MNIST, SVHN, CIFAR, CelebA, Market1501
HE	Gilad-Bachrach et al. (2016)	Dataset	Centralized	MNIST
	Chabanne et al. (2017)	Dataset	Centralized	MNIST
	Juvekar et al. (2018)	Dataset	Centralized	CIFAR, MNIST
	Sanyal et al. (2018)	Dataset	Centralized	MNIST, LFW
	Phong et al. (2018)	Dataset	Centralized	MNIST, SVHN
	Bian et al. (2020)	Dataset	Centralized	CIFAR, MNIST
	Lou et al. (2020)	Dataset	Centralized	CIFAR
	Zhang et al. (2020a)	Dataset	Distributed	CIFAR, FMNIST
	Lou and Jiang (2021)	Dataset	Centralized	CIFAR
SMC	Mohassel and Zhang (2017)	Dataset	Distributed	MNIST, Gisette
	Liu et al. (2017a)	Dataset	Distributed	CIFAR, MNIST
	Mohassel and Rindal (2018)	Dataset	Distributed	MNIST
	Pentyala et al. (2021)	Dataset	Distributed	RAVDESS
Other	Shokri et al. (2017)	Dataset	MLaaS	CIFAR, MNIST
	Nasr et al. (2018)	Dataset	Centralized	CIFAR
	Rahman et al. (2018)	Dataset	Centralized	CIFAR, MNIST
	Jia et al. (2019)	Dataset	Centralized	CH-MNIST
	Juuti et al. (2019)	Model	Centralized	MNIST, GTSRB
	Lee et al. (2019)	Model	Centralized	CIFAR, MNIST, FMNIST, STL-10
	Salem et al. (2019)	Dataset	MLaaS	CIFAR, MNIST, LFW
	Zhu et al. (2019)	Dataset	Distributed	CIFAR, MNIST, SVHN, LFW
	Kariyappa and Qureshi (2020)	Model	Centralized	CIFAR, MNIST, FMNIST, Flowers-17
	Orekondy et al. (2020)	Model	Centralized	CIFAR, MNIST, FMNIST, CUBS-200, Caltech-256
	Wei et al. (2020)	Dataset	Distributed	CIFAR, MNIST, LFW
	Yang et al. (2020)	Dataset	Centralized	CIFAR, FaceScrub
	He and Zhang (2021)	Dataset	Centralized	CIFAR, STL-10, UTKFace, CelebA
	Shejwalkar and Houmansadr (2021)	Dataset	Centralized	CIFAR
	Sun et al. (2021)	Dataset	Distributed	CIFAR, MNIST
	Wang et al. (2021b)	Dataset	Centralized	CIFAR, FaceScrub, CelebA
	Wang et al. (2021c)	Dataset	Centralized	CIFAR, MNIST

[1] DP: Differential privacy; HE: Homomorphic encryption; SMC: Secure multi-party computation.

The defense target in visual data can be either file-level or object-level. The file-level target means that the whole file (image or video) is the defense target, while the object-level target means that only private objects in the file need to be protected. For example, when using adversarial examples as a defense mechanism, adversarial perturbations are added in all pixels of the image or some specific objects.

There are many different datasets adopted in the experiments in these research papers. As faces are often the privacy preservation targets in visual data, many face datasets can be seen in the table. However, there has been a lack of consensus when choosing a face dataset. CelebA and FFHQ are more common than others.

Table 8 contains all the literature of defenses mechanisms in visual deep learning systems. These papers are grouped by different technologies, including differential privacy, homomorphic encryption, secure multi-party computation, and other mechanisms. The information in the table includes defense targets, system settings, and datasets.

In terms of defense targets, most research papers focus on defending training datasets, while only a few papers are about defending models, which are defense mechanisms against model extraction attacks. The system settings can be not only centralized, distributed but also MLaaS. Centralized systems account for the majority. Most of the literature on secure multi-party computation uses distributed systems. Different from papers regarding privacy in visual data, these papers mainly conducted experiments on image classification datasets. CIFAR and MNIST are the prominent datasets in this table.

Next, we discuss several differences among defense mechanisms for visual data privacy and visual deep learning system privacy. One of the main differences between different visual data privacy defense mechanisms is the generated data output. GAN-based and differential privacy-based mechanisms generate a whole new image or part of an image, which is recognizable by human eyes. Adversarial example-based mechanisms add perturbations to the original image, and the output image is indistinguishable by human eyes. Being the only method that can achieve both privacy preservation targets, differential privacy measures how much privacy is protected in visual data and datasets. Different from other mechanisms, homomorphic encryption, and secure multi-party computation are two encryption-based mechanisms. In the next section, we discuss future directions of visual privacy protection.

5 Future Directions

In the previous sections, we have reviewed recent research achievements on attack methods and corresponding defense mechanisms for visual data privacy and visual deep learning system privacy. Nevertheless, the research work of these attack methods and defense mechanisms is not done, as some methods are not applicable in real-world scenarios. In this section, several potential future directions concerning visual privacy are proposed.

Designing Privacy Attacks with More Relaxed Assumptions. For example, membership inference attacks were initially based on strong assumptions, such as using prediction confidence scores and prior information on the victim model to differentiate member samples from non-member samples. The shadow model should be trained based on the same data distribution as the victim model. Recent work relaxed these assumptions and still successfully launch the attacks. The prediction confidence scores were replaced with prediction labels. The shadow model was not required to be trained by the same data distributions. Designing more relaxed assumptions for privacy attacks for more realistic scenarios is a growing trend in future research.

Applying Privacy Attacks in More Domains. Membership inference attacks, model inversion attacks, and model extraction attacks were all originally demonstrated on image classification applications. With the tremendous growth in computer vision, other applications

like face recognition, object detection, and semantic segmentation are all emerged, and private information in these applications also needs protection. The fast development of deep learning makes new paradigms spawn, such as contrastive learning, transfer learning, and federated learning. The models in these paradigms are also vulnerable to various attacks. Research on attack methods in more domains is a growing trend.

Attacks on Visual Deep Learning System Privacy in Both White-Box and Black-Box Settings. The white-box and black-box settings for privacy attacks mean how much access the attacker can get to the deep learning model. Intuitively, white-box attacks are more straightforward to launch because of more prior information on the model. However, there is a growing interest in making few assumptions about the attacks in the research field, which enables more papers to concentrate on black-box attacks. Another point to notice is that some white-box/black-box settings actually depend on the attack scenarios. For example, membership inference attacks are launched in a black-box setting mainly because the attacks rely on model queries. These attacks have also been demonstrated on MLaaS, which is also a black-box setting. In future work, whether to choose a white-box/black-box setting is more about selecting the assumptions and scenarios.

Towards Personalized Privacy Detection. In the area of visual privacy, privacy detection often occurs in image sharing in online social networks. The result of privacy detection has evolved from simple output like public or private to complex output with personalization. The sharing scheme enables different privacy recommendations to different users based on their individual preferences. In the future, this feature can give a more complex recommendation to satisfy any user's requirements.

Leveraging New Deep Learning Technologies to Preserve Visual Privacy. With the fast development of deep learning in computer vision, methods such as GANs and adversarial examples have achieved remarkable results these years. Applications like face recognition are urgently needed for privacy preservation with the popularity of photo sharing in online social networks. The achievements of deep learning technologies can be applied in these applications to solve visual privacy issues. GANs are applied to generate synthetic data such as face images. The recent GAN technology ensures synthetic data with high fidelity, resulting in natural-looking and privacy-preserving face images. Adversarial examples were discovered initially as an attack tool to fail classification models. Leveraging this feature, they are adopted to defend attack models. Applying adversarial perturbations to an image privacy-preserving framework, privacy in visual data can be well protected against the attacker while keeping the application's utility at the same time. Utilizing cutting-edge deep learning technologies can achieve a better privacy preservation result.

Trade-off Between Model Privacy and Utility When Using Differential Privacy. It is an ongoing challenge to develop new deep learning frameworks and models that accommodate the needs of visual privacy preservation. The trade-off between model privacy and utility using differential privacy should be well balanced. When a certain amount of differentially private noise is added to the model, the model's performance is also affected. The model performance drops significantly as more noises are added. Keeping the model utility while providing a privacy-preserving model is still an unsolved problem.

Evaluating Larger Datasets for Visual Privacy. Most of the current works have conducted their experiments in CIFAR and MNIST, which are two popular visual datasets. However, these datasets both have small image sizes and only ten classes. The algorithms tested on these datasets do not have good generalization in real-world applications. Larger, more realistic datasets are beginning to be applied in experiments for visual privacy preservation. CIFAR and MNIST are currently mainstream datasets. This is mainly because researchers want to compare previous works with their methods. We can also see that researchers are beginning to test their algorithms in larger datasets such as ImageNet.

Utilizing Multiple Different Technologies in One Task and Utilizing One Kind of Technology in Multiple Tasks. Nowadays, multiple different technologies are adopted together in one task to preserve privacy in visual data or visual deep learning systems. A framework to solve privacy violation in datasets in distributed systems can require multiple different technologies, including differential privacy (for privacy preservation), GANs (for generating synthetic datasets), and federated learning (for structuring the whole system). What also needs to be addressed is the use of one kind of technology in multiple tasks. Adversarial examples are developed for attacking deep learning models, which on the other hand, are adopted in preserving privacy in visual data. How to apply adversarial examples is based on whether they are in the hands of an attacker or a defender. Being a double-edged sword, deep learning can break or protect visual privacy. In the future, more deep learning technologies will be utilized by visual privacy attacks and defenses.

Breaking the Never-Ending Attack-Defense Cycle. An attacker violates privacy in an unprotected application. A defense mechanism is used in the application to protect privacy. An advanced attacker can break this defense mechanism. This is a never-ending attack-defense cycle. Breaking the cycle requires novel research on attack methods or defense mechanisms, which remains an ongoing challenge.

6 Conclusion

In this survey, we have discussed visual privacy attacks and defenses in the context of deep learning. Related research papers during the past few years have been collected and reviewed. We have divided papers into attack methods and corresponding defense mechanisms, and then further into those focus on visual data or visual deep learning systems as the privacy preservation target.

In terms of attacks, we found that:

- Major attack methods include membership inference attacks, model inversion attacks, and model extraction attacks, while a small group of research focuses on property inference attacks, model memorization attacks, and privacy violation in data aggregation.
- Most attack methods target on datasets, while only a few target on deep learning models. This is because the unprotected training datasets directly contain sensitive information that the attacker wants to retrieve or reconstruct. There are attack methods targeting datasets in a wide range of system settings, including centralized, distributed systems, and MLaaS. On the other hand, attack methods targeting models (model extraction attacks) can only be deployed in centralized systems or MLaaS.
- The major research trends on attack methods include: relaxing attack assumptions, applying attacks in more domains, launching attacks in both white-box and black-box settings.

For defenses, the main findings include:

- Differential privacy is the only measurable defending method with strict mathematical guarantee, but applying differential privacy is still challenging for images and videos;
- Some emerging technologies, including GANs, adversarial examples, differential privacy, homomorphic encryption, and secure multi-party computation, can be utilized to defend the attacks.
- Most privacy preservation mechanisms focus on images, while a few others focus on videos. This is because a video clip is a sequence of images; it makes sense that a large group of research focuses on images in the early stage. Privacy preservation in videos is more difficult due to its additional requirements on temporal and frame consistency.
- Major research trends on defense include: personalized privacy detection, leveraging newly developed deep learning methods in privacy preservation, balancing model privacy and utility, evaluating larger datasets for visual privacy, utilizing multiple technologies in one task and utilizing one kind of technology in multiple tasks, breaking the never-ending attack-defense cycle.

Compared with general privacy, the significance of visual privacy can be summarized into three folds.

- The first is the objective. Visual privacy aims to study the privacy issues in vision (images and videos). Unlike numerical data or tabular data, visual data is a kind of unstructured data. The privacy definition is given in our survey, which can be object-level/file-level in visual data, and dataset-level/model-level in visual deep learning systems.
- The second is the attack and defense mechanisms. Visual privacy attacks and defenses require different technologies that can be applied to images and videos. With the recent booming of deep learning, the most effective attack methods and defense mechanisms are both based on deep learning technologies.
- The third is the applications. Visual privacy issues are discovered in vision-related applications. The tasks and experiment datasets for the mentioned research papers are all summarized and compared in our survey.

We believe our timely study will shed valuable light on the research problems associated with visual privacy attacks and defenses. With the increasing attention paid to this topic, we expect to see increasing research activities in this area.

Acknowledgements This work is supported by an ARC Linkage Project (LP180101150) from the Australian Research Council, Australia.

References

Abadi M, Chu A, Goodfellow I, McMahan HB, Mironov I, Talwar K, Zhang L (2016) Deep Learning with Differential Privacy. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Association for Computing Machinery, pp 308–318

Abdulrahman S, Tout H, Ould-Slimane H, Mourad A, Talhi C, Guizani M (2021) A Survey on Federated Learning: The Journey From Centralized to Distributed On-Site Learning and Beyond. IEEE Internet of Things Journal 8:5476–5497

Acar A, Aksu H, Uluagac AS, Conti M (2018) A Survey on Homomorphic Encryption Schemes: Theory and Implementation. ACM Computing Surveys 51:79:1–79:35

Aifanti N, Papachristou C, Delopoulos A (2010) The MUG facial expression database. In: 11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10, pp 1–4

Aïvodji U, Gambs S, Ther T (2019) GAMIN: An Adversarial Approach to Black-Box Model Inversion. arXiv:190911835 [cs, stat]

AmericanDataPortal (2021) https://www.data.gov
Amiri-Zarandi M, Dara RA, Fraser E (2020) A survey of machine learning-based solutions to protect privacy in the Internet of Things. Computers & Security 96:101921
Aneja D, Colburn A, Faigin G, Shapiro L, Mones B (2017) Modeling Stylized Character Expressions via Deep Learning. In: Computer Vision – ACCV 2016, Springer International Publishing, pp 136–153
Ateniese G, Mancini LV, Spognardi A, Villani A, Vitali D, Felici G (2015) Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers. International Journal of Security and Networks 10:137
Augenstein S, McMahan HB, Ramage D, Ramaswamy S, Kairouz P, Chen M, Mathews R, y Arcas BA (2020) Generative models for effective ML on private, decentralized datasets. In: 8th International Conference on Learning Representations, ICLR 2020, OpenReview.net
Barbalau A, Cosma A, Ionescu RT, Popescu M (2020) Black-Box Ripper: Copying black-box models using generative evolutionary algorithms. Advances in Neural Information Processing Systems 33:20120–20129
Bian S, Wang T, Hiromoto M, Shi Y, Sato T (2020) ENSEI: Efficient Secure Inference via Frequency-Domain Homomorphic Convolution for Privacy-Preserving Visual Recognition. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 9400–9409
Bottou L (1998) Online algorithms and stochastic approximations. In: Online Learning and Neural Networks, Cambridge University Press
Boulemtafes A, Derhab A, Challal Y (2020) A review of privacy-preserving techniques for deep learning. Neurocomputing 384:21–45
Caldas S, Duddu SMK, Wu P, Li T, Konečný J, McMahan HB, Smith V, Talwalkar A (2019) LEAF: A Benchmark for Federated Settings. arXiv:181201097 [cs, stat]
Cao Q, Shen L, Xie W, Parkhi OM, Zisserman A (2018) VGGFace2: A dataset for recognising faces across pose and age. In: 13th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2018, Xi'an, China, May 15-19, 2018, IEEE Computer Society, pp 67–74
Carlini N, Wagner DA (2017) Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017, IEEE Computer Society, pp 39–57
Chabanne H, de Wargny A, Milgram J, Morel C, Prouff E (2017) Privacy-preserving classification on deep neural network. IACR Cryptol ePrint Arch 2017:35
Chen D, Orekondy T, Fritz M (2020a) GS-WGAN: A Gradient-Sanitized Approach for Learning Differentially Private Generators. Advances in Neural Information Processing Systems 33:12673–12684
Chen D, Yu N, Zhang Y, Fritz M (2020b) GAN-Leaks: A Taxonomy of Membership Inference Attacks against Generative Models. In: Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security, Association for Computing Machinery, pp 343–362
Chen J, Konrad J, Ishwar P (2018) VGAN-Based Image Representation Learning for Privacy-Preserving Facial Expression Recognition. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp 1651–165109
Chen T, Kornblith S, Norouzi M, Hinton GE (2020c) A simple framework for contrastive learning of visual representations. In: Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event, PMLR, vol 119, pp 1597–1607
Chen X, Fang H, Lin TY, Vedantam R, Gupta S, Dollar P, Zitnick CL (2015) Microsoft COCO Captions: Data Collection and Evaluation Server. arXiv:150400325 [cs]
Chen X, Wu SZ, Hong M (2020d) Understanding Gradient Clipping in Private SGD: A Geometric Perspective. Advances in Neural Information Processing Systems 33:13773–13782
Chen Z, Zhu T, Xiong P, Wang C, Ren W (2021) Privacy preservation for image data: A GAN-based method. International Journal of Intelligent Systems 36:1668–1685
Chhabra S, Singh R, Vatsa M, Gupta G (2018) Anonymizing k facial attributes via adversarial perturbations. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden, ijcai.org, pp 656–662
Choe J, Park S, Kim K, Park JH, Kim D, Shim H (2017) Face Generation for Low-Shot Learning Using Generative Adversarial Networks. In: 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), pp 1940–1948
Choi Y, Choi M, Kim M, Ha JW, Kim S, Choo J (2018) StarGAN: Unified Generative Adversarial Networks for Multi-domain Image-to-Image Translation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 8789–8797
Choquette-Choo CA, Tramer F, Carlini N, Papernot N (2021) Label-Only Membership Inference Attacks. In: International Conference on Machine Learning

Coates A, Ng AY, Lee H (2011) An analysis of single-layer networks in unsupervised feature learning. In: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011, JMLR.org, vol 15, pp 215–223

Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P (2011) Natural Language Processing (Almost) from Scratch. Journal of Machine Learning Research 12:2493–2537

Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, Franke U, Roth S, Schiele B (2016) The Cityscapes Dataset for Semantic Urban Scene Understanding. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 3213–3223

Correia-Silva JR, Berriel RF, Badue C, de Souza AF, Oliveira-Santos T (2018) Copycat CNN: Stealing Knowledge by Persuading Confession with Random Non-Labeled Data. In: 2018 International Joint Conference on Neural Networks (IJCNN), pp 1–8

Correia-Silva JR, Berriel RF, Badue C, De Souza AF, Oliveira-Santos T (2021) Copycat CNN: Are random non-Labeled data enough to steal knowledge from black-box models? Pattern Recognition 113:107830

Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), IEEE, vol 1, pp 886–893

Dong H, Yu S, Wu C, Guo Y (2017) Semantic Image Synthesis via Adversarial Learning. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp 5707–5715

Dwork C, Kenthapadi K, McSherry F, Mironov I, Naor M (2006) Our Data, Ourselves: Privacy Via Distributed Noise Generation. In: Advances in Cryptology - EUROCRYPT 2006, Springer, pp 486–503

Enthoven D, Al-Ars Z (2020) An Overview of Federated Deep Learning Privacy Attacks and Defensive Strategies. arXiv:200404676 [cs, stat]

European Parliament (2016) EU directive 2016/679 - general data protection regulation (GDPR). Official Journal of the European Union 2014

Everingham M, Van Gool L, Williams CKI, Winn J, Zisserman A (2012) The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results

Fan L (2018) Image Pixelization with Differential Privacy. In: Data and Applications Security and Privacy XXXII, Springer International Publishing, pp 148–162

Fan L (2019) Practical Image Obfuscation with Provable Privacy. In: 2019 IEEE International Conference on Multimedia and Expo (ICME), pp 784–789

Feng Wc (2020) On the (Im)Practicality of Adversarial Perturbation for Image Privacy. Proceedings on Privacy Enhancing Technologies 2021:85–106

Ferryman J, Ellis A (2010) PETS2010: Dataset and Challenge. In: 2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance, pp 143–150

Fredrikson M, Lantz E, Jha S, Lin S, Page D, Ristenpart T (2014) Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing. In: 23rd USENIX Security Symposium (USENIX Security 14), pp 17–32

Fredrikson M, Jha S, Ristenpart T (2015) Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. In: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, Association for Computing Machinery, pp 1322–1333

Gafni O, Wolf L, Taigman Y (2019) Live Face De-Identification in Video. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp 9377–9386

Ganju K, Wang Q, Yang W, Gunter CA, Borisov N (2018) Property Inference Attacks on Fully Connected Neural Networks using Permutation Invariant Representations. In: Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, Association for Computing Machinery, pp 619–633

Geiping J, Bauermeister H, Dröge H, Moeller M (2020) Inverting Gradients - How easy is it to break privacy in federated learning? Advances in Neural Information Processing Systems 33:16937–16947

Gentry C (2009) Fully homomorphic encryption using ideal lattices. In: Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing, Association for Computing Machinery, pp 169–178

Geyer RC, Klein T, Nabi M (2018) Differentially Private Federated Learning: A Client Level Perspective. arXiv:171207557 [cs, stat]

Gilad-Bachrach R, Dowlin N, Laine K, Lauter K, Naehrig M, Wernsing J (2016) CryptoNets: Applying Neural Networks to Encrypted Data with High Throughput and Accuracy. In: International Conference on Machine Learning, PMLR, pp 201–210

Goldreich O, Micali S, Wigderson A (1987) How to play ANY mental game. In: Proceedings of the Nineteenth Annual ACM Symposium on Theory of Computing, Association for Computing Machinery, pp 218–229

Gong M, Xie Y, Pan K, Feng K, Qin A (2020) A Survey on Differentially Private Machine Learning [Review Article]. IEEE Computational Intelligence Magazine 15:49–64

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative Adversarial Nets. Advances in Neural Information Processing Systems 27

Goodfellow IJ, Shlens J, Szegedy C (2015) Explaining and harnessing adversarial examples. In: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings

Griffin G, Holub A, Perona P (2007) Caltech-256 object category dataset

Guo Y, Zhang L, Hu Y, He X, Gao J (2016) MS-Celeb-1M: A dataset and benchmark for large-scale face recognition. In: Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, the Netherlands, October 11-14, 2016, Proceedings, Part III, Springer, vol 9907, pp 87–102

Gupta A, Vedaldi A, Zisserman A (2016) Synthetic data for text localisation in natural images. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, IEEE Computer Society, pp 2315–2324

Guyon I, Gunn S, Ben-Hur A, Dror G (2004) Result Analysis of the NIPS 2003 Feature Selection Challenge. Advances in Neural Information Processing Systems 17

Ha T, Dang TK, Le H, Truong TA (2020) Security and Privacy Issues in Deep Learning: A Brief Review. SN Computer Science 1:253

Hayes J, Melis L, Danezis G, Cristofaro ED (2019) LOGAN: Membership Inference Attacks Against Generative Models. Proceedings on Privacy Enhancing Technologies 2019:133–152

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, IEEE Computer Society, pp 770–778

He X, Zhang Y (2021) Quantifying and Mitigating Privacy Risks of Contrastive Learning. arXiv:210204140 [cs]

He Y, Meng G, Chen K, Hu X, He J (2020a) Towards Security Threats of Deep Learning Systems: A Survey. IEEE Transactions on Software Engineering pp 1–1

He Y, Rahimian S, Schiele B, Fritz M (2020b) Segmentations-Leak: Membership Inference Attacks and Defenses in Semantic Image Segmentation. In: Computer Vision – ECCV 2020, Springer International Publishing, pp 519–535

Hinton GE, Srivastava N, Krizhevsky A, Sutskever I, Salakhutdinov RR (2012) Improving neural networks by preventing co-adaptation of feature detectors. arXiv:12070580 [cs]

Hitaj B, Ateniese G, Perez-Cruz F (2017) Deep Models Under the GAN: Information Leakage from Collaborative Deep Learning. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, Association for Computing Machinery, pp 603–618

Hu H, Salcic Z, Dobbie G, Zhang X (2021) Membership Inference Attacks on Machine Learning: A Survey. arXiv:210307853 [cs]

Huang G, Mattar M, Berg T, Learned-Miller E (2008) Labeled faces in the wild: A database for studying face recognition in unconstrained environments

Ioffe S, Szegedy C (2015) Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In: International Conference on Machine Learning, PMLR, pp 448–456

Issa I, Wagner AB, Kamath S (2020) An Operational Approach to Information Leakage. IEEE Transactions on Information Theory 66:1625–1657

Jagielski M, Carlini N, Berthelot D, Kurakin A, Papernot N (2020) High Accuracy and High Fidelity Extraction of Neural Networks. In: 29th USENIX Security Symposium (USENIX Security 20), pp 1345–1362

Jere MS, Farnan T, Koushanfar F (2021) A Taxonomy of Attacks on Federated Learning. IEEE Security Privacy 19:20–28

Jetchev N, Bergmann U, Vollgraf R (2017) Texture Synthesis with Spatial Generative Adversarial Networks. arXiv:161108207 [cs, stat]

Jhuang H, Gall J, Zuffi S, Schmid C, Black MJ (2013) Towards Understanding Action Recognition. In: 2013 IEEE International Conference on Computer Vision, pp 3192–3199

Jia J, Salem A, Backes M, Zhang Y, Gong NZ (2019) MemGuard: Defending against Black-Box Membership Inference Attacks via Adversarial Examples. In: Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, Association for Computing Machinery, pp 259–274

Jordon J, Yoon J, van der Schaar M (2019) PATE-GAN: Generating synthetic data with differential privacy guarantees. In: 7th International Conference on Learning Representations, ICLR 2019, OpenReview.net

Juuti M, Szyller S, Marchal S, Asokan N (2019) PRADA: Protecting Against DNN Model Stealing Attacks. In: 2019 IEEE European Symposium on Security and Privacy (EuroS P), pp 512–527

Juvekar C, Vaikuntanathan V, Chandrakasan A (2018) GAZELLE: A Low Latency Framework for Secure Neural Network Inference. In: 27th USENIX Security Symposium (USENIX Security 18), pp 1651–1669

Kaggle (2021a) 10 Monkey Species. https://www.kaggle.com/slothkong/10-monkey-species

Kaggle (2021b) Diabetic Retinopathy Detection. https://www. kaggle.com/c/diabetic-retinopathy-detection#references

Kairouz P, McMahan HB, Avent B, Bellet A, Bennis M, Bhagoji AN, Bonawitz K, Charles Z, Cormode G, Cummings R, D'Oliveira RGL, Eichner H, Rouayheb SE, Evans D, Gardner J, Garrett Z, Gascón A, Ghazi B, Gibbons PB, Gruteser M, Harchaoui Z, He C, He L, Huo Z, Hutchinson B, Hsu J, Jaggi M, Javidi T, Joshi G, Khodak M, Konecný J, Korolova A, Koushanfar F, Koyejo S, Lepoint T, Liu Y, Mittal P, Mohri M, Nock R, Özgür A, Pagh R, Qi H, Ramage D, Raskar R, Raykova M, Song D, Song W, Stich SU, Sun Z, Suresh AT, Tramèr F, Vepakomma P, Wang J, Xiong L, Xu Z, Yang Q, Yu FX, Yu H, Zhao S (2021) Advances and Open Problems in Federated Learning. Foundations and Trends® in Machine Learning 14:1–210

Kalantidis Y, Pueyo LG, Trevisiol M, van Zwol R, Avrithis Y (2011) Scalable triangulation-based logo recognition. In: Proceedings of the 1st International Conference on Multimedia Retrieval, ICMR 2011, Trento, Italy, April 18 - 20, 2011, ACM, p 20

Kariyappa S, Qureshi MK (2020) Defending Against Model Stealing Attacks With Adaptive Misinformation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 767–775

Karras T, Aila T, Laine S, Lehtinen J (2018) Progressive growing of GANs for improved quality, stability, and variation. In: 6th International Conference on Learning Representations, ICLR 2018, OpenReview.net

Karras T, Laine S, Aila T (2019) A Style-Based Generator Architecture for Generative Adversarial Networks. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 4396–4405

Kaya Y, Dumitras T (2021) When Does Data Augmentation Help With Membership Inference Attacks? In: International Conference on Machine Learning

Kesarwani M, Mukhoty B, Arya V, Mehta S (2018) Model Extraction Warning in MLaaS Paradigm. In: Proceedings of the 34th Annual Computer Security Applications Conference, Association for Computing Machinery, pp 371–380

Konečný J, McMahan HB, Ramage D, Richtárik P (2016) Federated Optimization: Distributed Machine Learning for On-Device Intelligence. arXiv:161002527 [cs]

Konečný J, McMahan HB, Yu FX, Richtárik P, Suresh AT, Bacon D (2017) Federated Learning: Strategies for Improving Communication Efficiency. arXiv:161005492 [cs]

Krizhevsky A (2009) Learning multiple layers of features from tiny images

Kuehne H, Jhuang H, Garrote E, Poggio T, Serre T (2011) HMDB: A large video database for human motion recognition. In: 2011 International Conference on Computer Vision, pp 2556–2563

Kumar N, Berg AC, Belhumeur PN, Nayar SK (2009) Attribute and simile classifiers for face verification. In: IEEE 12th International Conference on Computer Vision, ICCV 2009, Kyoto, Japan, September 27 - October 4, 2009, IEEE Computer Society, pp 365–372

Leal-Taixé L, Milan A, Reid I, Roth S, Schindler K (2015) MOTChallenge 2015: Towards a Benchmark for Multi-Target Tracking. arXiv:150401942 [cs]

LeCun Y, Cortes C (2005) The mnist database of handwritten digits

LeCun Y, Kavukcuoglu K, Farabet C (2010) Convolutional networks and applications in vision. In: Proceedings of 2010 IEEE International Symposium on Circuits and Systems, pp 253–256

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521:436–444

Ledig C, Theis L, Huszár F, Caballero J, Cunningham A, Acosta A, Aitken A, Tejani A, Totz J, Wang Z, Shi W (2017) Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 105–114

Lee KC, Ho J, Kriegman DJ (2005) Acquiring linear subspaces for face recognition under variable lighting. IEEE Transactions on Pattern Analysis and Machine Intelligence 27:684–698

Lee T, Edwards B, Molloy I, Su D (2019) Defending Against Neural Network Model Stealing Attacks Using Deceptive Perturbations. In: 2019 IEEE Security and Privacy Workshops (SPW), pp 43–49

Leino K, Fredrikson M (2020) Stolen Memories: Leveraging Model Memorization for Calibrated White-Box Membership Inference. In: 29th USENIX Security Symposium (USENIX Security 20), pp 1605–1622

Leung MKK, Xiong HY, Lee LJ, Frey BJ (2014) Deep learning of the tissue-regulated splicing code. Bioinformatics 30:i121–i129

Li F, Sun Z, Li A, Niu B, Li H, Cao G (2019a) HideMe: Privacy-Preserving Photo Sharing on Social Networks. In: IEEE INFOCOM 2019 - IEEE Conference on Computer Communications, pp 154–162

Li T, Choi MS (2021) DeepBlur: A Simple and Effective Method for Natural Image Obfuscation. arXiv:210402655 [cs]

Li T, Clifton C (2021) Differentially Private Imaging via Latent Space Manipulation. arXiv:210305472 [cs]

Li T, Lin L (2019) AnonymousNet: Natural Face De-Identification With Measurable Privacy. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp 56–65

Li T, Sahu AK, Talwalkar A, Smith V (2020) Federated Learning: Challenges, Methods, and Future Directions. IEEE Signal Processing Magazine 37:50–60

Li Y, Schwing A, Wang KC, Zemel R (2017) Dualing GANs. Advances in Neural Information Processing Systems 30

Li Y, Li L, Wang L, Zhang T, Gong B (2019b) NATTACK: Learning the distributions of adversarial examples for an improved black-box attack on deep neural networks. In: Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, PMLR, vol 97, pp 3866–3876

Li Z, Zhang Y (2021) Membership Leakage in Label-Only Exposures. In: ACM SIGSAC Conference on Computer and Communications Security (CCS 2021)

Lim WYB, Luong NC, Hoang DT, Jiao Y, Liang YC, Yang Q, Niyato D, Miao C (2020) Federated Learning in Mobile Edge Networks: A Comprehensive Survey. IEEE Communications Surveys Tutorials 22:2031–2063

Lin Y, Han S, Mao H, Wang Y, Dally B (2018) Deep gradient compression: Reducing the communication bandwidth for distributed training. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings, OpenReview.net

Liu B, Ding M, Zhu T, Xiang Y, Zhou W (2019a) Adversaries or allies? Privacy and deep learning in big data era. Concurrency and Computation: Practice and Experience 31

Liu B, Xiong J, Wu Y, Ding M, Wu CM (2019b) Protecting Multimedia Privacy from Both Humans and AI. In: 2019 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), pp 1–6

Liu B, Ding M, Shaham S, Rahayu W, Farokhi F, Lin Z (2021a) When Machine Learning Meets Privacy: A Survey and Outlook. ACM Computing Surveys 54:31:1–31:36

Liu B, Ding M, Xue H, Zhu T, Ye D, Song L, Zhou W (2021b) DP-Image: Differential Privacy for Image Data in Feature Space. arXiv:210307073 [cs]

Liu C, Zhu T, Zhang J, Zhou W (2021c) Privacy Intelligence: A Survey on Image Privacy in Online Social Networks. arXiv:200812199 [cs]

Liu J, Juuti M, Lu Y, Asokan N (2017a) Oblivious Neural Network Predictions via MiniONN Transformations. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, Association for Computing Machinery, pp 619–631

Liu X, Xie L, Wang Y, Zou J, Xiong J, Ying Z, Vasilakos AV (2021d) Privacy and Security Issues in Deep Learning: A Survey. IEEE Access 9:4566–4593

Liu Y, Chen X, Liu C, Song D (2017b) Delving into transferable adversarial examples and black-box attacks. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings, OpenReview.net

Liu Y, Zhang W, Yu N (2017c) Protecting Privacy in Shared Photos via Adversarial Examples Based Stealth. Security and Communication Networks 2017:e1897438

Liu Z, Luo P, Wang X, Tang X (2015) Deep Learning Face Attributes in the Wild. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp 3730–3738

Livingstone S, Russo F (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE

Lo SK, Lu Q, Wang C, Paik HY, Zhu L (2021) A Systematic Literature Review on Federated Machine Learning: From a Software Engineering Perspective. ACM Computing Surveys 54:95:1–95:39

Lou Q, Jiang L (2021) HEMET: A Homomorphic-Encryption-Friendly Privacy-Preserving Mobile Neural Network Architecture. In: International Conference on Machine Learning

Lou Q, Bian S, Jiang L (2020) AutoPrivacy: Automated Layer-wise Parameter Selection for Secure Neural Network Inference. Advances in Neural Information Processing Systems 33:8638–8647

McMahan B, Moore E, Ramage D, Hampson S, y Arcas BA (2017) Communication-Efficient Learning of Deep Networks from Decentralized Data. In: Artificial Intelligence and Statistics, PMLR, pp 1273–1282

Melis L, Song C, De Cristofaro E, Shmatikov V (2019) Exploiting Unintended Feature Leakage in Collaborative Learning. In: 2019 IEEE Symposium on Security and Privacy (SP), pp 691–706

Mikolov T, Deoras A, Povey D, Burget L, Černocký J (2011) Strategies for training large scale neural network language models. In: 2011 IEEE Workshop on Automatic Speech Recognition Understanding, pp 196–201

Milli S, Schmidt L, Dragan AD, Hardt M (2019) Model Reconstruction from Model Explanations. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, pp 1–9

Mireshghallah F, Taram M, Vepakomma P, Singh A, Raskar R, Esmaeilzadeh H (2020) Privacy in Deep Learning: A Survey. arXiv:200412254 [cs, stat]

Mohassel P, Rindal P (2018) ABY3: A Mixed Protocol Framework for Machine Learning. In: Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, Association for

Computing Machinery, pp 35–52
Mohassel P, Zhang Y (2017) SecureML: A System for Scalable Privacy-Preserving Machine Learning. In: 2017 IEEE Symposium on Security and Privacy (SP), pp 19–38
Moosavi-Dezfooli SM, Fawzi A, Fawzi O, Frossard P (2017) Universal adversarial perturbations. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, IEEE Computer Society, pp 86–94
Nasr M, Shokri R, Houmansadr A (2018) Machine Learning with Membership Privacy using Adversarial Regularization. In: Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, Association for Computing Machinery, pp 634–646
Nasr M, Shokri R, Houmansadr A (2019) Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning. In: 2019 IEEE Symposium on Security and Privacy (SP), pp 739–753
Netzer Y, Wang T, Coates A, Bissacco A, Wu B, Ng A (2011) Reading digits in natural images with unsupervised feature learning
Neuhold G, Ollmann T, Bulò SR, Kontschieder P (2017) The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp 5000–5009
Ng H, Winkler S (2014) A data-driven approach to cleaning large face datasets. In: 2014 IEEE International Conference on Image Processing, ICIP 2014, IEEE, pp 343–347
Nguyen H, Zhuang D, Wu PY, Chang M (2020) AutoGAN-based dimension reduction for privacy preservation. Neurocomputing 384:94–103
Nilsback ME, Zisserman A (2006) A Visual Vocabulary for Flower Classification. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), vol 2, pp 1447–1454
Oh SJ, Benenson R, Fritz M, Schiele B (2016) Faceless Person Recognition: Privacy Implications in Social Media. In: Computer Vision – ECCV 2016, Springer International Publishing, pp 19–35
Oh SJ, Augustin M, Fritz M, Schiele B (2018) Towards reverse-engineering black-box neural networks. In: 6th International Conference on Learning Representations, ICLR 2018, OpenReview.net
Orekondy T, Schiele B, Fritz M (2017) Towards a Visual Privacy Advisor: Understanding and Predicting Privacy Risks in Images. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp 3706–3715
Orekondy T, Schiele B, Fritz M (2019) Knockoff Nets: Stealing Functionality of Black-Box Models. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 4949–4958
Orekondy T, Schiele B, Fritz M (2020) Prediction poisoning: Towards defenses against DNN model stealing attacks. In: 8th International Conference on Learning Representations, ICLR 2020, OpenReview.net
Pal S, Gupta Y, Shukla A, Kanade A, Shevade S, Ganapathy V (2020) ActiveThief: Model Extraction Using Active Learning and Unannotated Public Data. Proceedings of the AAAI Conference on Artificial Intelligence 34:865–872
Papernot N, McDaniel P, Goodfellow I (2016) Transferability in Machine Learning: From Phenomena to Black-Box Attacks using Adversarial Samples. arXiv:160507277 [cs]
Papernot N, Abadi M, Erlingsson Ú, Goodfellow IJ, Talwar K (2017) Semi-supervised knowledge transfer for deep learning from private training data. In: 5th International Conference on Learning Representations, ICLR 2017, OpenReview.net
Papernot N, Song S, Mironov I, Raghunathan A, Talwar K, Erlingsson Ú (2018) Scalable private learning with PATE. In: 6th International Conference on Learning Representations, ICLR 2018, OpenReview.net
Park C, Hong D, Seo C (2019) An Attack-Based Evaluation Method for Differentially Private Learning Against Model Inversion Attack. IEEE Access 7:124988–124999
Park Y, Kang M (2020) Membership Inference Attacks Against Object Detection Models. arXiv:200104011 [cs]
Pavan Kumar MR, Jayagopal P (2021) Generative adversarial networks: A survey on applications and challenges. International Journal of Multimedia Information Retrieval 10:1–24
Pentyala S, Dowsley R, De Cock M (2021) Privacy-Preserving Video Classification with Convolutional Neural Networks. In: International Conference on Machine Learning
Phan N, Wu X, Dou D (2017a) Preserving differential privacy in convolutional deep belief networks. Machine Learning 106:1681–1704
Phan N, Wu X, Hu H, Dou D (2017b) Adaptive Laplace Mechanism: Differential Privacy Preservation in Deep Learning. In: 2017 IEEE International Conference on Data Mining (ICDM), pp 385–394
Phong LT, Aono Y, Hayashi T, Wang L, Moriai S (2018) Privacy-Preserving Deep Learning via Additively Homomorphic Encryption. IEEE Transactions on Information Forensics and Security 13:1333–1345
Poursaeed O, Katsman I, Gao B, Belongie SJ (2018) Generative adversarial perturbations. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June

18-22, 2018, IEEE Computer Society, pp 4422–4431
Quattoni A, Torralba A (2009) Recognizing indoor scenes. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp 413–420
Rahman MA, Rahman T, Laganière R, Mohammed N (2018) Membership inference attack against differentially private deep learning model. Trans Data Priv 11:61–79
Ren Z, Lee YJ, Ryoo MS (2018) Learning to Anonymize Faces for Privacy Preserving Action Detection. In: Computer Vision – ECCV 2018, Springer International Publishing, pp 639–655
Rivest RL, Dertouzos ML (1978) On data banks and privacy homomorphisms
Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein MS, Berg AC, Li FF (2015) ImageNet large scale visual recognition challenge. International Journal of Computer Vision 115:211–252
Sablayrolles A, Douze M, Schmid C, Ollivier Y, Jegou H (2019) White-box vs Black-box: Bayes Optimal Strategies for Membership Inference. In: International Conference on Machine Learning, PMLR, pp 5558–5567
Saeidian S, Cervia G, Oechtering TJ, Skoglund M (2021) Quantifying Membership Privacy via Information Leakage. IEEE Transactions on Information Forensics and Security 16:3096–3108
Salem A, Zhang Y, Humbert M, Berrang P, Fritz M, Backes M (2019) ML-Leaks: Model and data independent membership inference attacks and defenses on machine learning models. In: 26th Annual Network and Distributed System Security Symposium, NDSS 2019, The Internet Society
Salem A, Bhattacharya A, Backes M, Fritz M, Zhang Y (2020) Updates-Leak: Data Set Inference and Reconstruction Attacks in Online Learning. In: 29th USENIX Security Symposium (USENIX Security 20), p 1291
Samaraweera GD, Chang JM (2021) Security and Privacy Implications on Database Systems in Big Data Era: A Survey. IEEE Transactions on Knowledge and Data Engineering 33:239–258
Samaria F, Harter A (1994) Parameterisation of a stochastic model for human face identification. Proceedings of 1994 IEEE Workshop on Applications of Computer Vision pp 138–142
Samarzija B, Ribaric S (2014) An approach to the de-identification of faces in different poses. In: 37th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2014, IEEE, pp 1246–1251
Sanyal A, Kusner M, Gascon A, Kanade V (2018) TAPAS: Tricks to Accelerate (encrypted) Prediction As a Service. In: International Conference on Machine Learning, PMLR, pp 4490–4499
Sattar H, Krombholz K, Pons-Moll G, Fritz M (2020) Body Shape Privacy in Images: Understanding Privacy and Preventing Automatic Shape Extraction. In: Computer Vision – ECCV 2020 Workshops, Springer International Publishing, pp 411–428
Serban A, Poll E, Visser J (2020) Adversarial Examples on Object Recognition: A Comprehensive Survey. ACM Computing Surveys 53:66:1–66:38
Shan S, Wenger E, Zhang J, Li H, Zheng H, Zhao BY (2020) Fawkes: Protecting Privacy against Unauthorized Deep Learning Models. In: 29th USENIX Security Symposium (USENIX Security 20), pp 1589–1604
Shejwalkar V, Houmansadr A (2021) Membership Privacy for Machine Learning Models Through Knowledge Transfer. Proceedings of the AAAI Conference on Artificial Intelligence 35:9549–9557
Shen Z, Fan S, Wong Y, Ng TT, Kankanhalli M (2019) Human-imperceptible Privacy Protection Against Machines. In: Proceedings of the 27th ACM International Conference on Multimedia, Association for Computing Machinery, pp 1119–1128
Shetty RR, Fritz M, Schiele B (2018) Adversarial Scene Editing: Automatic Object Removal from Weak Supervision. Advances in Neural Information Processing Systems 31
Shokri R, Shmatikov V (2015) Privacy-Preserving Deep Learning. In: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, Association for Computing Machinery, pp 1310–1321
Shokri R, Stronati M, Song C, Shmatikov V (2017) Membership Inference Attacks Against Machine Learning Models. In: 2017 IEEE Symposium on Security and Privacy (SP), pp 3–18
Song C, Raghunathan A (2020) Information Leakage in Embedding Models. In: Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security, Association for Computing Machinery, pp 377–390
Song C, Ristenpart T, Shmatikov V (2017) Machine Learning Models that Remember Too Much. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, Association for Computing Machinery, pp 587–601
Song L, Shokri R, Mittal P (2019a) Membership Inference Attacks Against Adversarially Robust Deep Learning Models. In: 2019 IEEE Security and Privacy Workshops (SPW), pp 50–56
Song L, Shokri R, Mittal P (2019b) Privacy Risks of Securing Machine Learning Models against Adversarial Examples. In: Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications

Security, Association for Computing Machinery, pp 241–257
Soomro K, Zamir AR, Shah M (2012) UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. arXiv:12120402 [cs]
Spyromitros-Xioufis E, Papadopoulos S, Popescu A, Kompatsiaris Y (2016) Personalized Privacy-aware Image Classification. In: Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, Association for Computing Machinery, pp 71–78
Stallkamp J, Schlipsing M, Salmen J, Igel C (2012) Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. Neural Networks 32:323–332
Sun J, Li A, Wang B, Yang H, Li H, Chen Y (2021) Soteria: Provable Defense Against Privacy Leakage in Federated Learning From Representation Perspective. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 9311–9319
Sun Q, Ma L, Joon Oh S, Gool LV, Schiele B, Fritz M (2018a) Natural and Effective Obfuscation by Head Inpainting. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 5050–5059
Sun Q, Tewari A, Xu W, Fritz M, Theobalt C, Schiele B (2018b) A Hybrid Model for Identity Obfuscation by Face Replacement. In: Computer Vision – ECCV 2018, Springer International Publishing, pp 570–586
Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow IJ, Fergus R (2014) Intriguing properties of neural networks. In: 2nd International Conference on Learning Representations, ICLR 2014
Tajik K, Gunasekaran A, Dutta R, Ellis B, Bobba RB, Rosulek M, Wright CV, Feng Wc (2019) Balancing image privacy and usability with thumbnail-preserving encryption. In: 26th Annual Network and Distributed System Security Symposium, NDSS 2019, The Internet Society
Tanuwidjaja HC, Choi R, Kim K (2019) A Survey on Deep Learning Techniques for Privacy-Preserving. In: Machine Learning for Cyber Security, Springer, Cham, pp 29–46
Tanuwidjaja HC, Choi R, Baek S, Kim K (2020) Privacy-Preserving Deep Learning on Machine Learning as a Service—a Comprehensive Survey. IEEE Access 8:167425–167447
Tonge A (2018) Identifying Private Content for Online Image Sharing. In: Thirty-Second AAAI Conference on Artificial Intelligence
Tonge A, Caragea C (2019) Dynamic Deep Multi-modal Fusion for Image Privacy Prediction. In: The World Wide Web Conference, Association for Computing Machinery, pp 1829–1840
Tonge A, Caragea C (2020) Image Privacy Prediction Using Deep Neural Networks. ACM Transactions on the Web 14:7:1–7:32
Tonge A, Caragea C, Squicciarini AC (2018) Uncovering scene context for predicting privacy of online shared images. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, AAAI Press, pp 8167–8168
Tonge AK, Caragea C (2016) Image Privacy Prediction Using Deep Features. In: Thirtieth AAAI Conference on Artificial Intelligence
Torkzadehmahani R, Kairouz P, Paten B (2019) DP-CGAN: Differentially Private Synthetic Data and Label Generation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp 98–104
Tramèr F, Zhang F, Juels A, Reiter MK, Ristenpart T (2016) Stealing Machine Learning Models via Prediction APIs. In: 25th USENIX Security Symposium (USENIX Security 16), pp 601–618
Tran L, Kong D, Jin H, Liu J (2016) Privacy-cnh: A framework to detect photo privacy with convolutional neural network using hierarchical features. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, AAAI Press, pp 1317–1323
Triastcyn A, Faltings B (2020) Federated Generative Privacy. IEEE Intelligent Systems 35:50–57
Truex S, Liu L, Gursoy ME, Yu L, Wei W (2019) Demystifying Membership Inference Attacks in Machine Learning as a Service. IEEE Transactions on Services Computing pp 1–1
Uittenbogaard R, Sebastian C, Vijverberg J, Boom B, Gavrila DM, de With PH (2019) Privacy Protection in Street-View Panoramas Using Depth and Multi-View Imagery. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 10573–10582
von Marcard T, Henschel R, Black MJ, Rosenhahn B, Pons-Moll G (2018) Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In: Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part X, Springer, vol 11214, pp 614–631
Wah C, Branson S, Welinder P, Perona P, Belongie SJ (2011) The caltech-ucsd birds-200-2011 dataset
Wang B, Gong NZ (2018) Stealing Hyperparameters in Machine Learning. In: 2018 IEEE Symposium on Security and Privacy (SP), pp 36–52
Wang HP, Orekondy T, Fritz M (2021a) InfoScrub: Towards Attribute Privacy by Targeted Obfuscation. arXiv:200510329 [cs]
Wang J, Bao W, Sun L, Zhu X, Cao B, Yu PS (2019) Private Model Compression via Knowledge Distillation. Proceedings of the AAAI Conference on Artificial Intelligence 33:1190–1197

Wang T, Zhang Y, Jia R (2021b) Improving Robustness to Model Inversion Attacks via Mutual Information Regularization. Proceedings of the AAAI Conference on Artificial Intelligence 35:11666–11673

Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM (2017) ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 3462–3471

Wang Y, Wang C, Wang Z, Zhou S, Liu H, Bi J, Ding C, Rajasekaran S (2021c) Against Membership Inference Attack: Pruning is All You Need. In: International Joint Conference on Artificial Intelligence, IJCAI

Wang Z, She Q, Ward TE (2021d) Generative Adversarial Networks in Computer Vision: A Survey and Taxonomy. ACM Computing Surveys 54:37:1–37:38

Wei W, Liu L, Loper M, Chow KH, Gursoy ME, Truex S, Wu Y (2020) A Framework for Evaluating Client Privacy Leakages in Federated Learning. In: Computer Security – ESORICS 2020, Springer International Publishing, pp 545–566

Weinzaepfel P, Martin X, Schmid C (2017) Human Action Localization with Sparse Spatial Supervision. arXiv:160505197 [cs]

Wen Y, Song L, Liu B, Ding M, Xie R (2021) IdentityDP: Differential Private Identification Protection for Face Images. arXiv:210301745 [cs]

Wu Z, Wang Z, Wang Z, Jin H (2018) Towards Privacy-Preserving Visual Recognition via Adversarial Training: A Pilot Study. In: Computer Vision – ECCV 2018, Springer International Publishing, pp 627–645

Wu Z, Wang H, Wang Z, Jin H, Wang Z (2020) Privacy-Preserving Deep Action Recognition: An Adversarial Learning Framework and A New Dataset. IEEE Transactions on Pattern Analysis and Machine Intelligence pp 1–1

Xiao H, Rasul K, Vollgraf R (2017) Fashion-MNIST: A Novel Image Dataset for Benchmarking Machine Learning Algorithms. arXiv:170807747 [cs, stat]

Xiao Y, Wang C, Gao X (2020) Evade Deep Image Retrieval by Stashing Private Images in the Hash Space. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 9648–9657

Xie L, Lin K, Wang S, Wang F, Zhou J (2018) Differentially Private Generative Adversarial Network. arXiv:180206739 [cs, stat]

Xu H, Ma Y, Liu HC, Deb D, Liu H, Tang JL, Jain AK (2020) Adversarial Attacks and Defenses in Images, Graphs and Text: A Review. International Journal of Automation and Computing 17:151–178

Xue H, Liu B, Din M, Song L, Zhu T (2020) Hiding Private Information in Images From AI. In: ICC 2020 - 2020 IEEE International Conference on Communications (ICC), IEEE, pp 1–6

Yang Q, Liu Y, Chen T, Tong Y (2019a) Federated Machine Learning: Concept and Applications. ACM Transactions on Intelligent Systems and Technology 10:12:1–12:19

Yang Z, Zhang J, Chang EC, Liang Z (2019b) Neural Network Inversion in Adversarial Setting via Background Knowledge Alignment. In: Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, Association for Computing Machinery, pp 225–240

Yang Z, Shao B, Xuan B, Chang EC, Zhang F (2020) Defending Model Inversion and Membership Inference Attacks via Prediction Purification. arXiv:200503915 [cs]

Yao ACC (1986) How to generate and exchange secrets. In: 27th Annual Symposium on Foundations of Computer Science (Sfcs 1986), pp 162–167

Yeom S, Giacomelli I, Fredrikson M, Jha S (2018) Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting. In: 2018 IEEE 31st Computer Security Foundations Symposium (CSF), pp 268–282

Yi D, Lei Z, Liao S, Li SZ (2014) Learning Face Representation from Scratch. arXiv:14117923 [cs]

Yinka-Banjo C, Ugot OA (2020) A review of generative adversarial networks and its application in cybersecurity. Artificial Intelligence Review 53:1721–1736

Yu D, Zhang H, Chen W, Yin J, Liu TY (2021a) How Does Data Augmentation Affect Privacy in Machine Learning? Proceedings of the AAAI Conference on Artificial Intelligence 35:10746–10753

Yu F, Chen H, Wang X, Xian W, Chen Y, Liu F, Madhavan V, Darrell T (2020) BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 2633–2642

Yu J, Zhang B, Kuang Z, Lin D, Fan J (2017) iPrivacy: Image Privacy Protection by Identifying Sensitive Objects via Deep Multi-Task Learning. IEEE Transactions on Information Forensics and Security 12:1005–1016

Yu J, Kuang Z, Zhang B, Zhang W, Lin D, Fan J (2018) Leveraging Content Sensitiveness and User Trustworthiness to Recommend Fine-Grained Privacy Settings for Social Image Sharing. IEEE Transactions on Information Forensics and Security 13:1317–1332

Yu J, Xue H, Liu B, Wang Y, Zhu S, Ding M (2021b) GAN-Based Differential Private Image Privacy Protection Framework for the Internet of Multimedia Things. Sensors 21:58

Yu L, Liu L, Pu C, Gursoy ME, Truex S (2019) Differentially Private Model Publishing for Deep Learning. In: 2019 IEEE Symposium on Security and Privacy (SP), pp 332–349

Yuan X, He P, Zhu Q, Li X (2019) Adversarial Examples: Attacks and Defenses for Deep Learning. IEEE Transactions on Neural Networks and Learning Systems 30:2805–2824

Yun K, Honorio J, Chattopadhyay D, Berg TL, Samaras D (2012) Two-person interaction detection using body-pose features and multiple instance learning. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp 28–35

Zerr S, Siersdorfer S, Hare J, Demidova E (2012) Privacy-aware image classification and search. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, Association for Computing Machinery, pp 35–44

Zhang, Zhifei Y Song, Qi H (2017) Age Progression/Regression by conditional adversarial autoencoder. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE

Zhang C, Li S, Xia J, Wang W, Yan F, Liu Y (2020a) BatchCrypt: Efficient Homomorphic Encryption for Cross-Silo Federated Learning. In: 2020 USENIX Annual Technical Conference (USENIX ATC 20), pp 493–506

Zhang C, Xie Y, Bai H, Yu B, Li W, Gao Y (2021a) A survey on federated learning. Knowledge-Based Systems 216:106775

Zhang J, Li C (2020) Adversarial Examples: Opportunities and Challenges. IEEE Transactions on Neural Networks and Learning Systems 31:2578–2593

Zhang N, Paluri M, Taigman Y, Fergus R, Bourdev L (2015) Beyond frontal faces: Improving Person Recognition using multiple cues. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 4804–4813

Zhang X, Chen C, Xie Y, Chen X, Zhang J, Xiang Y (2021b) Privacy Inference Attacks and Defenses in Cloud-based Deep Neural Network: A Survey. arXiv:210506300 [cs]

Zhang Y, Jia R, Pei H, Wang W, Li B, Song D (2020b) The Secret Revealer: Generative Model-Inversion Attacks Against Deep Neural Networks. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 250–258

Zhao B, Mopuri KR, Bilen H (2020a) iDLG: Improved Deep Leakage from Gradients. arXiv:200102610 [cs, stat]

Zhao L, Wang Q, Zou Q, Zhang Y, Chen Y (2020b) Privacy-Preserving Collaborative Deep Learning With Unreliable Participants. IEEE Transactions on Information Forensics and Security 15:1486–1500

Zheng L, Shen L, Tian L, Wang S, Wang J, Tian Q (2015) Scalable Person Re-identification: A Benchmark. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp 1116–1124

Zhong H, Squicciarini AC, Miller DJ, Caragea C (2017) A group-based personalized model for image privacy classification and labeling. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, ijcai.org, pp 3952–3958

Zhu L, Liu Z, Han S (2019) Deep Leakage from Gradients. Advances in Neural Information Processing Systems 32

Zhu T, Ye D, Wang W, Zhou W, Yu P (2020a) More Than Privacy: Applying Differential Privacy in Key Areas of Artificial Intelligence. IEEE Transactions on Knowledge and Data Engineering pp 1–1

Zhu Y, Yu X, Chandraker M, Wang YX (2020b) Private-kNN: Practical Differential Privacy for Computer Vision. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, pp 11851–11859

Zou Y, Zhang Z, Backes M, Zhang Y (2020) Privacy Analysis of Deep Learning in the Wild: Membership Inference Attacks against Transfer Learning. arXiv:200904872 [cs, stat]