

IMAGE TAMPERING DETECTION

TECHNICAL FIELD

This specification relates to the field of computer technologies, and in particular, to an image tampering detection method and apparatus, a storage medium, and an electronic device.

BACKGROUND

In a consumer finance scenario, financial consumers and financial institutions involve a large number of financial vouchers with user identity information, such as salary certificates, occupation certificates, and identity documents. Usually, such materials are mostly uploaded by a user to a financial institution platform in a form of an image. After the financial institution platform completes material verification and an information review, a related transaction can be processed.

SUMMARY

This specification provides an image tampering detection method and apparatus, a storage medium, and an electronic device. The technical solutions are as follows.

According to a first aspect, this specification provides an image tampering detection method, applied to an electronic device. The method includes: determining a first model gradient matrix corresponding to a local image tampering detection model, and performing partial gradient differential privacy calculation processing based on the first model gradient matrix, to obtain a second model gradient matrix; uploading the second model gradient matrix to a service platform, where the second model gradient matrix is used for instructing the service platform to determine a target model parameter based on the second model gradient matrix uploaded by at least one electronic device; receiving the target model parameter sent by the service platform, and updating the local image tampering detection model based on the target model parameter, to obtain a target image tampering detection model; and performing image detection processing on a target privacy image based on the target image tampering detection model, to obtain a target heat detection map corresponding to the target privacy image, and performing tampering recognition processing on the target heat detection map, to obtain a tampering recognition result corresponding to the target privacy image.

According to a second aspect, this specification provides an image tampering detection method, applied to a service platform. The method includes: receiving a second model gradient

matrix uploaded by at least one electronic device, where the second model gradient matrix is obtained through partial gradient differential privacy calculation processing by the at least one electronic device based on a first model gradient matrix corresponding to a local image tampering detection model; determining a target model parameter based on each second model gradient matrix; and sending the target model parameter to each electronic device, where the target model parameter is used for instructing the electronic device to update the local image tampering detection model based on the target model parameter, to obtain a target image tampering detection model, performing image detection processing on a target privacy image based on the target image tampering detection model, to obtain a target heat detection map corresponding to the target privacy image, and performing tampering recognition processing on the target heat detection map, to obtain a tampering recognition result corresponding to the target privacy image.

According to a third aspect, this specification provides an image tampering detection apparatus, used in an electronic device. The apparatus includes: a gradient processing module, configured to determine a first model gradient matrix corresponding to a local image tampering detection model, and perform partial gradient differential privacy calculation processing based on the first model gradient matrix, to obtain a second model gradient matrix; a gradient sending module, configured to upload the second model gradient matrix to a service platform, where the second model gradient matrix is used for instructing the service platform to determine a target model parameter based on the second model gradient matrix uploaded by at least one electronic device; a data receiving module, configured to receive the target model parameter sent by the service platform, and update the local image tampering detection model based on the target model parameter, to obtain a target image tampering detection model; and an image processing module, configured to perform image detection processing on a target privacy image based on the target image tampering detection model, to obtain a target heat detection map corresponding to the target privacy image, and perform tampering recognition processing on the target heat detection map, to obtain a tampering recognition result corresponding to the target privacy image.

According to a fourth aspect, this specification provides an image tampering detection apparatus, used in a service platform. The apparatus includes: a data receiving module, configured to receive a second model gradient matrix uploaded by at least one electronic device, where the second model gradient matrix is obtained through partial gradient differential privacy calculation processing by the at least one electronic device based on a first model gradient matrix corresponding to a local image tampering detection model; a data processing module,

configured to determine a target model parameter based on each second model gradient matrix; and a data sending module, configured to send the target model parameter to each electronic device, where the target model parameter is used for instructing the electronic device to update the local image tampering detection model based on the target model parameter, to obtain a target image tampering detection model, perform image detection processing on a target privacy image based on the target image tampering detection model, to obtain a target heat detection map corresponding to the target privacy image, and perform tampering recognition processing on the target heat detection map, to obtain a tampering recognition result corresponding to the target privacy image.

According to a fifth aspect, this specification provides a computer storage medium. The computer storage medium has a plurality of instructions stored therein, where the instructions are adapted to be loaded by a processor and perform the foregoing method steps.

According to a sixth aspect, this specification provides a computer program product. The computer program product has at least one instruction stored therein, where the at least one instruction is executed by a processor and perform the foregoing method steps.

According to a seventh aspect, this specification provides an electronic device, which may include: a memory and a processor, where the memory has a computer program stored therein, and the computer program is adapted to be loaded by the processor and perform the foregoing method steps.

The technical solutions provided in this specification include at least the following beneficial effects. In embodiments of this specification, the electronic device determines a first model gradient matrix corresponding to a local image tampering detection model, and performs partial gradient differential privacy calculation processing based on the first model gradient matrix, to obtain a second model gradient matrix. The electronic device uploads the second model gradient matrix to a service platform, where the second model gradient matrix is used by the service platform to determine a target model parameter based on a second model gradient matrix uploaded by at least one electronic device. The electronic device receives the target model parameter sent by the service platform. The electronic device updates the local image tampering detection model based on the target model parameter, to obtain a target image tampering detection model. The electronic device performs image detection processing on a target privacy image based on the target image tampering detection model, to obtain a target heat detection map corresponding to the target privacy image, and performs tampering recognition processing on the target heat detection map, to obtain a tampering recognition result corresponding to the target privacy image. By using the foregoing method steps, the electronic

device performs preprocessing on gradient data that is to be uploaded to the service platform, to improve a reverse attack resistance capability of uploaded gradient data. The target image tampering detection model used by the electronic device to perform image detection processing on a privacy image is a model obtained based on model parameters obtained through processing by the service platform with reference to gradient data of a plurality of electronic devices. The model is trained with reference to data of a plurality of parties. Therefore, recognition accuracy of a privacy image by the model is improved. In this way, tampering recognition processing is further performed on a recognition result of the model, thereby improving accuracy of a tampering recognition result of a privacy image.

BRIEF DESCRIPTION OF THE DRAWINGS

To describe the technical solutions in embodiments of the present invention or in a related technology more clearly, the accompanying drawings required for describing the embodiments or related technology are briefly described below. Apparently, the accompanying drawings described below show merely some embodiments of the present invention, and a person of skill in the art may still derive other drawings from these accompanying drawings without creative efforts.

FIG. 1 is a schematic diagram of a system architecture of an image tampering detection method according to an embodiment of this specification.

FIG. 2 is a schematic flowchart of an image tampering detection method according to an embodiment of this specification.

FIG. 3 is a schematic flowchart of another image tampering detection method according to an embodiment of this specification.

FIG. 4 is a schematic diagram of a target heat detection map according to an embodiment of this specification.

FIG. 5 is a schematic flowchart of an image tampering detection method according to an embodiment of this specification.

FIG. 6 is a schematic flowchart of another image tampering detection method according to an embodiment of this specification.

FIG. 7 is a schematic structural diagram of an image tampering detection apparatus according to an embodiment of this specification.

FIG. 8 is a schematic structural diagram of an image tampering detection apparatus according to an embodiment of this specification.

FIG. 9 is a schematic structural diagram of an electronic device according to an embodiment of this specification.

FIG. 10 is a schematic structural diagram of a service platform according to an embodiment of this specification.

DETAILED DESCRIPTION

To make the invention objectives, features, technical solutions, and advantages of embodiments of this specification clearer and more understandable, the technical solutions in the embodiments of this specification are to be clearly and completely described with reference to the accompanying drawings in the embodiments of this specification. Apparently, the described embodiments are merely some but not all of the embodiments of this specification. All other embodiments obtained by a person of skill in the art based on the embodiments of this specification without creative efforts shall fall within the protection scope of this specification.

In the descriptions of this specification, it should be understood that, the terms "first", "second", and the like are only used for the purpose of description and are not to be construed as indicating or implying relative importance. In the descriptions of this specification, it should be noted that, unless otherwise expressly stated and limited, the terms "include", "have", and any variations thereof are intended to cover a non-exclusive inclusion unless. For example, a process, a method, a system, a product, or a device that includes a series of steps or units is not limited to the listed steps or units, and instead, further alternatively includes a step or unit that is not listed, or further alternatively includes another step or unit that is intrinsic to the process, method, product, or device. For a person of ordinary skill in the art, the specific meanings of the foregoing terms in this specification can be understood according to specific conditions. Moreover, in the descriptions of this specification, unless otherwise described, "a plurality of" means two or more. The term "and/or" describes an association relationship of associated objects and represents that three relationships may exist. For example, A and/or B may represent the following three cases: Only A exists, both A and B exist, and only B exists. The character "/" generally represents an "or" relationship between associated objects.

In a consumer finance scenario, a financial institution needs to review materials uploaded by a user, such as a financial voucher with user identity information, and handle related transactions upon approval. Tampering detection performed by a financial institution on a material uploaded by a user is a key to financial voucher review. In a related technology, a

financial institution relies on a tampering detection model obtained through training by the financial institution to perform tampering detection on a material uploaded by a user. A single financial institution has limited data, and the data possessed by the financial institution is not allowed to flow. A single financial institution has a small amount of usable training data for training a tampering detection model. Consequently, the tampering detection model of the financial institution has a problem of a poor tampering recognition effect.

To resolve the above technical problems, this specification is described in detail below in combination with specific embodiments.

FIG. 1 is a schematic diagram of a system architecture of an image tampering detection method according to an embodiment of this specification.

As shown in FIG. 1, FIG. 1 includes a service platform and an electronic device. A quantity of electronic devices and a quantity of nodes in the service platform that are shown in FIG. 1 are merely exemplary, and the quantities are not limited in the embodiments of this specification.

In this embodiment of this specification, the service platform may be a server including one node or a server cluster including a plurality of nodes.

When the service platform is a server cluster including a plurality of nodes, each node in the service platform may be an independent server device, for example, a rack, server device, tower, or cabinet server device, or a hardware device having a relatively strong computing capability, such as a workstation or a mainframe. Alternatively, the service platform may be a server cluster formed by a plurality of servers. Servers in the service cluster may be formed in a symmetrical manner, where each server has an equivalent function and an equivalent position in a transaction link, and each server may provide a service separately to the outside. That the server separately provides a service to the outside may be understood as that assistance of an additional server is not required.

In this embodiment of this specification, the electronic device may be a computer device having functions such as performing image detection processing and tampering recognition processing on a privacy image. Different electronic devices are devices of different consumer finance institutions.

It should be noted that the electronic device establishes a communication connection with at least one node in the service platform through a network for interactive communication. The network may be a wireless network or a wired network. The wireless network includes, but is not limited to, a cellular network, a wireless local area network, an infrared network, or a Bluetooth network. The wired network includes, but is not limited to, an Ethernet, a universal

serial bus (USB), or a controller area network. In one or more embodiments of this specification, technologies and/or formats including hypertext markup language (HTML), extensible markup language (XML), and the like, are used to represent data (for example, a target compressed package) exchanged through a network. In addition, conventional encryption technologies such as a secure socket layer (SSL), transport layer security (TLS), a virtual private network (VPN), and Internet protocol security (IPsec) may be used for encrypting all or some links. In some other embodiments, customized and/or dedicated data communication technologies may alternatively be used for replacing or supplementing the foregoing data communication technologies.

In embodiments of this specification, an electronic device determines a first model gradient matrix corresponding to a local image tampering detection model, and performs partial gradient differential privacy calculation processing based on the first model gradient matrix, to obtain a second model gradient matrix. The electronic device uploads the second model gradient matrix to a service platform, where the second model gradient matrix is used by the service platform to determine a target model parameter based on a second model gradient matrix uploaded by at least one electronic device. The electronic device receives the target model parameter sent by the service platform. The electronic device updates the local image tampering detection model based on the target model parameter, to obtain a target image tampering detection model. The electronic device performs image detection processing on a target privacy image based on the target image tampering detection model, to obtain a target heat detection map corresponding to the target privacy image, and performs tampering recognition processing on the target heat detection map, to obtain a tampering recognition result corresponding to the target privacy image. By using the foregoing method steps, the electronic device performs preprocessing on gradient data that is to be uploaded to the service platform, to improve a reverse attack resistance capability of uploaded gradient data. The target image tampering detection model used by the electronic device to perform image detection processing on a privacy image is a model obtained based on model parameters obtained through processing by the service platform with reference to gradient data of a plurality of electronic devices. The model is trained with reference to data of a plurality of parties. Therefore, recognition accuracy of a privacy image by the model is improved. In this way, tampering recognition processing is further performed on a recognition result of the model, thereby improving accuracy of a tampering recognition result of a privacy image.

FIG. 2 is a schematic flowchart of an image tampering detection method according to an embodiment of this specification. The method in this embodiment of this specification is

performed by an electronic device. As shown in FIG. 2, the method in this embodiment of this specification may include the following steps.

S202: Determine a first model gradient matrix corresponding to a local image tampering detection model, and perform partial gradient differential privacy calculation processing based on the first model gradient matrix, to obtain a second model gradient matrix.

It is easy to understand that the local image tampering detection model is a model that has an image detection capability and that is obtained based on a machine learning model. The image detection capability of the local image tampering detection model may be understood as a capability of recognizing a tampered area that may exist in an image. The tampered area in the image is an area obtained by modifying an original image area in the image through one or more image operations. The image operations may include operations such as image compositing, copying and pasting, deleting, and decorating. Specifically, in this embodiment of this specification, the image recognized by the local image tampering detection model may be a privacy image such as an image of a personal identity document or an image of a credential with user identity information. The credential may be a credential such as a salary certificate, an occupation certificate, or a medical certificate.

The first model gradient matrix is a matrix formed by gradients of model parameters of the local image tampering detection model. The gradients in the first model gradient matrix may be obtained by training the local image tampering detection model using training data. The model parameters are variables used for describing data inside a model, and the variables are usually learned during model training, to minimize an error between model prediction and real data. In the electronic device, the model parameter and a model structure of a local image tampering detection model may be obtained from the service platform. Specifically, the model parameter may be a weight, a deviation, or another adjustable variable. The parameters are continuously adjusted by iterating on the training data, so that the model predicts new data more accurately. For example, in a neural network, the model parameter is a weight and a deviation of each neuron. In linear regression, the model parameter is a regression coefficient and an intercept. In a support vector machine, the model parameter is a decision boundary and a support vector.

The second model gradient matrix is used for instructing the service platform to determine the target model parameter based on the second model gradient matrix uploaded by at least one electronic device. The second model gradient matrix is a gradient matrix obtained by performing differential privacy calculation on a partial gradient in the first model gradient matrix.

In one or more embodiments of this specification, the step of determining a first model gradient matrix corresponding to a local image tampering detection model may be specifically: performing model training processing on the local image tampering detection model based on the sample privacy image, to obtain the first model gradient matrix. The sample privacy image may include a first privacy image that is acquired in a historical time period and that has been authorized by a user, and may further include a second privacy image obtained by performing data expansion processing on the first privacy image that has been authorized by the user. A quantity and types of first privacy images that are acquired by a single electronic device and that have been authorized by a user are limited. Therefore, data expansion processing such as splicing, deleting, and scaling may be performed on the first privacy image, to obtain the second privacy image, thereby expanding sample data. When the model training processing is performed on the local image tampering detection model by using the sample privacy image, the local image tampering detection model may be trained for a plurality of rounds through a loss function and a backpropagation algorithm, to obtain the first model gradient matrix of the model parameter of the local image tampering detection model.

The step of performing partial gradient differential privacy calculation processing based on the first model gradient matrix, to obtain a second model gradient matrix may be specifically: determining a first gradient subset in the first model gradient matrix, performing noise addition processing on all gradients in the first gradient subset to obtain a second gradient subset, and updating the first model gradient matrix based on the second gradient subset to obtain the second model gradient matrix.

In a process of determining the first gradient subset in the first model gradient matrix, a gradient of at least one specified position in the first model gradient matrix may be used as a gradient forming the first gradient subset. The specified position may be determined based on a position of a specified parameter in a network structure. The network structure is a structure of a network used by the local image tampering detection model, and the specified parameter is a network parameter in a network used by the local image tampering detection model. The specified parameter may be obtained by querying a network parameter mapping relationship. The network parameter mapping relationship may store correspondence between reference models and reference parameters. When the local image tampering detection model is used as a reference model, a reference parameter corresponding to the local image tampering detection model is the specified parameter. The correspondence between the reference models and the reference parameters stored in the network parameter mapping relationship may be obtained through configuration by technical personnel according to previous experience.

In a process of performing noise addition processing on all the gradients in the first gradient subset to obtain the second gradient subset, noise addition processing may be performed on all the gradients in the first gradient subset by using random noise such as Gaussian noise, to obtain the second gradient subset.

In a process of updating the first model gradient matrix based on the second gradient subset to obtain the second model gradient matrix, a corresponding position of each second gradient in the second gradient subset in the first model gradient matrix may be first determined, and a first gradient at each corresponding position in the first model gradient matrix is then replaced with a second gradient corresponding to the corresponding position, to obtain the second model gradient matrix.

S204: Upload the second model gradient matrix to a service platform.

It is easy to understand that the service platform is a server that forms a federated learning architecture with the electronic device. The federated learning architecture refers to a system architecture shown in FIG. 1. That is, the federated learning architecture includes one service platform and a plurality of electronic devices. In the federated learning architecture, the electronic device sends the second model gradient matrix to the service platform, the service platform aggregates the second model gradient matrices sent by the electronic devices, and then the service platform sends a model parameter obtained through the aggregation processing to the electronic devices. During communication between the service platform and each electronic device, image data used by the electronic device to calculate the second model gradient matrix does not flow out of the electronic device, that is, the image data used by each electronic device is only inside the electronic device. Therefore, the federated learning architecture is used to jointly train a model while ensuring that data does not exit a domain.

In one or more embodiments of this specification, performing step S204 may specifically include: encrypting the second model gradient matrix by using a preset encryption key, to obtain encrypted gradient data, and uploading the encrypted gradient data to the service platform. The preset encryption key may be an encryption key agreed on in advance between the electronic device and the service platform in a trusted environment. Alternatively, the preset encryption key may be a public key of the service platform. Because the electronic device uploads an encrypted second model gradient matrix, the service platform may decrypt the encrypted gradient data by using a preset decryption key corresponding to the preset encryption key, to obtain the second model gradient matrix. When the preset encryption key is the encryption key agreed on in advance between the electronic device and the service platform in the trusted environment, the preset decryption key used by the service platform is the

decryption key corresponding to the preset encryption key. When the preset encryption key is the public key of the service platform, the preset decryption key used by the service platform may be a private key of the service platform. In this way, the encrypted second model gradient matrix is sent to the service platform, so that a problem of insecurity due to leakage of the second model gradient matrix is avoided.

Optionally, based on that the encrypted second model gradient matrix is sent to the service platform, the encrypted second model gradient matrix may be further sent to the service platform when current time reaches preset sending time. The preset sending time may be sending time agreed on with the service platform. For example, the preset sending time is time such as 8 PM or 10 PM every day. In this way, the second model gradient matrix is sent to the service platform at specific time, to ensure joint training efficiency of a model.

S206: Receive a target model parameter sent by the service platform, and update the local image tampering detection model based on the target model parameter, to obtain a target image tampering detection model.

It is easy to understand that the target model parameter is a latest model parameter of the local image tampering detection model obtained through aggregation processing performed by the service platform based on the second model gradient matrix sent by each electronic device.

In one or more embodiments of this specification, the receiving the target model parameter sent by the service platform may be that the service platform actively sends the target model parameter to the electronic device, so that the electronic device can receive the target model parameter; or may be that the electronic device sends a parameter query request to the service platform, and the service platform sends the target model parameter to the electronic device based on the parameter query request, so that the electronic device can receive the target model parameter.

Specifically, that the electronic device receives the target model parameter sent by the service platform may include: receiving, by the electronic device, encrypted parameter data sent by the service platform, where the encrypted parameter data is encrypted data obtained by the service platform by encrypting the target model parameter, and decrypting, by the electronic device, the encrypted parameter data to obtain the target model parameter. When the service platform encrypts the target model parameter to obtain the encrypted parameter data, a used encryption key may be a preset encryption key agreed on in advance between the electronic device and the service platform in a trusted environment, or may be a public key of the electronic device. Correspondingly, the service platform encrypts the target model parameter, and the electronic device needs to decrypt the encrypted parameter data. When the encryption

key is the preset encryption key, the decryption key used by the electronic device is a preset decryption key corresponding to the preset encryption key. When the encryption key is the public key of the electronic device, the decryption key used by the electronic device is a private key of the electronic device. In this way, the data that is obtained by encrypting the target model parameter and that is sent by the service platform is received, so that a problem of insecurity due to leakage of the target model parameter is avoided.

The update the local image tampering detection model based on the target model parameter, to obtain a target image tampering detection model may be specifically: updating the model parameter of the local image tampering detection model to the target model parameter, to obtain the target image tampering detection model.

S208: Perform image detection processing on a target privacy image based on the target image tampering detection model, to obtain a target heat detection map corresponding to the target privacy image, and perform tampering recognition processing on the target heat detection map, to obtain a tampering recognition result corresponding to the target privacy image.

It is easy to understand that the target privacy image may be a to-be-detected privacy image authorized by the user. The target privacy image may include a privacy image such as an image of a personal identity document or an image of a credential with user identity information. The credential may be a credential such as a salary certificate, an occupation certificate, or a medical certificate.

The target heat detection map may be image data displaying a distribution status of a tampered area by using color encoding. Specifically, in the target heat detection map, different probabilities of tampered areas may be displayed by using different colors. The probability is a probability that a tampered area detected by the target image tampering detection model has been tampered with.

The tampering recognition result may include a result of whether the target privacy image is a tampered image.

In one or more embodiments of this specification, the target privacy image is input to the target image tampering detection model, and the target heat detection map is output. Further, tampering recognition processing may be performed on the target heat detection map by using a device-dedicated classifier, to obtain the tampering recognition result corresponding to the target privacy image. The device-dedicated classifier may be determined based on data features of privacy images possessed by different electronic devices, and device-dedicated classifiers in different electronic devices may not be completely the same. For example, different electronic devices may use classifiers having a same network structure, and different electronic devices

may configure different discrimination thresholds in the classifiers based on data features of privacy images of the different electronic devices, so that different electronic devices perform targeted discrimination on a public model recognition result. The tampering recognition result may be that the target privacy image is a tampered image, or the tampering recognition result may be that the target privacy image is not a tampered image. In this way, different device-dedicated classifiers are configured for different electronic devices, to accurately determine a tampering recognition result of a privacy image that needs to be detected by each electronic device.

Optionally, the tampering recognition result may include a result of whether the target privacy image is a tampered image, and may further include a tampered area existing in the target privacy image. In this case, the step of performing tampering recognition processing on the target heat detection map, to obtain the tampering recognition result corresponding to the target privacy image may be specifically: performing verification processing on a reference tampered area in the target heat detection map based on a preset tamper verification method, to obtain a real tampering result corresponding to each reference tampered area; and determining, based on the real tampering result, the tampering recognition result corresponding to the target privacy image.

The preset tampering verification method may be a tampering verification method determined by the electronic device based on a data feature of a privacy image possessed by the electronic device. Different electronic devices have different preset tampering verification methods.

The determining, based on the real tampering result, the tampering recognition result corresponding to the target privacy image may be understood as that, when at least one real tampering result is a first tampering result of a type of including tampering, a reference tampered area corresponding to the first tampering result is determined as a tampered area in the target privacy image, and a tampering recognition result including the tampered area is generated; or when all real tampering results are second tampering results of a type of including no tampering, a tampering recognition result indicating that the target privacy image is not a tampered image is generated. In this way, different electronic devices use respective preset tampering verification methods to verify a possibly-existing tampered area in a target heat detection map, so that a tampering recognition result of a privacy image that needs to be detected by each electronic device can be accurately determined.

In the embodiments of this specification, by using the foregoing method steps, the electronic device performs preprocessing on gradient data that is to be uploaded to the service

platform, to improve a reverse attack resistance capability of uploaded gradient data. The target image tampering detection model used by the electronic device to perform image detection processing on a privacy image is a model obtained based on model parameters obtained through processing by the service platform with reference to gradient data of a plurality of electronic devices. The model is trained with reference to data of a plurality of parties. Therefore, recognition accuracy of a privacy image by the model is improved. In this way, tampering recognition processing is further performed on a recognition result of the model, thereby improving accuracy of a tampering recognition result of a privacy image.

FIG. 3 is a schematic flowchart of an image tampering detection method according to an embodiment of this specification. The method in this embodiment of this specification is performed by an electronic device. As shown in FIG. 3, the method in this embodiment of this specification may include the following steps.

S302: Determine a first model gradient matrix corresponding to a local image tampering detection model.

In one or more embodiments of this specification, performing S302 may be specifically: obtaining a historical privacy image; and performing model training processing on the local image tampering detection model based on the historical privacy image, to obtain the first model gradient matrix.

The historical privacy image may be an image related to identity information or a credential with user identity information that is acquired with user authorization in a financial anti-fraud identification scenario. The identity information may refer to a personal identity document. The credential may be a credential such as a salary certificate, an occupation certificate, or a medical certificate.

The obtaining the historical privacy image may be: determining a current training transaction scenario, and obtaining a historical privacy image corresponding to the current training transaction scenario from a preset database. The privacy images stored in the preset database may be privacy images respectively corresponding to a plurality of training transaction scenarios. A privacy image corresponding to each training transaction scenario may be a privacy image uploaded by the user terminal in a historical time period. The historical time period may be a time period such as last 24 hours or last 7 days.

The current training transaction scenario is a transaction scenario of current joint training agreed on between each electronic device and the service platform. The scenario involved in this embodiment of this specification is a financial anti-fraud identification scenario. The financial anti-fraud identification scenario may specifically include a plurality of different

transaction scenarios, such as a poverty certificate identification scenario, a medical certificate identification scenario, and a loan certificate identification scenario. The current training transaction scenario may be any one of the financial anti-fraud identification scenarios. In a process of jointly training models using the federated learning architecture by each electronic device and the service platform, the model may be jointly trained in different scenarios. For example, if a number of joint training rounds reaches a preset number of rounds, model training for a transaction scenario can be completed. In this embodiment of this specification, the privacy image is obtained by using the current training transaction scenario, thereby ensuring consistency of data content used when each electronic device jointly trains the model, and facilitating improving identification accuracy of a model obtained through subsequent joint modeling.

The performing model training processing on the local image tampering detection model based on the historical privacy image, to obtain the first model gradient matrix may be understood as: performing data expansion processing on the historical privacy image to obtain a reference privacy image, marking the reference privacy image to obtain training data, and performing model training processing on the local image tampering detection model by using the training data, to obtain the first model gradient matrix.

Specifically, the performing data expansion processing on the historical privacy image to obtain the reference privacy image may be processing the historical privacy image by using a data expansion method, to obtain the reference privacy image. The data expansion method may include, but is not limited to, methods such as image copying and pasting, image cutting and splicing, image generation, image resampling, and scaling. The image copying and pasting means copying and pasting data at a position of a historical privacy image to another position of the historical privacy image. The image cutting and splicing means separately cutting some images from a plurality of historical privacy images and splicing the cut some images. The image generation means replacing a piece of data in a historical privacy image with similar data similar to the piece of data by using a digital image processing technology; or inputting a historical privacy image to a generative model to obtain a reference privacy image, where data of a position in the reference privacy image is similar data of initial data of the position in the historical privacy image. The image resampling and scaling means changing a resolution, a size, or a sampling rate of a historical privacy image, to adjust a definition, a ratio, or a visual effect of the historical privacy image. In this embodiment of this specification, a sample quantity of a model is increased by using the data expansion method, thereby improving a generalization capability of the model.

Specifically, the marking the reference privacy image to obtain the training data may include marking operations such as marking a tampered area in the reference privacy image, marking a tampering label of the reference privacy image, and marking a privacy area in the reference privacy image, to obtain the training data including marked information such as the reference privacy image, the tampered area corresponding to the reference privacy image, the tampering label corresponding to the reference privacy image, and the privacy area corresponding to the reference privacy image. The tampering label marking the reference privacy image may be represented by using a preset character. For example, the preset character is 1 or 0. When the tampering label is 1, it may indicate that there is a tampered area in the reference privacy image. When the tampering label is 0, it may indicate that there is no tampered area in the reference privacy image.

Specifically, the performing model training processing on the local image tampering detection model by using the training data, to obtain the first model gradient matrix may be understood as: inputting the training data as batch data to the local image tampering detection model to perform forward propagation calculation, to obtain an output of the local image tampering detection model, then calculating a gradient of a model parameter of the local image tampering detection model based on the output of the local image tampering detection model by using an error backpropagation algorithm, and forming a gradient matrix by using the gradients, to obtain the first model gradient matrix.

S304: Determine an initial partial gradient subset in the first model gradient matrix.

In one or more embodiments of this specification, performing step S304 may be specifically: A2: determining a model feature of the local image tampering detection model; A4: determining, based on the model feature, a candidate network parameter corresponding to the local image tampering detection model; and A6: determining the initial partial gradient subset corresponding to the candidate network parameter in the first model gradient matrix.

The model feature is a feature of a network structure used by the model. For example, the model feature may include a feature of a network layer in the network structure, such as a quantity of network layers and a function of the network layer.

The candidate network parameter is a network parameter selected, based on the model feature, from a network structure used by the local image tampering detection model.

The initial partial gradient subset is a gradient set formed by some gradients of the first model gradient matrix.

Specifically, performing step A2 may be: obtaining the model feature of the local image tampering detection model from a model configuration file. The model configuration file may

store model configuration information such as a model feature and a model parameter of a model.

Specifically, performing step A4 may be: a2: determining the candidate network layer in the local image tampering detection model based on the model feature; a4: determining a sample mask image corresponding to the sample privacy image, and inputting the sample mask image to the local image tampering detection model, to obtain a learnable network parameter in the candidate network layer; and a6: determining, based on the learnable network parameter, the candidate network parameter corresponding to the local image tampering detection model.

In a process of performing step a2, a network-layer mapping relationship is obtained, where the network-layer mapping relationship is used for storing correspondence between a reference model feature and a reference candidate network layer; and the network-layer mapping relationship is searched for a target candidate network layer corresponding to the model feature, where the target candidate network layer is the candidate network layer of the local image tampering detection model in this embodiment of this specification. The network-layer mapping relationship may be configured and generated by model network experts, or may be configured and generated by technical personnel according to previous experience. For example, if a network used by the local image tampering detection model is a CATNet, a reference candidate network layer corresponding to a model feature of the CATNet in the network-layer mapping relationship may be the first convolutional layer in an RGB branch, the last convolutional layer in the RGB branch, and the first convolutional layer in a DCT branch. When the network-layer mapping relationship is queried based on the model feature of the CATNet, it can be obtained that the candidate network layers are the first convolutional layer in the RGB branch, the last convolutional layer in the RGB branch, and the first convolutional layer in the DCT branch.

In a process of performing step a4, the determining the sample mask image corresponding to the sample privacy image may be understood as: selecting a preset quantity of sample privacy images from all privacy images of the electronic device, where the preset quantity may be set based on an actual requirement, using at least one privacy area in the sample privacy image as a mask area, and performing a masking operation on the mask area, to obtain the sample mask image including only the mask area. The inputting the sample mask image to the local image tampering detection model, to obtain the learnable network parameter in the candidate network layer may be understood as: inputting the sample mask image to the local image tampering detection model to obtain network parameters in the candidate network layer, and determining the learnable network parameter from the network parameters. The learnable

network parameter herein includes a name of the learnable network parameter and a parameter value of the learnable network parameter. The learnable network parameter is a parameter of a learning parameter value during training, that is, a parameter whose parameter value is a variable value. A parameter value of the learnable parameter is a variable value, which may be understood as that, the parameter value of the learnable parameter usually starts from being set to a group of random values, and the values are updated in an iterative manner as a network learns. For example, the learnable parameter may be a parameter of a convolution layer, including a weight of a convolution kernel, a bias of each channel, and the like.

In a process of performing step a6, the determining, based on the learnable network parameter, the candidate network parameter corresponding to the local image tampering detection model may be understood as: determining a parameter having a largest absolute value in the learnable network parameter as the candidate network parameter corresponding to the local image tampering detection model. This is because, the candidate network parameter has a strong response output for an input of the privacy area in the privacy image, and the candidate network parameter may be used for positioning a gradient corresponding to the privacy area.

Specifically, performing step A4 may be: b2: determining at least one gradient corresponding to each candidate network parameter in the first model gradient matrix; b4: performing filtering processing on the at least one gradient, to obtain a target gradient corresponding to the candidate network parameter; and b6: determining, based on the target gradient, the initial partial gradient subset corresponding to the candidate network parameter.

In a process of performing step b2, after the candidate network parameter is obtained, it is easy to determine position information of the candidate network parameter in the network used by the local image tampering detection model. Because gradients at different positions in the first model gradient matrix are in correspondence with parameters at different positions in the network, at least one gradient corresponding to each candidate network parameter may be selected from the first model gradient matrix based on the position information of the candidate network parameter in the network.

In a process of performing step b4, for the at least one gradient corresponding to each candidate network parameter, a gradient threshold corresponding to the candidate network parameter may be obtained from a gradient filtering mapping table. Filtering processing is performed on the at least one gradient by using the gradient threshold, that is, a gradient less than the gradient threshold in the at least one gradient is filtered, to obtain the target gradient corresponding to the candidate network parameter. Gradient thresholds corresponding to different candidate network parameters are preconfigured in the gradient filtering mapping

table. A function of the gradient threshold is to filter out a relatively small gradient in the gradients corresponding to the candidate network parameters.

In a process of performing step b6, each candidate network parameter corresponds to a target gradient, and target gradients corresponding to all the candidate network parameters are stored in a gradient set, to obtain the initial partial gradient subset.

S306: Perform noise addition processing on the initial partial gradient subset, to obtain a target partial gradient subset, and update the first model gradient matrix based on the target partial gradient subset, to obtain the second model gradient matrix.

In one or more embodiments of this specification, the performing noise addition processing on the initial partial gradient subset, to obtain the target partial gradient subset may be specifically: B2: performing gradient clipping processing on the initial partial gradient subset to obtain a reference partial gradient subset; and B4: performing noise addition processing on the reference partial gradient subset, to obtain the target partial gradient subset.

In a process of step B2, gradient clipping may be performed on the initial partial gradient subset by using a gradient boundary, to obtain the reference partial gradient subset. A gradient clipping formula may be:

$$(G)' = \frac{G}{\max\left(1, \frac{\|G\|_F^2}{C}\right)}$$

where C represents a gradient boundary, G represents an initial partial gradient subset, $(G)'$ represents a reference partial gradient subset, and $\|G\|_F^2$ represents a 2-norm. A value of C may be set based on an actual application. None of gradients calculated by using the foregoing gradient clipping formula is greater than C, that is, none of gradients in the reference partial gradient subset is greater than C. In this embodiment of this specification, clipping processing is performed on the gradient subset, thereby ensuring stability of the model during training.

In a process of performing step B4, noise addition processing is performed on each gradient in the reference partial gradient subset, and all gradients subjected to noise addition form the target partial gradient subset. For example, a noise addition formula may be:

$$(g)' = g + N\left(0, \frac{\sigma^2 C^2}{m}\right)$$

where m is a quantity of selected sample privacy images in step a4, σ^2 is a variance of Gaussian noise, C is the gradient boundary, g represents a reference partial gradient subset, and

$(g)'$ represents a target partial gradient subset. After setting a privacy budget ε and a relaxation item δ , it can be learned that when $\sigma = \frac{\sqrt[2]{2\log \frac{1.25}{\delta}}}{\varepsilon}$, a partial gradient subset G

positioned by using a privacy area reaches a differential privacy protection level of (ε, δ) . The privacy budget ε is a hyperparameter for controlling a differential privacy protection effect. A smaller value of the privacy budget indicates a better privacy protection effect, but louder noise is introduced.

The updating the first model gradient matrix based on the target partial gradient subset, to obtain the second model gradient matrix may be specifically: determining corresponding positions of gradients in the target partial gradient subset in the first model gradient matrix, and replacing the gradients at the corresponding positions in the first model gradient matrix with gradients corresponding to the corresponding positions in the target partial gradient subset, to obtain the second model gradient matrix. For example, one gradient in the target partial gradient subset is g_1 , a corresponding position of g_1 in the first model gradient matrix is at the 3rd row, the 4th column, a gradient at the 3rd row, the 4th column in the first model gradient matrix is G_1 , and G_1 in the first model gradient matrix is replaced with g_1 . Deduced by analogy, a gradient at a corresponding position in the first model gradient matrix is replaced with a corresponding gradient in the target partial gradient subset, to obtain the second model gradient matrix.

S308: Upload the second model gradient matrix to a service platform.

S310: Receive a target model parameter sent by the service platform, and update the local image tampering detection model based on the target model parameter, to obtain a target image tampering detection model.

Specifically, for implementations of S308 and S310, refer to explanations and descriptions of S204 and S206 in the embodiment shown in FIG. 2. Details are not described herein again.

S312: Perform image detection processing on a target privacy image based on the target image tampering detection model, to obtain a target heat detection map corresponding to the target privacy image.

For details of the target privacy image and the target heat detection map, refer to the explanations and descriptions of S208 in the embodiment shown in FIG. 2. Details are not described herein again.

In one or more embodiments of this specification, the target privacy image is input to the target image tampering detection model, to obtain the target heat detection map output by the

target image tampering detection model. For example, refer to a schematic diagram of a target heat detection map shown in FIG. 4. In FIG. 4, marks "a red area", "an orange area", and "a yellow area" are tampered areas detected by the target image tampering detection model that may be tampered with. Tampered areas distinguished by different colors represent different tampering probabilities. For example, a tampering probability of the "red area" is greater than a tampering probability of the "orange area", and a tampering probability of the "orange area" is greater than a tampering probability of the "yellow area".

S314: Obtain a device-dedicated classifier, and perform tampering recognition processing on the target heat detection map based on the device-dedicated classifier, to obtain a tampering recognition result corresponding to the target privacy image.

It is easy to understand that the device-dedicated classifier is a dedicated classifier set for a data feature of a privacy image possessed by an electronic device. Device-dedicated classifiers of different electronic devices are not all the same.

Because data features of privacy images possessed by different electronic devices are not all the same, and preferences of different electronic devices for a recall rate and accuracy of a local image tampering detection model are different, device-dedicated classifiers with different discrimination thresholds may be set for different electronic devices, to adapt to tampering recognition requirements of different electronic devices for privacy images. For example, the target heat detection map indicates that a character may be tampered with. However, for an electronic device, it is considered that tampering occurs only when consecutive characters are tampered with, and it is not considered that tampering occurs when a single character is tampered with. Therefore, different electronic devices are configured with classifiers adapting to recognition requirements of the electronic devices to determine a final tampering recognition result.

In one or more embodiments of this specification, the device-dedicated classifier may be obtained through selection from a model library locally configured in the electronic device. Further, the target heat detection map is input to the device-dedicated classifier, to obtain the output tampering recognition result. The tampering recognition result may be a binary classification result, that is, the tampering recognition result may include two results: a result of including tampering and a result of including no tampering. The result of including tampering represents that the target privacy image is tampered with, and the result of including no tampering may represent that the target privacy image is not been tampered with.

Optionally, the device-dedicated classifier in this embodiment of this specification may include a three-layer fully-connected network, and the classifier obtains a final classification

output by using a sigmoid activation function. The electronic device trains the classifier by using a binary classification cross-entropy loss and the error backpropagation algorithm. A distribution interval of a model output value using the sigmoid activation function is 0 to 1, and the model output value may be used for representing a probability of image being tampered with. Different electronic devices may select a proper discrimination threshold based on progress of a transaction and importance of data verification, to obtain the tampering recognition result.

For example, when the determining threshold of the device-dedicated classifier is set to 0.7, a target privacy image whose model output value is equal to or greater than 0.7 is recognized as an image that is tampered with. The schematic diagram of the target heat detection map shown in FIG. 4 is still used. Assuming that model output values of a device classifier in this embodiment of this specification for the "red area", the "orange area", and the "yellow area" are respectively 0.81, 0.67, and 0.32, for the device classifier, a tampering recognition result of the "red area" is including tampering, a tampering recognition result of the "orange area" is including no tampering, and a tampering recognition result of the "yellow area" is including no tampering.

In this embodiment of this specification, a larger set discrimination threshold of the device-dedicated classifier indicates higher accuracy of recognition of the device-dedicated classifier and a lower recall rate.

In this embodiment of this specification, locally at the electronic device, the electronic device first determines a first model gradient matrix of the local image tampering detection model, and then performs noise addition processing on a gradient perceptual and obvious to a privacy area in the first model gradient matrix, to obtain the second model gradient matrix. In this way, noise perturbation is added to the gradient matrix uploaded to the service platform for joint training, to defend against an external attacker implementing a reverse attack by stealing an uploaded gradient, thereby improving security of privacy data. In addition, the model is updated by using the target model parameter sent by the service platform, to obtain a latest target image tampering detection model, and second tampering recognition processing is further performed, by using the device-dedicated classifier for the electronic device, on a detection result obtained by the target image tampering detection model from the target privacy image. While tampering detection requirements of different electronic devices are adapted, accuracy of tampering recognition results of the electronic devices for a privacy image is improved.

FIG. 5 is a schematic flowchart of an image tampering detection method according to an

embodiment of this specification. The method in this embodiment of this specification is performed by a service platform. As shown in FIG. 5, the method in this embodiment of this specification may include the following steps.

S502. Receive a second model gradient matrix uploaded by at least one electronic device.

The second model gradient matrix is obtained through partial gradient differential privacy calculation processing by the at least one electronic device based on a first model gradient matrix corresponding to a local image tampering detection model. For details of the step in which the electronic device calculates the second model gradient matrix based on the first model gradient matrix, refer to the descriptions of S202 in the embodiment shown in FIG. 2. Details are not described herein again.

In some embodiments, the step of performing S502 may specifically include: receiving encrypted gradient data uploaded by the at least one electronic device, and decrypting the encrypted gradient data by using a preset decryption key, to obtain a second model gradient matrix corresponding to each electronic device. Preset decryption keys used for decrypting the encrypted gradient data sent by different electronic devices may be different. The preset decryption key may be a decryption key agreed on between the electronic device and the service platform in a trusted environment in advance. The preset decryption key may alternatively be a public key of the electronic device. The service platform decrypts the encrypted gradient data. Correspondingly, the electronic device needs to encrypt the second model gradient matrix. When the preset decryption key is the decryption key agreed on between the electronic device and the service platform in the trusted environment in advance, the preset encryption key used by the electronic platform is an encryption key corresponding to the preset decryption key. When the preset decryption key is the public key of the electronic device, the encryption key used by the electronic device may be a private key of the electronic device. In this way, the encrypted second model gradient matrix is sent to the service platform, so that a problem of insecurity due to leakage of the second model gradient matrix is avoided.

Optionally, before performing S502, the service platform may share a local image tampering detection model with each electronic device. Specifically, the service platform may send a model architecture and an initial model parameter of the local image tampering detection model to the electronic device. The local image tampering detection model may be a model obtained by the service platform by performing model training based on an open-source privacy data training set.

S504: Determine a target model parameter based on each second model gradient matrix.

It is easy to understand that the target model parameter is a latest model parameter of the

local image tampering detection model obtained through aggregation processing performed by the service platform based on the second model gradient matrix sent by each electronic device.

In one or more embodiments of this specification, the service platform may perform gradient aggregation processing on each second model gradient matrix of a same batch to obtain an aggregation gradient, and update a model parameter of the local image tampering detection model based on the aggregation gradient to obtain the target model parameter. The each second model gradient matrix of the same batch may refer to each second model gradient matrix received within a same time period. The service platform and the electronic devices may agree in advance on a time period of uploading the second model gradient matrix in each round. In each round of calculating the target model parameter, calculation may be performed by using the second model gradient matrix within the time period, to obtain the target model parameter.

Optionally, an algorithm for obtaining the aggregation gradient by the service platform may be a federated average (FedAvg) algorithm.

S506: Send the target model parameter to each electronic device.

The target model parameter is used for instructing the electronic device to update the local image tampering detection model based on the target model parameter, to obtain a target image tampering detection model. Image detection processing is performed on a target privacy image based on the target image tampering detection model, to obtain a target heat detection map corresponding to the target privacy image. Tampering recognition processing is performed on the target heat detection map, to obtain a tampering recognition result corresponding to the target privacy image. For a process in which each electronic device updates the local image tampering detection model based on the target model parameter to obtain the target image tampering detection model, and performs subsequent processing based on the target image tampering detection model, refer to the descriptions of S206 and S208 in the embodiment shown in FIG. 2. Details are not described herein again.

In some embodiments, performing step S506 may specifically include: encrypting the target model parameter by using a preset encryption key to obtain encrypted parameter data, and sending the encrypted parameter data to each electronic device. Encrypted parameter data may be sent to different electronic devices by using different encryption keys. The preset encryption key used by the service platform may be an encryption key agreed on in advance between the electronic device and the service platform in a trusted environment. The preset encryption key may alternatively be a public key of the electronic device. The service platform encrypts the target model parameter, and correspondingly, the electronic device needs to decrypt the encrypted parameter data to obtain the target model parameter. When the preset

encryption key used by the service platform is the encryption key agreed on in advance between the electronic device and the service platform in the trusted environment, the decryption key used by the electronic device may be a preset decryption key corresponding to the preset encryption key. When the preset encryption key used by the service platform is the public key of the electronic device, a decryption key used by the electronic device for decryption may be a private key of the electronic device. In this way, the service platform sends the encrypted target model parameter to the electronic device, thereby avoiding a problem of insecurity due to leakage of the target model parameter.

It should be noted that the service platform may calculate target model parameters for a plurality of rounds by using second model gradient matrices of a plurality of batches, to continuously update the local image tampering detection model shared with each electronic device until the local image tampering detection model converges. Each time the service platform updates the local image tampering detection model, the service platform may send the latest model parameter, that is, the target model parameter, of the local image tampering detection model to each electronic device.

In this embodiment of this specification, the service platform receives the second model gradient matrix uploaded by each electronic device, where the second model gradient matrix is a gradient matrix obtained by the electronic device through partial differential privacy calculation, and then determines the target model parameter based on the second model gradient matrix, to train the model through integration of multi-party data, and further improve a model recognition capability. Because a privacy level of the second model gradient matrix uploaded by the electronic device is higher, a reverse attack from an external attacker is well defended against, and security of data in a joint model training process is ensured. The service platform then sends the target model parameter to each electronic device, so that the electronic device updates, based on the target model parameter, the local image tampering detection model shared by the service platform, to obtain the target image tampering detection model having a better recognition capability. Therefore, an effect of detecting a privacy image by the electronic device using the target image tampering detection model is improved, thereby improving accuracy of a recognition result obtained by the electronic device through performing recognition again on the detection result of the target image tampering detection model.

FIG. 6 is a schematic flowchart of an image tampering detection method according to an embodiment of this specification. The method in this embodiment of this specification is performed by a service platform. As shown in FIG. 6, the method in this embodiment of this

specification may include the following steps.

S602. Receive a second model gradient matrix uploaded by at least one electronic device.

The second model gradient matrix is obtained through partial gradient differential privacy calculation processing by the at least one electronic device based on a first model gradient matrix corresponding to a local image tampering detection model. For details of the step in which the electronic device calculates the second model gradient matrix based on the first model gradient matrix, refer to the descriptions of S302 to S306 in the embodiment shown in FIG. 3. Details are not described herein again.

S604: Calculate a target gradient based on each second model gradient matrix.

In some embodiments, performing S604 may specifically include: C2: obtaining a matrix training data volume uploaded by each electronic device, where the matrix training data volume is a training data consuming value when the second model gradient matrix is generated at a model training stage; C4: determining a data volume sum based on all matrix training data volumes; C6: determining a ratio of the matrix training data volume to the data volume sum, and determining the ratio as a gradient weight corresponding to the electronic device; and C8: performing weighted summation based on the gradient weight and the second model gradient matrix, to obtain the target gradient.

The matrix training data volume is a training data consuming value of the electronic device when the second model gradient matrix is obtained. The training data consuming value means a quantity of sample privacy images used by the electronic device when the second model gradient matrix is obtained. Because the second model gradient matrix is obtained based on the first model gradient matrix, the training data consuming value may also be understood as a quantity of sample privacy images used by the electronic device when the first model gradient matrix is obtained by training.

In a process of performing step C4, a sum of all matrix training data volumes may be used as the data volume sum.

In a process of performing step C6, for each matrix training data volume, a ratio of the matrix training data volume to the data volume sum may be calculated, to obtain the gradient weight corresponding to the electronic device.

In a process of performing step C8, for each electronic device, a product of the gradient weight corresponding to the electronic device and the second model gradient matrix uploaded by the electronic device may be calculated, and then a sum calculation is performed on the products to obtain the target gradient.

In this embodiment of this application, a gradient average value is calculated through

weighted summation, to comprehensively evaluate each second model gradient matrix, thereby ensuring accuracy of the gradient used for updating the model parameter.

S606: Obtain a learning rate corresponding to the local image tampering detection model, and calculate a product of the learning rate and the target gradient.

S608: Obtain an initial model parameter corresponding to the local image tampering detection model, calculate a sum of the product and the initial model parameter, and determine the sum as the target model parameter.

It is easy to understand that the initial model parameter is a model parameter obtained by the service platform and the electronic device in a previous round of aggregation training.

In this embodiment of this specification, a formula for calculating the target model parameter may be:

$$w_{t+1} = w_t + \eta \sum_{k=1}^K \frac{n_k}{n} G_i^t *$$

where w_t represents an initial model parameter, w_{t+1} represents a target model parameter, η represents a learning rate, K represents a total quantity of electronic devices, n_k represents a matrix training data volume uploaded by a k^{th} electronic device, n represents a data volume sum, and $G_i^t *$ represents a second model gradient matrix of an i^{th} electronic device.

S610: Send the target model parameter to each electronic device.

The target model parameter is used for instructing the electronic device to update the local image tampering detection model based on the target model parameter, to obtain a target image tampering detection model. Image detection processing is performed on a target privacy image based on the target image tampering detection model, to obtain a target heat detection map corresponding to the target privacy image. Tampering recognition processing is performed on the target heat detection map, to obtain a tampering recognition result corresponding to the target privacy image. For a process in which each electronic device updates the local image tampering detection model based on the target model parameter to obtain the target image tampering detection model, and performs subsequent processing based on the target image tampering detection model, refer to the descriptions of S310 to S314 in the embodiment shown in FIG. 3. Details are not described herein again.

S612: Determine a current number of training rounds of the local image tampering detection model.

S614: Send a training end instruction to each electronic device if the current number of training rounds is greater than or equal to a preset number of rounds.

S612 and S614 are explained and described below.

It is easy to understand that the current number of training rounds is a number of rounds for which the local image tampering detection model has been trained currently.

The preset number of rounds may refer to a preset maximum number of iteration rounds of a model.

The training end instruction is used for instructing each electronic device to stop uploading the second model gradient matrix.

In one or more embodiments of this specification, the determining the current number of training rounds of the local image tampering detection model may be understood as: determining the current number of training rounds based on a current quantity of calculations of the target model parameter, or determining the current number of training rounds based on a quantity of times of receiving the second model gradient matrix corresponding to each electronic device. Further, whether the current number of training rounds is greater than or equal to the preset number of rounds may be determined. When the current number of training rounds is greater than or equal to the preset number of rounds, it represents that a quantity of iterations of the model parameter reaches the maximum number of iteration rounds, and the service platform and the electronic device may stop jointly training the local image tampering detection model. Then, the service platform may send the training end instruction to each electronic device, to instruct, by using the training end instruction, the electronic device to stop jointly training the model. When the current number of training rounds is less than the preset number of rounds, it represents that a quantity of iterations of the model parameter does not reach the maximum number of iteration rounds, and the service platform and the electronic device further need to jointly train the model. Therefore, the service platform does not need to send the training end instruction to the electronic device.

In this embodiment of this specification, the service platform receives the second model gradient matrix uploaded by each electronic device, to determine the target gradient based on the second model gradient matrix, and then determines the target model parameter based on the target gradient. While the model is trained through integration of multi-party data, the accuracy of the gradient is improved through calculation of the average of the target model gradient, and the target model parameter with higher accuracy is calculated by using the gradient with higher accuracy, thereby updating the model by using the target model parameter with higher accuracy, to better improve a model recognition capability. Because noise is added to the second model gradient matrix uploaded by each electronic device, the second model gradient matrix has good privacy. Therefore, the target model parameter calculated by the

service platform based on the second model gradient matrix also has good privacy. The service platform then sends the target model parameter with higher accuracy to each electronic device, so that the electronic device updates, based on the target model parameter with higher accuracy, the local image tampering detection model shared by the service platform, to obtain the target image tampering detection model having a better recognition capability. Therefore, an effect of detecting a privacy image by the electronic device using the target image tampering detection model is improved, thereby improving accuracy of a tampering recognition result obtained by the electronic device through performing recognition again on the detection result of the target image tampering detection model.

An image tampering detection apparatus provided in an embodiment of this application is described in detail below with reference to FIG. 7. It should be noted that the image tampering detection apparatus shown in FIG. 7 is configured to perform the methods in the embodiments shown in FIG. 2 and FIG. 3 of this application. For ease of description, only a part related to this embodiment of this specification is shown. For specific technical details that are not disclosed, refer to the embodiments shown in FIG. 2 and FIG. 3 of this application.

FIG. 7 is a schematic structural diagram of an image tampering detection apparatus according to an embodiment of this specification. The image tampering detection apparatus 1 may be implemented as all or a part of an apparatus by using software, hardware, or a combination of software and hardware. According to some embodiments, the image tampering detection apparatus 1 includes a gradient processing module 11, a gradient sending module 12, a data receiving module 13, and an image processing module 14. Specifically, the gradient processing module 11 is configured to determine a first model gradient matrix corresponding to a local image tampering detection model, and perform partial gradient differential privacy calculation processing based on the first model gradient matrix, to obtain a second model gradient matrix; the gradient sending module 12 is configured to upload the second model gradient matrix to a service platform, where the second model gradient matrix is used for instructing the service platform to determine a target model parameter based on the second model gradient matrix uploaded by at least one electronic device; the data receiving module 13 is configured to: receive the target model parameter sent by the service platform, and update the local image tampering detection model based on the target model parameter, to obtain a target image tampering detection model; and the image processing module 14 is configured to: perform image detection processing on a target privacy image based on the target image tampering detection model, to obtain a target heat detection map corresponding to the target privacy image, and perform tampering recognition processing on the target heat detection map,

to obtain a tampering recognition result corresponding to the target privacy image.

Optionally, the gradient processing module 11 includes: a first processing unit, configured to determine an initial partial gradient subset in the first model gradient matrix; and a second processing unit, configured to: perform noise addition processing on the initial partial gradient subset, to obtain a target partial gradient subset, and update the first model gradient matrix based on the target partial gradient subset, to obtain the second model gradient matrix.

Optionally, the first processing unit includes: a parameter determining unit, configured to: determine a model feature of the local image tampering detection, and determine, based on the model feature, a candidate network parameter corresponding to the local image tampering detection model; and a gradient determining unit, configured to determine the initial partial gradient subset corresponding to the candidate network parameter in the first model gradient matrix.

Optionally, the parameter determining unit is specifically configured to: determine a candidate network layer in the local image tampering detection model based on the model feature; determine a sample mask image corresponding to the sample privacy image, and input the sample mask image to the local image tampering detection model, to obtain a learnable network parameter in the candidate network layer; and determine, based on the learnable network parameter, the candidate network parameter corresponding to the local image tampering detection model.

Optionally, the gradient determining unit is specifically configured to: determine at least one gradient corresponding to each candidate network parameter in the first model gradient matrix; perform filtering processing on the at least one gradient, to obtain a target gradient corresponding to the candidate network parameter; and determine, based on the target gradient, the initial partial gradient subset corresponding to the candidate network parameter.

Optionally, the second processing unit is specifically configured to: perform gradient clipping on the initial partial gradient subset, to obtain a reference partial gradient subset; and perform noise addition processing on the reference partial gradient subset, to obtain the target partial gradient subset.

Optionally, the gradient processing module 11 is specifically configured to: obtain a historical privacy image; and perform model training processing on the local image tampering detection model based on the historical privacy image, to obtain the first model gradient matrix.

Optionally, the image processing module 14 is specifically configured to: obtain a device-dedicated classifier, and perform tampering recognition processing on the target heat detection map based on the device-dedicated classifier, to obtain a tampering recognition result

corresponding to the target privacy image, where device-dedicated classifiers of different electronic devices are not all the same.

An image tampering detection apparatus provided in an embodiment of this application is described in detail below with reference to FIG. 8. It should be noted that the image tampering detection apparatus shown in FIG. 8 is configured to perform the methods in the embodiments shown in FIG. 5 and FIG. 6 of this application. For ease of description, only a part related to this embodiment of this specification is shown. For specific technical details that are not disclosed, refer to the embodiments shown in FIG. 5 and FIG. 6 of this application.

FIG. 8 is a schematic structural diagram of an image tampering detection apparatus according to an embodiment of this specification. The image tampering detection apparatus 1 may be implemented as all or a part of an apparatus by using software, hardware, or a combination of software and hardware. According to some embodiments, the image tampering detection apparatus 1 includes a data receiving module 11, a data processing module 12, and a data sending module 13. Specifically, the data receiving module 11 is configured to receive a second model gradient matrix uploaded by at least one electronic device, where the second model gradient matrix is obtained through partial gradient differential privacy calculation processing by the at least one electronic device based on a first model gradient matrix corresponding to a local image tampering detection model; the data processing module 12 is configured to determine a target model parameter based on each second model gradient matrix; and the data sending module 13 is configured to: send the target model parameter to each electronic device, where the target model parameter is used for instructing the electronic device to update the local image tampering detection model based on the target model parameter, to obtain a target image tampering detection model, perform image detection processing on a target privacy image based on the target image tampering detection model, to obtain a target heat detection map corresponding to the target privacy image, and perform tampering recognition processing on the target heat detection map, to obtain a tampering recognition result corresponding to the target privacy image.

Optionally, the data processing module 12 includes: a first calculating unit, configured to calculate a target gradient based on each second model gradient matrix; a second calculation unit, configured to: obtain a learning rate corresponding to the local image tampering detection model, and calculate a product of the learning rate and the target gradient; and a third calculation unit, configured to: obtain an initial model parameter corresponding to the local image tampering detection model, calculate a sum of the product and the initial model parameter, and determine the sum as the target model parameter.

Optionally, the first calculation unit is specifically configured to: obtain a matrix training data volume uploaded by each electronic device, where the matrix training data volume is a training data consuming value when the second model gradient matrix is generated at a model training stage; determine a data volume sum based on all matrix training data volumes; determine a ratio of the matrix training data volume to the data volume sum, and determine the ratio as a gradient weight corresponding to the electronic device; and perform weighted summation based on the gradient weight and the second model gradient matrix, to obtain the target gradient.

Optionally, the image tampering detection apparatus 1 is further configured to: determine a current number of training rounds of the local image tampering detection model; and send a training end instruction to each electronic device if the current number of training rounds is greater than or equal to a preset number of rounds, where the training end instruction is used for instructing the electronic device to stop uploading the second model gradient matrix.

FIG. 9 is a schematic structural diagram of an electronic device according to an exemplary embodiment of this specification. The electronic device in this specification may include one or more of the following components a processor 110, a memory 120, an input apparatus 130, an output apparatus 140, and a bus 150. The processor 110, the memory 120, the input apparatus 130, and the output apparatus 140 may be connected by the bus 150.

The processor 110 may include one or more processing cores. The processor 110 uses various interfaces and lines to connect various parts of the entire electronic device, and executes various functions of the electronic device 100 and processes data by running or executing an instruction, a program, a code set, or an instruction set stored in the memory 120, and invoking data stored in the memory 120. Optionally, the processor 110 may be implemented in at least one of the following hardware forms: digital signal processing (DSP), field-programmable gate array (FPGA), and programmable logic array (PLA). The processor 110 may integrate one or a combination of several of a central processing unit (CPU), a graphics processing unit (GPU), a modem, and the like. The CPU mainly processes an operating system, a user interface, an application program, and the like. The GPU is configured to be responsible for rendering and drawing display content. The modem is configured to process wireless communication. It may be understood that the modem may alternatively not be integrated into the processor 110, but implemented by using a communication chip independently.

The memory 120 may include a random access memory (RAM), and may alternatively include a read-only memory (ROM). Optionally, the memory 120 includes a non-transitory computer-readable storage medium. The memory 120 may be configured to store an instruction,

a program, code, a code set, or an instruction set. The memory 120 may include a program storage area and a data storage area. The program storage area may store an instruction for implementing an operating system, an instruction for implementing at least one function (such as a touch function, a sound playback function, and an image display function), an instruction for implementing the following method embodiments, and the like. The operating system may be an Android system, including a system deeply developed based on the Android system, an iOS system developed by Apple Inc., including a system deeply developed based on the iOS system, or another system.

To enable the operating system to distinguish specific application scenarios of a third-party application program, data communication between the third-party application program and the operating system needs to be enabled, so that the operating system can obtain current scenario information of the third-party application program at any time, thereby performing targeted system resource adaptation based on a current scenario.

The input apparatus 130 is configured to receive an input instruction or input data. The input apparatus 130 includes, but is not limited to, a keyboard, a mouse, a camera, a microphone, or a touch device. The output apparatus 140 is configured to output an instruction or data. The output apparatus 140 includes, but is not limited to, a display device, a speaker, and the like. In an example, the input apparatus 130 and the output apparatus 140 may be combined, and the input apparatus 130 and the output apparatus 140 are a touch screen.

The touch screen may be designed as an all-screen screen, a curved screen, or a special-shaped screen. The touch display screen may alternatively be designed to be a combination of an all-screen screen and a curved screen, or a combination of a special-shaped screen and a curved screen. This is not limited in this embodiment of this specification.

In addition, it may be understood by a person skilled in the art that the structure of the electronic device shown in the accompanying drawings does not constitute a limitation to the electronic device, and the electronic device may include more or fewer components than those shown, or some components may be combined, or different component arrangements are used. For example, the electronic device further includes components such as a radio frequency circuit, an input unit, a sensor, an audio circuit, a wireless fidelity (Wi-Fi) module, a power supply, and a Bluetooth module, which are not described in detail herein.

In the electronic device shown in FIG. 9, the processor 110 may be configured to invoke a program of an image tampering detection method stored in the memory 120, and specifically perform the following operations: determining a first model gradient matrix corresponding to a local image tampering detection model, and performing partial gradient differential privacy

calculation processing based on the first model gradient matrix, to obtain a second model gradient matrix; uploading the second model gradient matrix to a service platform, where the second model gradient matrix is used for instructing the service platform to determine a target model parameter based on the second model gradient matrix uploaded by at least one electronic device; receiving the target model parameter sent by the service platform, and updating the local image tampering detection model based on the target model parameter, to obtain a target image tampering detection model; and performing image detection processing on a target privacy image based on the target image tampering detection model, to obtain a target heat detection map corresponding to the target privacy image, and performing tampering recognition processing on the target heat detection map, to obtain a tampering recognition result corresponding to the target privacy image.

In an embodiment, when performing the step of performing partial gradient differential privacy calculation processing based on the first model gradient matrix, to obtain the second model gradient matrix, the processor 110 specifically performs the following operations: determining an initial partial gradient subset in the first model gradient matrix; and performing noise addition processing on the initial partial gradient subset, to obtain a target partial gradient subset, and updating the first model gradient matrix based on the target partial gradient subset, to obtain the second model gradient matrix.

In an embodiment, when performing the step of determining the initial partial gradient subset in the first model gradient matrix, the processor 110 specifically performs the following operations: determining a model feature of the local image tampering detection, and determining, based on the model feature, a candidate network parameter corresponding to the local image tampering detection model; and determining the initial partial gradient subset corresponding to the candidate network parameter in the first model gradient matrix.

In an embodiment, when performing the step of determining, based on the model feature, the candidate network parameter corresponding to the local image tampering detection model, the processor 110 specifically performs the following operations: determining a candidate network layer in the local image tampering detection model based on the model feature; determining a sample mask image corresponding to the sample privacy image, and inputting the sample mask image to the local image tampering detection model, to obtain a learnable network parameter in the candidate network layer; and determining, based on the learnable network parameter, the candidate network parameter corresponding to the local image tampering detection model.

In an embodiment, when performing the step of determining the initial partial gradient

subset corresponding to the candidate network parameter in the first model gradient matrix, the processor 110 specifically performs the following operations: determining at least one gradient corresponding to each candidate network parameter in the first model gradient matrix; performing filtering processing on the at least one gradient, to obtain a target gradient corresponding to the candidate network parameter; and determining, based on the target gradient, the initial partial gradient subset corresponding to the candidate network parameter.

In an embodiment, when performing the step of performing noise addition processing on the initial partial gradient subset, to obtain the target partial gradient subset, the processor 110 specifically performs the operations: performing gradient clipping on the initial partial gradient subset, to obtain a reference partial gradient subset; and perform noise addition processing on the reference partial gradient subset, to obtain the target partial gradient subset.

In an embodiment, when performing the step of determining the first model gradient matrix corresponding to the local image tampering detection model, the processor 110 specifically performs the following operations: obtaining a historical privacy image; and performing model training processing on the local image tampering detection model based on the historical privacy image, to obtain the first model gradient matrix.

In an embodiment, when performing the step of performing tampering recognition processing on the target heat detection map, to obtain the tampering recognition result corresponding to the target privacy image, the processor 110 specifically performs the following operations: obtaining a device-dedicated classifier, and performing tampering recognition processing on the target heat detection map based on the device-dedicated classifier, to obtain the tampering recognition result corresponding to the target privacy image, where device-dedicated classifiers of different electronic devices are not all the same.

FIG. 10 is a schematic structural diagram of a service platform according to an exemplary embodiment of this specification. The service platform in this specification may include one or more of the following components a processor 110, a memory 120, an input apparatus 130, an output apparatus 140, and a bus 150. The processor 110, the memory 120, the input apparatus 130, and the output apparatus 140 may be connected by the bus 150.

The processor 110 may include one or more processing cores. The processor 110 uses various interfaces and lines to connect various parts of the entire service platform, and executes various functions of the service platform 100 and processes data by running or executing an instruction, a program, a code set, or an instruction set stored in the memory 120, and invoking data stored in the memory 120. Optionally, the processor 110 may be implemented in at least one of the following hardware forms: digital signal processing (DSP), field-programmable gate

array (FPGA), and programmable logic array (PLA). The processor 110 may integrate one or a combination of several of a central processing unit (CPU), a graphics processing unit (GPU), a modem, and the like. The CPU mainly processes an operating system, a user interface, an application program, and the like. The GPU is configured to be responsible for rendering and drawing display content. The modem is configured to process wireless communication. It may be understood that the modem may alternatively not be integrated into the processor 110, but implemented by using a communication chip independently.

The memory 120 may include a random access memory (RAM), and may alternatively include a read-only memory (ROM). Optionally, the memory 120 includes a non-transitory computer-readable storage medium. The memory 120 may be configured to store an instruction, a program, code, a code set, or an instruction set. The memory 120 may include a program storage area and a data storage area. The program storage area may store an instruction for implementing an operating system, an instruction for implementing at least one function (such as a touch function, a sound playback function, and an image display function), an instruction for implementing the following method embodiments, and the like. The operating system may be an Android system, including a system deeply developed based on the Android system, an iOS system developed by Apple Inc., including a system deeply developed based on the iOS system, or another system.

To enable the operating system to distinguish specific application scenarios of a third-party application program, data communication between the third-party application program and the operating system needs to be enabled, so that the operating system can obtain current scenario information of the third-party application program at any time, thereby performing targeted system resource adaptation based on a current scenario.

The input apparatus 130 is configured to receive an input instruction or input data. The input apparatus 130 includes, but is not limited to, a keyboard, a mouse, a camera, a microphone, or a touch device. The output apparatus 140 is configured to output an instruction or data. The output apparatus 140 includes, but is not limited to, a display device, a speaker, and the like. In an example, the input apparatus 130 and the output apparatus 140 may be combined, and the input apparatus 130 and the output apparatus 140 are a touch screen.

The touch screen may be designed as an all-screen screen, a curved screen, or a special-shaped screen. The touch display screen may alternatively be designed to be a combination of an all-screen screen and a curved screen, or a combination of a special-shaped screen and a curved screen. This is not limited in this embodiment of this specification.

In addition, it may be understood by a person skilled in the art that the structure of the

service platform shown in the accompanying drawings does not constitute a limitation to the service platform, and the service platform may include more or fewer components than those shown, or some components may be combined, or different component arrangements are used. For example, the electronic device further includes components such as a radio frequency circuit, an input unit, a sensor, an audio circuit, a wireless fidelity (Wi-Fi) module, a power supply, and a Bluetooth module, which are not described in detail herein.

In the service platform shown in FIG. 10, the processor 110 may be configured to invoke a program of an image tampering detection method stored in the memory 120, and specifically performs the following operations: receiving a second model gradient matrix uploaded by at least one electronic device, where the second model gradient matrix is obtained through partial gradient differential privacy calculation processing by the at least one electronic device based on a first model gradient matrix corresponding to a local image tampering detection model; determining a target model parameter based on each second model gradient matrix; and sending the target model parameter to each electronic device, where the target model parameter is used for instructing the electronic device to update the local image tampering detection model based on the target model parameter, to obtain a target image tampering detection model, performing image detection processing on a target privacy image based on the target image tampering detection model, to obtain a target heat detection map corresponding to the target privacy image, and performing tampering recognition processing on the target heat detection map, to obtain a tampering recognition result corresponding to the target privacy image.

In an embodiment, when performing the step of determining the target model parameter based on each second model gradient matrix, the processor 110 specifically performs the following operations: calculating a target gradient based on each second model gradient matrix; obtaining a learning rate corresponding to the local image tampering detection model, and calculating a product of the learning rate and the target gradient; and obtaining an initial model parameter corresponding to the local image tampering detection model, calculating a sum of the product and the initial model parameter, and determining the sum as the target model parameter.

In an embodiment, when performing the step of calculating the target gradient based on each second model gradient matrix, the processor 110 specifically performs the following operations: obtaining a matrix training data volume uploaded by each electronic device, where the matrix training data volume is a training data consuming value when the second model gradient matrix is generated at a model training stage; determining a data volume sum based on all matrix training data volumes; determining a ratio of the matrix training data volume to

the data volume sum, and determining the ratio as a gradient weight corresponding to the electronic device; and performing weighted summation based on the gradient weight and the second model gradient matrix, to obtain the target gradient.

In an embodiment, the processor 110 further performs the following operations: determining a current number of training rounds of the local image tampering detection model; and sending a training end instruction to each electronic device if the current number of training rounds is greater than or equal to a preset number of rounds, where the training end instruction is used for instructing the electronic device to stop uploading the second model gradient matrix.

It should be noted that information (including but not limited to user equipment information, user personal information, and the like), data (including but not limited to data for analysis, stored data, displayed data, and the like) involved in the embodiments of this specification are all authorized by a user or information and data fully authorized by all parties. The collection, use, and processing of relevant data need to conform to relevant laws, regulations, and standards of relevant countries and regions. For example, the historical privacy image, the target privacy image, and the like involved in this specification are all obtained with sufficient authorization.

An embodiment of this specification further provides a computer program product, the computer program product has at least one instruction stored therein, and the at least one instruction is loaded by the processor to perform the image tampering detection method described in the embodiments.

A person skilled in the art should be aware that in the foregoing one or more examples, functions described in the embodiments of this specification may be implemented by hardware, software, firmware, or any combination thereof. When implemented by using software, the functions may be stored in a computer-readable medium or may be used as one or more instructions or code in a computer-readable medium for transmission. The computer-readable medium includes a computer storage medium and a communication medium, where the communication medium includes any medium that enables a computer program to be transmitted from one place to another. The storage medium may be any available medium accessible to a general-purpose or a dedicated computer.

Described above are merely optional embodiments of this specification, and are not intended to limit this specification. Within the spirit and principles of this specification, any modifications, equivalent replacements, improvements, and the like shall fall within the protection scope of this specification.

Particular embodiments of this specification are described above. Other embodiments fall

within a scope of the appended claims. In some implementations, the actions or steps recited in the claims may be performed in an order different from that described in the embodiments while still achieving the desired results. In addition, the processes depicted in the accompanying drawings are not necessarily performed in a specific order or sequential order to achieve the desired results. In some implementations, multitasking and parallel processing are also possible or may be advantageous.

CLAIMS

What is claimed is:

1. An image tampering detection method, applied to an electronic device, wherein the method comprises:

determining a first model gradient matrix corresponding to a local image tampering detection model, and performing partial gradient differential privacy calculation processing based on the first model gradient matrix, to obtain a second model gradient matrix;

uploading the second model gradient matrix to a service platform, wherein the second model gradient matrix is used for instructing the service platform to determine a target model parameter based on the second model gradient matrix uploaded by at least one electronic device;

receiving the target model parameter sent by the service platform, and updating the local image tampering detection model based on the target model parameter, to obtain a target image tampering detection model; and

performing image detection processing on a target privacy image based on the target image tampering detection model, to obtain a target heat detection map corresponding to the target privacy image, and performing tampering recognition processing on the target heat detection map, to obtain a tampering recognition result corresponding to the target privacy image.

2. The method according to claim 1, wherein the performing partial gradient differential privacy calculation processing based on the first model gradient matrix, to obtain a second model gradient matrix comprises:

determining an initial partial gradient subset in the first model gradient matrix; and

performing noise addition processing on the initial partial gradient subset, to obtain a target partial gradient subset, and updating the first model gradient matrix based on the target partial gradient subset, to obtain the second model gradient matrix.

3. The method according to claim 2, wherein the determining an initial partial gradient subset in the first model gradient matrix comprises:

determining a model feature of the local image tampering detection, and determining, based on the model feature, a candidate network parameter corresponding to the local image tampering detection model; and

determining the initial partial gradient subset corresponding to the candidate network parameter in the first model gradient matrix.

4. The method according to claim 3, wherein the determining, based on the model feature,

a candidate network parameter corresponding to the local image tampering detection model comprises:

- determining a candidate network layer in the local image tampering detection model based on the model feature;

- determining a sample mask image corresponding to the sample privacy image, and inputting the sample mask image to the local image tampering detection model, to obtain a learnable network parameter in the candidate network layer; and

- determining, based on the learnable network parameter, the candidate network parameter corresponding to the local image tampering detection model.

5. The method according to claim 3, wherein the determining the initial partial gradient subset corresponding to the candidate network parameter in the first model gradient matrix comprises:

- determining at least one gradient corresponding to each candidate network parameter in the first model gradient matrix;

- performing filtering processing on the at least one gradient, to obtain a target gradient corresponding to the candidate network parameter; and

- determining, based on the target gradient, the initial partial gradient subset corresponding to the candidate network parameter.

6. The method according to claim 2, wherein the performing noise addition processing on the initial partial gradient subset, to obtain a target partial gradient subset comprises:

- performing gradient clipping on the initial partial gradient subset, to obtain a reference partial gradient subset; and

- performing noise addition processing on the reference partial gradient subset, to obtain the target partial gradient subset.

7. The method according to claim 1, wherein the determining a first model gradient matrix corresponding to a local image tampering detection model comprises:

- obtaining a historical privacy image; and

- performing model training processing on the local image tampering detection model based on the historical privacy image, to obtain the first model gradient matrix.

8. The method according to claim 1, wherein the performing tampering recognition processing on the target heat detection map, to obtain a tampering recognition result corresponding to the target privacy image comprises:

- obtaining a device-dedicated classifier, and performing tampering recognition processing on the target heat detection map based on the device-dedicated classifier, to obtain the

tampering recognition result corresponding to the target privacy image, wherein device-dedicated classifiers of different electronic devices are not all the same.

9. An image tampering detection method, applied to a service platform, wherein the method comprises:

receiving a second model gradient matrix uploaded by at least one electronic device, wherein the second model gradient matrix is obtained through partial gradient differential privacy calculation processing by the electronic device based on a first model gradient matrix corresponding to a local image tampering detection model;

determining a target model parameter based on each second model gradient matrix; and

sending the target model parameter to each electronic device, wherein the target model parameter is used for instructing the electronic device to update the local image tampering detection model based on the target model parameter, to obtain a target image tampering detection model, performing image detection processing on a target privacy image based on the target image tampering detection model, to obtain a target heat detection map corresponding to the target privacy image, and performing tampering recognition processing on the target heat detection map, to obtain a tampering recognition result corresponding to the target privacy image.

10. The method according to claim 9, wherein the determining a target model parameter based on each second model gradient matrix comprises:

calculating a target gradient based on each second model gradient matrix;

obtaining a learning rate corresponding to the local image tampering detection model, and calculating a product of the learning rate and the target gradient; and

obtaining an initial model parameter corresponding to the local image tampering detection model, calculating a sum of the product and the initial model parameter, and determining the sum as the target model parameter.

11. The method according to claim 10, wherein the calculating a target gradient based on each second model gradient matrix comprises:

obtaining a matrix training data volume uploaded by each electronic device, wherein the matrix training data volume is a training data consuming value when the second model gradient matrix is generated at a model training stage;

determining a data volume sum based on all matrix training data volumes;

determining a ratio of the matrix training data volume to the data volume sum, and determining the ratio as a gradient weight corresponding to the electronic device; and

performing weighted summation based on the gradient weight and the second model

gradient matrix, to obtain the target gradient.

12. The method according to claim 9, further comprising:

determining a current number of training rounds of the local image tampering detection model; and

sending a training end instruction to each electronic device if the current number of training rounds is greater than or equal to a preset number of rounds, wherein the training end instruction is used for instructing the electronic device to stop uploading the second model gradient matrix.

13. An image tampering detection apparatus, used in an electronic device, wherein the apparatus comprises:

a gradient processing module, configured to determine a first model gradient matrix corresponding to a local image tampering detection model, and perform partial gradient differential privacy calculation processing based on the first model gradient matrix, to obtain a second model gradient matrix;

a gradient sending module, configured to upload the second model gradient matrix to a service platform, wherein the second model gradient matrix is used for instructing the service platform to determine a target model parameter based on the second model gradient matrix uploaded by at least one electronic device;

a data receiving module, configured to receive the target model parameter sent by the service platform, and update the local image tampering detection model based on the target model parameter, to obtain a target image tampering detection model; and

an image processing module, configured to perform image detection processing on a target privacy image based on the target image tampering detection model, to obtain a target heat detection map corresponding to the target privacy image, and perform tampering recognition processing on the target heat detection map, to obtain a tampering recognition result corresponding to the target privacy image.

14. An image tampering detection apparatus, used in a service platform, wherein the apparatus comprises:

a data receiving module, configured to receive a second model gradient matrix uploaded by at least one electronic device, wherein the second model gradient matrix is obtained through partial gradient differential privacy calculation processing by the at least one electronic device based on a first model gradient matrix corresponding to a local image tampering detection model;

a data processing module, configured to determine a target model parameter based on each

second model gradient matrix; and

a data sending module, configured to send the target model parameter to each electronic device, wherein the target model parameter is used for instructing the electronic device to update the local image tampering detection model based on the target model parameter, to obtain a target image tampering detection model, perform image detection processing on a target privacy image based on the target image tampering detection model, to obtain a target heat detection map corresponding to the target privacy image, and perform tampering recognition processing on the target heat detection map, to obtain a tampering recognition result corresponding to the target privacy image.

15. A computer storage medium, having a plurality of instructions stored therein, wherein the instructions are adapted to be loaded by a processor and perform the steps in the method according to any one of claims 1 to 8 or 9 to 12.

16. A computer program product, having at least one instruction stored therein, wherein the at least one instruction is executed by a processor and perform the steps in the method according to any one of claims 1 to 8 or 9 to 12.

17. An electronic device, comprising: a processor and a memory, wherein the memory has a computer program stored therein, and the computer program is adapted to be loaded by the processor and perform the steps in the method according to any one of claims 1 to 8 or 9 to 12.