

MULTI-MODEL FUSION BASED METHOD AND DEVICE FOR IMAGE ANNOTATION, COMPUTER DEVICE, AND STORAGE MEDIUM

CROSS-REFERENCE TO RELATED APPLICATION(S)

[1] This application claims priority to Chinese Patent Application No. 201910268876.7, filed on April 4, 2019, and entitled “MULTI-MODEL FUSION BASED METHOD AND DEVICE FOR IMAGE ANNOTATION, COMPUTER DEVICE, AND STORAGE MEDIUM”, the entire disclosure of which is hereby incorporated by reference.

TECHNICAL FIELD

[2] This disclosure relates to the field of image processing, and particularly to a multi-model fusion based method and device for image annotation, a computer device, and a storage medium.

BACKGROUND

[3] In some application scenarios, it is necessary to segment a target region of interest in an image. However, due to fine-grained changes in some images (e.g., a skin lesion image), it becomes challenging for a system to automatically classify regions of the image. Although the existing deep convolutional neural network (CNN) is often used to classify various fine-grained objects and performs well in multiple tasks, it still has the problem of low accuracy.

SUMMARY

[4] According to implementations of the disclosure, a multi-model fusion based method for image annotation, a multi-model fusion based device for image annotation, a computer device, and a storage medium are provided, which can well realize annotation of a region of interest in an image to-be-annotated. Moreover, Dice and intersection over union (IoU) of the annotation are relatively high.

[5] In a first aspect of the disclosure, a multi-model fusion based method for image annotation is provided. An image to-be-annotated is acquired and preprocessed to obtain a plurality of instance images. The instance images are inputted into a DenseNet network-based classification model for binary classification. Binary-classification results corresponding to the instance images are connected to obtain a classification-result image. The instance images are inputted into a Vnet network-FPN network-based segmentation model for binarization

segmentation. Binarization-segmentation results corresponding to the instance images are connected to obtain a segmentation-result image. A binarization fusion image is calculated according to the classification-result image and the segmentation-result image. A contour of the fusion image is extracted, and a region of interest is annotated in the image to-be-annotated according to the contour.

[6] In a second aspect of the disclosure, a multi-model fusion based device for image annotation is provided. The device includes a preprocessing module, a binary-classification module, a classification-connecting module, a binarization-segmentation module, a segmentation-connecting module, a fusion module, and an annotation module. The preprocessing module is configured to acquire and preprocess an image to-be-annotated to obtain a plurality of instance images. The binary-classification module is configured to input the instance images into a DenseNet network-based classification model for binary classification. The classification-connecting module is configured to connect binary-classification results corresponding to the instance images to obtain a classification-result image. The binarization-segmentation module is configured to input the instance images into a Vnet network-FPN network-based segmentation model for binarization segmentation. The segmentation-connecting module is configured to connect binarization-segmentation results corresponding to the instance images to obtain a segmentation-result image. The fusion module is configured to calculate a binarization fusion image according to the classification-result image and the segmentation-result image. The annotation module is configured to extract a contour of the fusion image and annotate a region of interest in the image to-be-annotated according to the contour.

[7] In a third aspect of the disclosure, a computer device is provided. The computer device includes a memory and a processor. The memory is configured to store computer programs. The processor is configured to execute the computer programs which, when executed by the processor, cause the processor to carry out the foregoing multi-model fusion based method for image annotation.

[8] In a fourth aspect of the disclosure, a computer readable storage medium is provided. The computer readable storage medium is configured to store computer programs which, when executed by a processor, cause the processor to carry out the foregoing multi-model fusion based method for image annotation.

[9] According to the multi-model fusion based method and device for image annotation, the computer device, and the storage medium of the disclosure, the DenseNet network-based classification model can perform binary classification on the instance images

of the image to-be-annotated to obtain the classification-result image, and the Vnet network-FPN network-based segmentation model can perform binarization segmentation on the instance images of the image to-be-annotated to obtain the segmentation-result image. Then the classification-result image and the segmentation-result image are fused and the contour used to annotate the region of interest in the image to-be-annotated is extracted. In this way, the Dice and the IoU of the annotation can be increased, and the false positive rate can be reduced.

BRIEF DESCRIPTION OF THE DRAWINGS

[10] In order to describe technical solutions of implementations more clearly, the following will give a brief description of accompanying drawings used for describing the implementations. Apparently, accompanying drawings described below are merely some implementations. Those of ordinary skill in the art can also obtain other accompanying drawings based on the accompanying drawings described below without creative efforts.

[11] FIG. 1 is a schematic flowchart illustrating a multi-model fusion based method for image annotation according to an implementation.

[12] FIG. 2 is a schematic diagram illustrating an application scenario of the method for image annotation according to an implementation.

[13] FIG. 3 is a schematic diagram illustrating an image to-be-annotated.

[14] FIG. 4 is a schematic diagram illustrating instance images obtained by processing the image to-be-annotated in FIG. 3.

[15] FIG. 5 is a schematic structural diagram illustrating a DenseNet network.

[16] FIG. 6 is a specific schematic flowchart illustrating binary classification performed by a classification model according to an implementation.

[17] FIG. 7 is a schematic structural diagram illustrating a dense block in a DenseNet network.

[18] FIG. 8 is a specific schematic flowchart illustrating binary classification performed by a classification model according to an implementation.

[19] FIG. 9 is a schematic diagram illustrating a classification-result image.

[20] FIG. 10 is a schematic structural diagram illustrating a Vnet network of a segmentation model.

[21] FIG. 11 is a schematic structural diagram illustrating an FPN network.

[22] FIG. 12 is a specific schematic flowchart of performing binarization segmentation through a segmentation model according to an implementation.

- [23] FIG. 13 is a specific schematic flowchart of outputting by an FPN network a binarization-segmentation result according to a result of up-sampling on a Vnet network.
- [24] FIG. 14 is a schematic diagram illustrating binarization-segmentation results corresponding to instance images.
- [25] FIG. 15 is a schematic diagram illustrating a segmentation-result image.
- [26] FIG. 16 is a schematic flowchart illustrating a multi-model fusion based method for image annotation according to another implementation.
- [27] FIG. 17 is a specific schematic flowchart of obtaining a binarization fusion image by taking intersection of a classification-result image and a segmentation-result image.
- [28] FIG. 18 is a schematic flowchart illustrating a multi-model fusion based method for image annotation according to still another implementation.
- [29] FIG. 19 is a schematic diagram illustrating annotating a region of interest in an image to-be-annotated.
- [30] FIG. 20 is a schematic structural diagram illustrating a multi-model fusion based device for image annotation according to an implementation.
- [31] FIG. 21 is a schematic structural diagram illustrating a multi-model fusion based device for image annotation according to another implementation.
- [32] FIG. 22 is a schematic structural diagram illustrating a computer device according to an implementation.

DETAILED DESCRIPTION

- [33] Hereinafter, technical solutions embodied by the implementations of the disclosure will be described in a clear and comprehensive manner with reference to the accompanying drawings intended for the implementations. It is evident that the implementations described herein constitute merely some rather than all of the implementations of the disclosure, and that those of ordinary skill in the art will be able to derive other implementations based on these implementations without making creative efforts, which all such derived implementations shall all fall in the protection scope of the disclosure.
- [34] The flowcharts illustrated in the accompanying drawings are merely illustrative, all contents and operations/steps in the flowcharts are not necessarily included, nor does it have to be executed in the described order. For example, some operations/steps can also be decomposed, combined or partially combined, so an actual execution order may be changed according to actual conditions. In addition, although functional modules are divided in device schematic diagrams, in some cases, they may be divided in a manner different from division

of the modules of the device schematic diagrams.

[35] According to implementations of the disclosure, a multi-model fusion based method and device for image annotation, a computer device, and a storage medium are provided. The multi-model fusion based method for image annotation can be applicable to a terminal or a server to implement annotation of a region of interest in an image.

[36] As an example, the multi-model fusion based method for image annotation is applicable to a server. As another example, the multi-model fusion based method for image annotation is applicable to a terminal, such as a mobile phone, a notebook, a desktop, and so on. In order to facilitate understanding, implementations of the disclosure will be described in detail with reference to a scenario where the multi-model fusion based method for image annotation is applied to a server.

[37] Hereinafter, the implementations of the disclosure will be described in detail with reference to the accompanying drawings. The following implementations and features of the implementations can be combined with each other without conflict.

[38] FIG. 1 a schematic flowchart illustrating a multi-model fusion based method for image annotation according to some implementations.

[39] As illustrated in FIG. 1, the multi-model fusion based method for image annotation includes the following.

[40] At block S110, an image to-be-annotated is acquired and preprocessed to obtain a plurality of instance images.

[41] In some implementations, the image to-be-annotated is an image stored locally in a device for implementing the multi-model fusion based method for image annotation, an image obtained by the device from a network, an image obtained by the device from an input device which is coupled thereto, an image obtained by the device from other electronic devices, or the like.

[42] As illustrated in FIG. 2, a server obtains an image to-be-annotated from a terminal. Both the server and the terminal are connected to Internet. The image to-be-annotated is uploaded by a user at the terminal to the server. The server annotates a region of interest in the image to-be-annotated by adopting a multi-model fusion based method for image annotation. An annotation result can be sent back to the terminal.

[43] The image to-be-annotated can be a color image, a grayscale image, and so on. In one implementation, after obtaining the image to-be-annotated, some preprocessing can be done for the image to-be-annotated, such as size adjustment, cropping, rotation, and/or image-algorithm processing, etc. The image-algorithm processing may include color

temperature adjustment algorithm, exposure adjustment algorithm, contrast adjustment algorithm, highlight restoration algorithm, low light compensation algorithm, white balance algorithm, sharpness adjustment algorithm, fog algorithm index, natural saturation adjustment algorithm, etc.

[44] For example, as illustrated in FIG. 3, the image to-be-annotated obtained by the server is a pathological image of skin basal cell carcinoma.

[45] In some implementations, if the size of the image to-be-annotated is relatively large, the image to-be-annotated can be processed into several instance images of preset sizes (i.e., a patch) through preprocessing such as image clipping. As illustrated in FIG. 4, two instance images are illustrated. It should be noted that, the image to-be-annotated of FIG. 3 and the instance image of FIG. 4 are not drawn to scale and actual size.

[46] At block S120, the instance images are inputted into a DenseNet network-based classification model for binary classification.

[47] According to a pre-trained DenseNet network-based classification model (e.g., a DenseNet121 network-based classification model), the instance images are predicted.

[48] The DenseNet network is a densely connected convolutional neural network, which can alleviate the vanishing-gradient problem, strengthen feature propagation, encourage feature reuse, and greatly reduce the number of parameters. As illustrated in FIG. 5, a DenseNet network includes multiple dense blocks. There is a transition block (i.e., a transition layer) between two adjacent dense blocks.

[49] In some implementations, as illustrated in FIG. 6, the operation of inputting the instance images into the DenseNet network-based classification model for binary classification at block S120 includes the following.

[50] At block S121, for each of the instance images, when the DenseNet network-based classification model predicts that the instance image has a region of interest, determine that a binary-classification result corresponding to the instance image is a first-value block.

[51] If the DenseNet network-based classification model detects that an instance image has the region of interest (e.g., a cancerous region), the classification model classifies the instance image as the first-value block during classification. As an example, the first-value block has a same size (that is, length, width, and the like) with the instance image.

[52] As an example, a first value is 255, and in a grayscale image or a black-and-white image, the first-value block is a white block.

[53] At block S122, when the DenseNet network-based classification model predicts that the instance image has no region of interest, determine that a binary-classification result

corresponding to the instance image is a second-value block.

[54] If the DenseNet network-based classification model detects that an instance image has no region of interest (e.g., a cancerous region), the classification model classifies the instance image as the second-value block during classification. As an example, the second-value block has a same size with the instance image, that is, the length and width of the second-value block are equal to these of the instance image.

[55] As an example, a second value is 0, and in a grayscale image or a black-and-white image, the second-value block is a black block.

[56] The region of interest for example is a cancerous region. If the DenseNet network-based classification model detects that an instance image has the region of interest, a binary-classification result corresponding to the instance image is determined to be a positive block. If the DenseNet network-based classification model detects that an instance image has no region of interest, a binary-classification result corresponding to the instance image is determined to be a negative block.

[57] In other implementations, if the DenseNet network-based classification model predicts that proportion of a region of interest of an instance image (i.e., a ratio of the region of interest to the instance image) is greater than a preset threshold, a binary-classification result corresponding to the instance image is determined to be the first-value block. If the DenseNet network-based classification model predicts that the proportion of the region of interest of the instance image is not greater than the preset threshold, the binary-classification result corresponding to the instance image is determined to be the second-value block.

[58] FIG. 7 is a schematic structural diagram illustrating a dense block in a DenseNet network. There is a direct connection between any two layers in each dense block. That is, for each layer of the DenseNet network, an input of the layer is concatenation of outputs of all layers prior to the layer (i.e., preceding layers), and a feature map learned in the layer is directly passed to all layers subsequent to the layer (i.e., subsequent layers) as inputs of the subsequent layers. In this way, the DenseNet network can maximize information flow between all layers of the DenseNet network.

[59] In some implementations, by adopting the above-mentioned architecture, on the one hand, because each layer receives gradient signals of all subsequent layers during back propagation, it is possible to avoid the problem that a gradient near an input layer becomes smaller with the increase of network depth. As such, features of a lower layer can be reflected in a higher layer. Therefore, vanishing-gradient during training can be alleviated. On the other hand, since a large number of features are multiplexed, a relatively small number of

convolution kernels are needed to generate a large number of features. Therefore, the size of the model is relatively small.

[60] In some implementations, as illustrated in FIG. 8, the DenseNet network-based classification model performs binary classification on the inputted instance images as follows.

[61] At block S11, for each of the inputted instance images, the instance image is preprocessed to obtain a feature map.

[62] As an example, the inputted instance image has a size of [512, 512, 3], that is, the length and width of the instance image are both 512 pixels, and the instance image has three channels. The instance image is an RGB three-channel image for example.

[63] In one implementation, for each of the inputted instance images, the instance image is preprocessed as follows.

[64] Zero padding and convolution: the instance image is processed through a zero-padding layer (e.g., ZeroPadding2D) to have a zero-padded image with size of [518, 518, 3], and the zero-padded image is convolved through a convolution layer (e.g., Conv2D) to output a feature map of [256, 256, 64]. The zero-padding layer is mainly used to pad zeros on both edges of a single matrix, to control the size of the feature map in subsequent processing (e.g., convolution).

[65] Batch normalization: the feature map of [256, 256, 64] is batch normalized through a normalization layer (e.g., Batch Normalization), to output a batch-normalized feature map of [256, 256, 64]. The normalization layer is mainly used to normalize a distribution state of data to a standard normal distribution, to speed up convergence speed when training the model. The position and the amount of the normalization layer are not limited herein, for example, the normalization layer can be set at the beginning of the DenseNet network, before the convolution layer, or after the convolution layer. In one implementation, the normalization layer is set after the convolution layer to increase convergence speed of structured feature data obtained by convolution.

[66] Non-linear processing: non-linear processing is performed on the batch-normalized feature map of [256, 256, 64] through an activation layer. In one implementation, the non-linear processing can be implemented based on a non-linear function ReLU. Non-linear function ReLU is an element-wise operation and can set all pixel values less than 0 in the feature map to be zero. Non-linear processing aims to introduce nonlinearity into the DenseNet network.

[67] Zero padding and pooling: a non-linearly processed feature map of [256, 256, 64] is processed through a zero-padding layer (e.g., ZeroPadding2D) to be a feature map of [258,

258, 64], and pooling processing is performed on the zero-padded feature map through a pooling layer (e.g., MaxPooling2D), to output a feature map of [128, 128, 64].

[68] In one implementation, the pooling is embodied as max pooling. Max pooling has a characteristic of invariance. The invariance includes translation invariance, rotation invariance, and scale invariance. Max pooling can also reduce the amount of parameters and calculations while retaining main features, prevent overfitting, and improve generalization ability of the model.

[69] At block S12, the number of channels of the preprocessed feature map is increased according to a first dense block in a DenseNet network.

[70] As an example, the number of channels of the feature map is increased according to a first dense block in FIG. 5. In one implementation, the following operations are performed in a loop at the first dense block in FIG. 7: batch normalization, a non-linear processing in an activation layer, a convolution processing in a convolution layer, batch normalization, a non-linear processing in an activation layer, a convolution processing in a convolution layer, and a concatenation processing in a concatenation layer. The concatenation layer establishes a concatenation between a layer after the concatenation layer and all layers before the concatenation layer, which can reduce the number of network layers and facilitate network training. Each time the loop is completed, the number of channels of the feature map increases by 32 while the size of the feature map remains unchanged.

[71] At block S13, the size of the feature map outputted by the first dense block is reduced according to a first transition block in the DenseNet network.

[72] In one implementation, the following operations are performed in a loop at the transition block: batch normalization, a non-linear processing in an activation layer, a convolution processing in a convolution layer, a pooling processing. The convolution layer has a convolution kernel of 1×1 , the pooling adopts average pooling, and pooling window is 2×2 . As such, while reducing dimensions, it is conducive to transfer information to a next step for feature extraction, reduce errors of feature extraction caused by the increase in variance of an estimated value due to a limited neighborhood size, and retain relatively more image background information. Each time the loop is completed, the size of the feature map becomes half of an initial size of the feature map while the number of the channels of the feature map remains unchanged.

[73] At block S14, the number of channels of the feature map outputted by the first transition block is increased according to a second dense block in the DenseNet network, and the size of the feature map outputted by the second dense block is reduced according to a

second transition block in the DenseNet network.

[74] At block S15, the number of channels of the feature map outputted by the second transition block is increased according to a third dense block in the DenseNet network, and the size of the feature map outputted by the third dense block is reduced according to a third transition block in the DenseNet network.

[75] At block S16, the number of channels of the feature map outputted by the third transition block is increased according to a fourth dense block in the DenseNet network.

[76] As an example, the feature map outputted by the fourth dense block (i.e., dense block 4) in the DenseNet network has a size of [16, 16, 1024], that is, the length and width of the feature map are 16 pixels, and the feature map has 1024 channels.

[77] At block S17, perform batch normalization and global average pooling on the feature map outputted by the fourth dense block.

[78] As an example, after performing batch normalization and global average pooling on the feature map of [16, 16, 1024] outputted by the fourth dense block, a one-dimensional feature vector with a length of 1024 is outputted.

[79] At block S18, according to a feature vector obtained by performing global average pooling, perform binary classification on the instance image corresponding to the feature vector.

[80] As an example, the instance image is predicted according to the one-dimensional feature vector with a length of 1024, and a binary-classification result may be the first-value block or the second-value block.

[81] At block S130, binary-classification results corresponding to the instance images are connected to obtain a classification-result image.

[82] As an example, multiple instance images are obtained after preprocessing the image to-be-annotated of FIG. 3. The DenseNet network-based classification model performs prediction on each of the instance images. Prediction results for respective instance images are connected to obtain a classification-result image in FIG. 9. A black region in FIG. 9 includes multiple second-value blocks, where the second-value block corresponds to an instance image which has no region of interest. A white region in FIG. 9 includes multiple first-value blocks, where the first-value block corresponds to an instance image which has the region of interest.

[83] At block S140, the instance images are inputted into a Vnet network-FPN network-based segmentation model for binarization segmentation.

[84] A Vnet network is suitable for medical image segmentation of three-dimensional

structure, and can realize end-to-end image semantic segmentation of 3D images. FIG. 10 is a schematic structural diagram illustrating a Vnet network. The Vnet network includes a down-sampling (encoder) part and an up-sampling (decoder) part, and the entire Vnet network is V-shaped.

[85] As illustrated in FIG. 11, a feature pyramid network (FPN) includes a bottom-up line, a top-down line, and a lateral connection. An enlarged area in a dashed box of FIG. 11 illustrates the lateral connection. A main function of 1×1 convolution kernel is to reduce the number of convolution kernels (i.e., reduce the number of feature maps) without changing the size of the feature maps.

[86] Specifically, a bottom-up part of the FPN actually corresponds to a normal forward propagation process of a neural network, and a feature map usually becomes smaller after it is calculated through a convolution kernel. A top-down process adopts up sampling to up-sample a more abstract and semantic high-level feature map. Then the features are laterally connected to features of a preceding layer, with such lateral connection, a result of the up-sampling is merged with a feature map of a same size generated after the bottom-up process. A feature map used for prediction in each layer contains features of different resolutions and different semantic strengths, which can realize detection of an object with a corresponding resolution, thereby ensuring that each layer has an appropriate resolution and a strong semantic feature. Two lateral connected features have a same spatial size, and can use the bottom layer to locate detailed information.

[87] FPN mainly focus on a multi-scale problem in object detection. By simply changing network connections, performance of small object detection can be greatly improved without any increase in calculation amount of an original model. The Vnet network-FPN network-based segmentation model improves effect of binarization segmentation based on a network structure that combines the Vnet network and the FPN network.

[88] In some implementations, as illustrated in FIG. 12, the operation of inputting the instance images into the Vnet network-FPN network-based segmentation model for binarization segmentation at block S140 includes the following.

[89] At block S141, for each of the inputted instance images, the instance image is inputted into a Vnet network of the segmentation model.

[90] As an example, the instance image has a size of $[1024, 1024, 3]$, where 3 represents 3 channels (e.g., RGB channels).

[91] At block S142, non-linear activation, a plurality of down-sampling, and a plurality

of up-sampling are sequentially performed on the instance image based on the Vnet network.

[92] As an example, the non-linear activation (processing) includes performing batch normalization and convolution on the inputted instance image, and then processing the inputted instance image through PReLU activation layer; the number of filters for the convolution is 16, the size of a convolution kernel is 5×5 , and padding can be used to ensure that the size of a feature map remains unchanged.

[93] As an example, five times of down-sampling (encoder) are sequentially performed on a feature map subjected to non-linear activation through a convolution layer. For each of previous four down-sampling, batch normalization and PReLU activation are performed, and then dimensionality reduction operation is performed after adding corresponding positions of an input feature map of the down-sampling and a feature map subjected to the down-sampling, to reduce the size of the image, where reduction factor is 2. No dimensionality reduction operation is performed during the fifth down-sampling. In one implementation, for the first down-sampling, an input feature map has a size of [1024, 1024, 16], and an output feature map has a size of [512, 512, 32]; for the second down-sampling, an input feature map has a size of [512, 512, 32], and an output feature map has a size of [256, 256, 64]; for the third down-sampling, an input feature map has a size of [256, 256, 64], and an output feature map has a size of [128, 128, 128]; for the fourth down-sampling, an input feature map has a size of [128, 128, 128], and an output feature map has a size of [64, 64, 256]; for the fifth down-sampling, an input feature map has a size of [64, 64, 256], and an output feature map has a size of [64, 64, 256], and no dimensionality reduction operation is performed.

[94] As an example, four times of up-sampling (decoder) are performed on a feature map outputted after several times of down-sampling. There are two inputs for each up-sampling: an input of a corresponding down-sampling passed through fine-grained features forwarding and an output of a previous process. First, the two inputs are merged together through a concatenation layer. Then batch normalization and PReLU activation are performed on a merged result. After performing element-wise sum on an output of batch normalization and PReLU activation and an output of a previous processing, up-sampling is performed through a deconvolution layer (deconvolutional).

[95] In one implementation, an input of the first up-sampling includes an input feature map of [128, 128, 128] of the fourth down-sampling and a feature map of [128, 128, 128] obtained by performing convolution on an output feature map of [64, 64, 256] of the fifth down-sampling, and an output feature map of the first up-sampling has a size of [256, 256, 64]; an input of the second up-sampling includes an input feature map of [256, 256, 64] of the

third down-sampling and the output feature map of [256, 256, 64] of the first up-sampling, and an output feature map of the second up-sampling has a size of [512, 512, 32]; an input of the third up-sampling includes an input feature map of [512, 512, 32] of the second down-sampling and the output feature map of [512, 512, 32] of the second up-sampling, and an output feature map of the third up-sampling has a size of [1024, 1024, 16]; an input of the fourth up-sampling includes an input feature map of [1024, 1024, 16] of the first down-sampling and the output feature map of [1024, 1024, 16] of the third up-sampling, and an output feature map of the fourth up-sampling has a size of [1024, 1024, 16].

[96] At block S143, based on a FPN network of the segmentation model, a binarization-segmentation result corresponding to the instance image is outputted according to a result of the plurality of up-sampling.

[97] As an example, based on the FPN network, prediction is performed on the feature map after up-sampling.

[98] In one implementation, as illustrated in FIG. 13, the operation of outputting, based on the FPN network of the segmentation model, the binarization-segmentation result corresponding to the instance image according to the result of the plurality of up-sampling at block S143 includes the following.

[99] At block S1431, a convolution operation is performed on an output of a first up-sampling in the Vnet network.

[100] As an example, the number of convolution kernels (filter) is 128, the size of the feature map inputted is [256, 256, 64], and the size of the feature map outputted is [256, 256, 128].

[101] At block S1432, a convolution operation is performed on an output of a second up-sampling in the Vnet network.

[102] As an example, the number of convolution kernels (filter) is 128.

[103] At block S1433, up-sampling is performed on a result of the convolution operation performed on the output of the first up-sampling, and a result of the up-sampling performed is added to a result of the convolution operation performed on the output of the second up-sampling.

[104] As an example, the result of the convolution operation performed on the output [512, 512, 32] of the second up-sampling is [512, 512, 128], and an output of the up-sampling performed on the output [256, 256, 128] obtained at block S1431 is [512, 512, 128], and the output [512, 512, 128] is added to another output [512, 512, 128] to obtain an output with a size of [512, 512, 128].

- [105]** At block S1434, a convolution operation is performed on an output of a third up-sampling in the Vnet network.
- [106]** As an example, the number of convolution kernels (filter) is 128.
- [107]** At block S1435, up-sampling is performed on a result of the convolution operation performed on the output of the second up-sampling, and a result the up-sampling performed is added to a result of the convolution operation performed on the output of the third up-sampling.
- [108]** As an example, the result of the convolution operation performed on the output [1024, 1024, 16] of the third up-sampling is [1024, 1024, 128], an output of the up-sampling performed on the output [512, 512, 128] obtained at block S1432 is [1024, 1024, 128], and the output [1024, 1024, 128] is added to another output [1024, 1024, 128] to obtain an output with a size of [1024, 1024, 128].
- [109]** At block S1436, a convolution operation is performed on an output of a fourth up-sampling in the Vnet network.
- [110]** As an example, the number of convolution kernels (filter) is 128.
- [111]** At block S1437, a result of the convolution operation performed on the output of the third up-sampling is added to a result of the convolution operation performed on the output of the fourth up-sampling.
- [112]** As an example, the result of the convolution operation performed on the output [1024, 1024, 16] of the fourth up-sampling is [1024, 1024, 128], and the output obtained at block S1434 is [1024, 1024, 128], and the output [1024, 1024, 128] is added to another output [1024, 1024, 128] to obtain an output with a size of [1024, 1024, 128].
- [113]** At block S1438, perform 1×1 convolution on added results to obtain a binarization segmentation-result image.
- [114]** As an example, perform 1×1 convolution on output [1024, 1024, 128] obtained at block S1437, to obtain the binarization segmentation-result image.
- [115]** In some implementations, the instance image inputted in the Vnet network of the segmentation model has a same length and width with a result (i.e., a segmentation-result image) of binarization segmentation.
- [116]** In some implementations, if the segmentation model predicts that an instance image corresponding to a binarization-segmentation result has a region of interest, determine that a region of the binarization-segmentation result corresponding to the region of interest is a first-value region. If the segmentation model predicts that an instance image corresponding to a binarization-segmentation result has a region of uninterested, determine that a region of

the binarization-segmentation result corresponding to the region of uninterested is a second-value region.

[117] As an example, if the segmentation model detects that an instance image corresponding to a binarization-segmentation result outputted by the segmentation model has the region of interest (e.g., a cancerous region), determine that a region of the binarization-segmentation result corresponding to the region of interest is the first-value region, and determine that a region of the binarization-segmentation result corresponding to the region of uninterested is the second-value region.

[118] As an example, a first value is 255, and in a grayscale image or a black-and-white image, the first-value region is a white region; a second value is 0, and in a grayscale image or a black-and-white image, the second-value region is a black region.

[119] The region of interest is a cancerous region for example. If the Vnet network-FPN network-based segmentation model predicts that the instance image has the region of interest, determine that a region of a binarization-segmentation result (outputted by the segmentation model) corresponding to the region of interest is a positive region, and determine that a region of the binarization-segmentation result (outputted by the segmentation model) not corresponding to the region of interest is a negative region.

[120] At block S150, binarization-segmentation results corresponding to the instance images are connected to obtain a segmentation-result image.

[121] As an example, multiple instance images are obtained after preprocessing the image to-be-annotated of FIG. 3. The Vnet network-FPN network-based segmentation model performs binarization segmentation on each of the instance images. Binarization-segmentation results of the instance images corresponding to an image in FIG. 3 are illustrated in FIG. 14. A black region in FIG. 14 corresponds to a second-value region in the binarization-segmentation result, where the second-value region is not a region of interest (in other words, the black region refers to a region of uninterested). A white region in FIG. 14 corresponds to a first-value region in the binarization-segmentation result, where the first-value region is a region of interest.

[122] FIG. 15 illustrates a segmentation-result image obtained by connecting binarization-segmentation results corresponding to the instance images.

[123] At block S160, a binarization fusion image is calculated according to the classification-result image and the segmentation-result image.

[124] By combining classification results predicted by the DenseNet network-based classification model and segmentation results predicted by the Vnet network-FPN network-

based segmentation model, effect of prediction of a region of interest in the image to-be-annotated can be improved.

[125] In some implementations, as illustrated in FIG. 16, the operation of calculating the binarization fusion image according to the classification-result image and the segmentation-result image at block S160 includes the following.

[126] At block S161, take intersection of the classification-result image and the segmentation-result image to obtain the binarization fusion image. As an example, if a position in the classification-result image corresponds to a region of interest and the position in the segmentation-result image corresponds to the region of interest, the position in the binarization fusion image corresponds to the region of interest.

[127] In some implementations, as illustrated in FIG. 17, the operation of taking the intersection of the classification-result image and the segmentation-result image to obtain the binarization fusion image at block S161 includes the following.

[128] At block S1611, when a pixel in the classification-result image has a first value and the pixel in the segmentation-result image has the first value, the pixel in the fusion image is set to be the first value.

[129] As an example, if a position in the classification-result image corresponds to white and the position in the segmentation-result image corresponds to white, the position in the binarization fusion image corresponds to white.

[130] At block S1612: when a pixel in the classification-result image has a second value, the pixel in the fusion image is set to be the second value.

[131] At block S1613: when a pixel in the segmentation-result image has the second value, the pixel in the fusion image is set to be the second value.

[132] As an example, if a position in the classification-result image corresponds to black or if the position in the segmentation-result image corresponds to black, the position in the binarization fusion image corresponds to black.

[133] In some implementations, as illustrated in FIG. 18, before taking the intersection of the classification-result image and the segmentation-result image to obtain the binarization fusion image at block S161, the multi-model fusion based method for image annotation further includes the following. At block S101, the first-value block in the classification-result image is dilated.

[134] In the implementation, the operation of taking the intersection of the classification-result image and the segmentation-result image to obtain the binarization fusion image at block S161 includes the following. At block S1601, take intersection of the dilated

classification-result image and the segmentation-result image to obtain the binarization fusion image.

[135] Dilation can smooth an edge of the first-value block in the classification-result image, so that a boundary between the first value and the second value in the fusion image obtained by taking the intersection can be smoothed.

[136] At block S170, a contour of the fusion image is extracted, and a region of interest is annotated in the image to-be-annotated according to the contour.

[137] A range defined by the boundary between the first value and the second value in the fusion image (i.e., a boundary of the intersection) can be used as a range of prediction of obtaining the region of interest.

[138] As an example, for a fusion image generated according to prediction, all contour points of a region of interest in the fusion image are extracted to form a contour of the fusion image. An .xml file is generated based on the contour points, the file is named with the name of the image to-be-annotated, and the extension of the file is .ndpa, that is, the file is an annotation file corresponding to the image to-be-annotated. As illustrated in FIG. 19, the annotation file is displayed in an original image. In this way, the region of interest in the image to-be-annotated can be annotated according to the contour.

[139] In some implementations, the server in FIG. 2 or the terminal in FIG. 2 can be configured to display the annotation file corresponding to the image to-be-annotated in the original image.

[140] In other implementations, prediction is performed on each of multiple instance images according to a classification model and a segmentation model. Binarization fusion images corresponding to the instance images are calculated according to binary-classification results and binarization-segmentation results of the instance images. Then the binarization fusion images corresponding to the instance images are connected to obtain a binarization fusion image corresponding to an image to-be-annotated.

[141] In some implementations, the DenseNet network-based classification model and the Vnet network-FPN network-based segmentation model can be trained together. Training samples may include a to-be-segmented image without the region of interest, and a to-be-segmented image annotated with the region of interest. Loss function includes weightedloss and focalloss, and weightedloss and focalloss have different functions. Combined loss function can integrate different factors into a model training, which can improve Dice and intersection over union (IoU) of the model of the model and reduce false positive rate (fpr, which refers to a percentage that an image has no region of interest is determined to be

positive by false). Moreover, evaluation of training effect is based on precision rather than accuracy, thereby enhancing suppression to the false positive. In addition, introduction of the combined loss function can further optimize the training effect.

[142] By fusing a prediction result of the DenseNet network-based classification model and a prediction result of the Vnet network-FPN network-based segmentation model, almost 100% suppression of the false positive can be achieved. Moreover, a prediction speed of the model can reach 90 seconds per sheet, which significantly reduces workload of doctors. As an example, the Dice of the model is 0.88, and the IoU of the model is 0.83.

[143] The definition of Dice is as follows:

$$\text{Dice}(A, B) = 2 \frac{|A \cap B|}{|A| + |B|}$$

Where A represents a target window generated by the model, (i.e., a segmentation result, SegResult), and B represents an original annotation window (i.e., ground truth).

[144] IoU represents a ratio of intersection of the segmentation result and the annotation window to union of the segmentation result and the annotation window.

[145] According to the multi-model fusion based method for image annotation of the foregoing implementations, the DenseNet network-based classification model can perform binary classification on the instance images of the image to-be-annotated to obtain the classification-result image, and the Vnet network-FPN network-based segmentation model can perform binarization segmentation on the instance images of the image to-be-annotated to obtain the segmentation-result image. Then the classification-result image and the segmentation-result image are fused and the contour used to annotate the region of interest in the image to-be-annotated is extracted. In this way, the Dice and the IoU of the annotation can be increased, thereby reducing the false positive rate.

[146] FIG. 20 is a schematic structural diagram illustrating a multi-model fusion based device for image annotation according to an implementation. As an example, the device is configured in a server which is configured to execute the foregoing multi-model fusion based method for image annotation.

[147] As illustrated in FIG. 20, the image tagging device based on multi-model fusion includes: a preprocessing module 110, a binary-classification module 120, a classification-connecting module 130, a binarization-segmentation module 140, a segmentation-connecting module 150, a fusion module 160, and an annotation module 170.

[148] The preprocessing module 110 is configured to acquire and preprocess an image

to-be-annotated to obtain a plurality of instance images.

[149] The binary-classification module 120 is configured to input the instance images into a DenseNet network-based classification model for binary classification.

[150] In some implementations, as illustrated in FIG. 21, the binary-classification module 120 includes a first classification sub-module 121 and a second classification sub-module 122. The first classification sub-module 121 is configured to determine, for each of the instance images, that a binary-classification result corresponding to the instance image is a first-value block, when the DenseNet network-based classification model predicts that the instance image has the region of interest. The second classification sub-module 122 is configured to determine, for each of the instance images, that the binary-classification result corresponding to the instance image is a second-value block, when the DenseNet network-based classification model predicts that the instance image has no region of interest.

[151] The classification-connecting module 130 is configured to connect binary-classification results corresponding to the instance images to obtain a classification-result image.

[152] The binarization-segmentation module 140 is configured to input the instance images into a Vnet network-FPN network-based segmentation model for binarization segmentation.

[153] In some implementations, if the segmentation model predicts that an instance image corresponding to a binarization-segmentation result has a region of interest, determine that a region of the binarization-segmentation result corresponding to the region of interest is a first-value region. If the segmentation model predicts that an instance image corresponding to a binarization-segmentation result has a region of uninterested, determine that a region of the binarization-segmentation result corresponding to the region of uninterested is a second-value region.

[154] In some implementations, as illustrated in FIG. 21, the binarization-segmentation module 140 includes an input sub-module 141, a Vnet sub-module 142, and an FPN sub-module 143. The input sub-module 141 is configured to input, for each of the instance images, the instance image into a Vnet network of the segmentation model. The Vnet sub-module 142 is configured to sequentially perform, for each of the instance images, based on the Vnet network, non-linear activation, a plurality of down-sampling, and a plurality of up-sampling on the instance image. The FPN sub-module 143 is configured to output, for each of the instance images, based on a FPN network of the segmentation model, a binarization-segmentation result corresponding to the instance image according to a result of the plurality

of up-sampling.

[155] The segmentation-connecting module 150 is configured to connect binarization-segmentation results corresponding to the instance images to obtain a segmentation-result image.

[156] The fusion module 160 is configured to calculate a binarization fusion image according to the classification-result image and the segmentation-result image.

[157] In some implementations, the fusion module 160 is configured to take intersection of the classification-result image and the segmentation-result image to obtain the binarization fusion image.

[158] In some implementations, as illustrated in FIG. 21, the fusion module 160 includes a first fusion sub-module 161, a second fusion sub-module 162, and a third fusion sub-module 163. The first fusion sub-module 161 is configured to set a pixel in the fusion image to be a first value, when the pixel in the classification-result image has the first value and the pixel in the segmentation-result image has the first value. The second fusion sub-module 162 is configured to set a pixel in the fusion image to be a second value, when the pixel in the classification-result image has the second value. The third fusion sub-module 163 is configured to set a pixel in the fusion image to be the second value, when the pixel in the segmentation-result image has the second value.

[159] The annotation module 170 is configured to extract a contour of the fusion image, and annotate a region of interest in the image to-be-annotated according to the contour.

[160] In some implementations, as illustrated in FIG. 21, the multi-model fusion based device for image annotation further includes a dilating module 101. The dilating module 101 is configured to dilate the first-value block in the classification-result image. The fusion module 160 is configured to take intersection of the dilated classification-result image and the segmentation-result image to obtain the binarization fusion image.

[161] It should be noted that, those skilled in the art can clearly understand that for the convenience and conciseness of description, for detailed working processes of the foregoing devices, modules, and units, reference may be made to corresponding processes in the foregoing method implementations, which will not be repeated herein.

[162] The method and device of the disclosure can be used in many general-purpose or special-purpose computing system environments or configurations. For example, personal computers, server computers, handheld or portable devices, tablet devices, multi-processor systems, microprocessor-based systems, programmable consumer electronic devices, network PCs, small computers, large computers, distributed computing environments including any of

the above systems or devices, and so on.

[163] In some implementations, the foregoing method and device may be implemented in a form of computer programs. The computer programs may be run on a computer device in FIG. 22.

[164] FIG. 22 is a schematic structural diagram illustrating a computer device according to some implementations. The computer device may be a server or a terminal.

[165] As illustrated in FIG. 22, the computer device includes a processor, a memory, and a network interface coupled with each other via a system bus. The memory includes a non-transitory storage medium and a memory storage.

[166] The non-transitory storage medium can store an operating system and computer programs. The computer programs include program instructions which, when executed by a processor, cause the processor to execute any multi-model fusion based method for image annotation.

[167] The processor is configured to provide computing and control capabilities, and support operations of the entire computer device.

[168] The memory storage is configured to provide environments for operations of the computer programs in the non-transitory storage medium. When the computer programs are executed by the processor, the processor can execute any image annotation method based on multi-model fusion.

[169] The network interface is configured for network communication, such as sending assigned tasks. Those skilled in the art should understand that, a structure of the computer device in a block diagram is only a part of a structure related to technical solutions of the disclosure, and does not constitute a limitation on the computer device to which the technical solutions of the disclosure is applied. The specific computer device may include more or fewer components than the diagram, or combine some components, or have different component arrangements.

[170] It should be understood that, the processor may be a central processing unit (CPU). The processor may also be other general-purpose processors, digital signal processors (DSP), application specific integrated circuits (ASIC), field-programmable gate array (FPGA) or other programmable logic devices, discrete or transistor logic devices, discrete hardware components, etc. The general-purpose processor may be a microprocessor, or the processor may be any conventional processor or the like.

[171] In one implementation, the processor is configured to run the computer programs stored in the memory to implement the following operations. An image to-be-annotated is

acquired and preprocessed to obtain a plurality of instance images. The instance images are inputted into a DenseNet network-based classification model for binary classification. Binary-classification results corresponding to the instance images are connected to obtain a classification-result image. The instance images are inputted into a Vnet network-FPN network-based segmentation model for binarization segmentation. Binarization-segmentation results corresponding to the instance images are connected to obtain a segmentation-result image. A binarization fusion image is calculated according to the classification-result image and the segmentation-result image. A contour of the fusion image is extracted, and a region of interest is annotated in the image to-be-annotated according to the contour.

[172] In one implementation, the processor configured to input the instance images into the DenseNet network-based classification model for binary classification is configured to: for each of the instance images, determine that a binary-classification result corresponding to the instance image is a first-value block, when the DenseNet network-based classification model predicts that the instance image has the region of interest; or determine that the binary-classification result corresponding to the instance image is a second-value block, when the DenseNet network-based classification model predicts that the instance image has no region of interest.

[173] In one implementation, the processor configured to input the instance images into the Vnet network-FPN network-based segmentation model for binarization segmentation is configured to: for each of the instance images, input the instance image into a Vnet network of the segmentation model; perform, based on the Vnet network, non-linear activation, a plurality of down-sampling, and a plurality of up-sampling on the instance image sequentially; and output, based on a FPN network of the segmentation model, a binarization-segmentation result corresponding to the instance image according to a result of the plurality of up-sampling.

[174] In one implementation, the processor configured to calculate the binarization fusion image according to the classification-result image and the segmentation-result image is configured to take intersection of the classification-result image and the segmentation-result image to obtain the binarization fusion image.

[175] In one implementation, the processor is further configured to dilate the first-value block in the classification-result image, before taking the intersection of the classification-result image and the segmentation-result image. The processor configured to take the intersection of the classification-result image and the segmentation-result image to obtain the binarization fusion image is configured to take intersection of the dilated classification-result

image and the segmentation-result image to obtain the binarization fusion image.

[176] According to the description of the foregoing implementations, those skilled in the art can clearly understand that technical solutions of the disclosure can be implemented by means of software plus a necessary general-purpose hardware platform. Based on this understanding, the technical solutions of the disclosure essentially or the part that contributes to the existing technology can be embodied in the form of a software product. The computer software product can be stored in a storage medium, such as ROM/RAM, magnetic disk, CD-ROM, and so on. The computer software product includes multiple instructions which are operable with a computer device (a personal computer, server, network device, or the like) to execute some or part of the operations of the method described in implementations of the disclosure, for example, a computer readable storage medium. The computer readable storage medium is configured to store computer programs. The computer programs include program instructions. The processor is configured to execute the program instructions to implement any of the multi-model fusion based methods for image annotation of the implementations of the disclosure.

[177] The computer readable storage medium may be an internal storage unit of the computer device described in the foregoing implementations, such as a hard disk or memory of the computer device. The computer readable storage medium may also be an external storage device of the computer device, such as a plug-in hard disk, a smart media card (SMC), and a secure digital (SD) card, flash card, or the like equipped on the computer device.

[178] The foregoing implementations are merely illustrative examples, and it is to be understood that the disclosure is not to be limited to the disclosed implementations. As will occur to those skilled in the art, any modifications or equivalent substitutions made thereto without departing from the spirits and principles of the disclosure shall all be encompassed within the protection of the disclosure. Therefore, the scope of the disclosure should be determined by the scope of the claims.

WHAT IS CLAIMED IS:

1. A computer-implemented multi-model fusion based method for image annotation, the method being executed by a processor and comprising:

acquiring and preprocessing an image to-be-annotated to obtain a plurality of instance images;

inputting the instance images into a DenseNet network-based classification model for binary classification;

connecting binary-classification results corresponding to the instance images to obtain a classification-result image;

inputting the instance images into a Vnet network-FPN network-based segmentation model for binarization segmentation;

connecting binarization-segmentation results corresponding to the instance images to obtain a segmentation-result image;

calculating a binarization fusion image according to the classification-result image and the segmentation-result image; and

extracting a contour of the fusion image, and annotating a region of interest in the image to-be-annotated according to the contour.

2. The multi-model fusion based method for image annotation of claim 1, wherein inputting the instance images into the DenseNet network-based classification model for binary classification comprises:

for each of the instance images:

determining that a binary-classification result corresponding to the instance image is a first-value block, when the DenseNet network-based classification model predicts that the instance image has the region of interest; or

determining that the binary-classification result corresponding to the instance image is a second-value block, when the DenseNet network-based classification model predicts that the instance image has no region of interest.

3. The multi-model fusion based method for image annotation of claim 2, wherein inputting the instance images into the Vnet network-FPN network-based segmentation model for binarization segmentation comprises:

for each of the instance images:

inputting the instance image into a Vnet network of the segmentation model;
performing, based on the Vnet network, non-linear activation, a plurality of
down-sampling, and a plurality of up-sampling on the instance image sequentially;
and

outputting, based on a FPN network of the segmentation model, a
binarization-segmentation result corresponding to the instance image according to a
result of the plurality of up-sampling.

4. The multi-model fusion based method for image annotation of claim 3, wherein
for each of the instance images:

determining that a region of a binarization-segmentation result corresponding
to the region of interest is a first-value region, when the segmentation model predicts
that the instance image corresponding to the binarization-segmentation result has the
region of interest; or

determining that a region of the binarization-segmentation result
corresponding to a region of uninterested is a second-value region, when the
segmentation model predicts that the instance image corresponding to the
binarization-segmentation result has the region of uninterested.

5. The multi-model fusion based method for image annotation of claim 4, wherein
calculating the binarization fusion image according to the classification-result image and the
segmentation-result image comprises:

taking intersection of the classification-result image and the segmentation-result
image to obtain the binarization fusion image.

6. The multi-model fusion based method for image annotation of claim 5, wherein
taking the intersection of the classification-result image and the segmentation-result image to
obtain the binarization fusion image comprises:

setting a pixel in the fusion image to be a first value, when the pixel in the
classification-result image has the first value and the pixel in the segmentation-result image
has the first value;

setting a pixel in the fusion image to be a second value, when the pixel in the
classification-result image has the second value; and

setting a pixel in the fusion image to be the second value, when the pixel in the

segmentation-result image has the second value.

7. The multi-model fusion based method for image annotation of claim 5, further comprising:

dilating the first-value block in the classification-result image, before taking the intersection of the classification-result image and the segmentation-result image; and
taking the intersection of the classification-result image and the segmentation-result image to obtain the binarization fusion image comprising:

taking intersection of the dilated classification-result image and the segmentation-result image to obtain the binarization fusion image.

8. A multi-model fusion based device for image annotation, comprising:

a preprocessing module, configured to acquire and preprocess an image to-be-annotated to obtain a plurality of instance images;

a binary-classification module, configured to input the instance images into a DenseNet network-based classification model for binary classification;

a classification-connecting module, configured to connect binary-classification results corresponding to the instance images to obtain a classification-result image;

a binarization-segmentation module, configured to input the instance images into a Vnet network-FPN network-based segmentation model for binarization segmentation;

a segmentation-connecting module, configured to connect binarization-segmentation results corresponding to the instance images to obtain a segmentation-result image;

a fusion module, configured to calculate a binarization fusion image according to the classification-result image and the segmentation-result image; and

an annotation module, configured to extract a contour of the fusion image and annotate a region of interest in the image to-be-annotated according to the contour.

9. A computer device, comprising:

a memory, configured to store computer programs; and

a processor, configured to execute the computer programs which, when executed by the processor, cause the processor to:

acquire and preprocess an image to-be-annotated to obtain a plurality of instance images;

input the instance images into a DenseNet network-based classification model

for binary classification;

connect binary-classification results corresponding to the instance images to obtain a classification-result image;

input the instance images into a Vnet network-FPN network-based segmentation model for binarization segmentation;

connect binarization-segmentation results corresponding to the instance images to obtain a segmentation-result image;

calculate a binarization fusion image according to the classification-result image and the segmentation-result image; and

extract a contour of the fusion image, and annotate a region of interest in the image to-be-annotated according to the contour.

10. The computer device of claim 9, wherein the processor configured to input the instance images into the DenseNet network-based classification model for binary classification is configured to:

for each of the instance images:

determine that a binary-classification result corresponding to the instance image is a first-value block, when the DenseNet network-based classification model predicts that the instance image has the region of interest; or

determine that the binary-classification result corresponding to the instance image is a second-value block, when the DenseNet network-based classification model predicts that the instance image has no region of interest.

11. The computer device of claim 10, wherein the processor configured to input the instance images into the Vnet network-FPN network-based segmentation model for binarization segmentation is configured to:

for each of the instance images:

input the instance image into a Vnet network of the segmentation model;

perform, based on the Vnet network, non-linear activation, a plurality of down-sampling, and a plurality of up-sampling on the instance image sequentially; and

output, based on a FPN network of the segmentation model, a binarization-segmentation result corresponding to the instance image according to a result of the plurality of up-sampling.

determine that a region of the binarization-segmentation result corresponding to the region of interest is a first-value region, when the segmentation model predicts that the instance image corresponding to the binarization-segmentation result has the region of interest; or

determine that a region of the binarization-segmentation result corresponding to a region of uninterested is a second-value region, when the segmentation model predicts that the instance image corresponding to the binarization-segmentation result has the region of uninterested.

12. The computer device of claim 11, wherein the processor configured to calculate the binarization fusion image according to the classification-result image and the segmentation-result image is configured to:

take intersection of the classification-result image and the segmentation-result image to obtain the binarization fusion image.

13. The computer device of claim 12, wherein the processor configured to take the intersection of the classification-result image and the segmentation-result image to obtain the binarization fusion image is configured to:

set a pixel in the fusion image to be a first value, when the pixel in the classification-result image has the first value and the pixel in the segmentation-result image has the first value;

set a pixel in the fusion image to be a second value, when the pixel in the classification-result image has the second value; and

set a pixel in the fusion image to be the second value, when the pixel in the segmentation-result image has the second value.

14. The computer device of claim 12, wherein the processor is further configured to:

dilate the first-value block in the classification-result image, before taking the intersection of the classification-result image and the segmentation-result image; and the processor configured to take the intersection of the classification-result image and the segmentation-result image to obtain the binarization fusion image is configured to:

take intersection of the dilated classification-result image and the segmentation-result image to obtain the binarization fusion image.

15. A computer readable storage medium, configured to store computer programs which, when executed by a processor, cause the processor to carry out actions, comprising:

- acquiring and preprocessing an image to-be-annotated to obtain a plurality of instance images;
- inputting the instance images into a DenseNet network-based classification model for binary classification;
- connecting binary-classification results corresponding to the instance images to obtain a classification-result image;
- inputting the instance images into a Vnet network-FPN network-based segmentation model for binarization segmentation;
- connecting binarization-segmentation results corresponding to the instance images to obtain a segmentation-result image;
- calculating a binarization fusion image according to the classification-result image and the segmentation-result image; and
- extracting a contour of the fusion image, and annotating a region of interest in the image to-be-annotated according to the contour.

16. The computer readable storage medium of claim 15, wherein the processor configured to carry out the action of inputting the instance images into the DenseNet network-based classification model for binary classification is configured to carry out actions, comprising:

- for each of the instance images:
 - determining that a binary-classification result corresponding to the instance image is a first-value block, when the DenseNet network-based classification model predicts that the instance image has the region of interest; or
 - determining that the binary-classification result corresponding to the instance image is a second-value block, when the DenseNet network-based classification model predicts that the instance image has no region of interest.

17. The computer readable storage medium of claim 16, wherein the processor configured to carry out the action of inputting the instance images into the Vnet network-FPN network-based segmentation model for binarization segmentation is configured to carry out actions, comprising:

- for each of the instance images:

inputting the instance image into a Vnet network of the segmentation model;
performing, based on the Vnet network, non-linear activation, a plurality of
down-sampling, and a plurality of up-sampling on the instance image sequentially;
and

outputting, based on a FPN network of the segmentation model, a
binarization-segmentation result corresponding to the instance image according to a
result of the plurality of up-sampling.

determining that a region of the binarization-segmentation result
corresponding to the region of interest is a first-value region, when the segmentation
model predicts that the instance image corresponding to the binarization-segmentation
result has the region of interest; or

determining that a region of the binarization-segmentation result
corresponding to a region of uninterested is a second-value region, when the
segmentation model predicts that the instance image corresponding to the
binarization-segmentation result has the region of uninterested.

18. The computer readable storage medium of claim 17, wherein the processor configured
to carry out the action of calculating the binarization fusion image according to the
classification-result image and the segmentation-result image is configured to carry out
actions, comprising:

taking intersection of the classification-result image and the segmentation-result
image to obtain the binarization fusion image.

19. The computer readable storage medium of claim 18, wherein the processor configured
to carry out the action of taking the intersection of the classification-result image and the
segmentation-result image to obtain the binarization fusion image is configured to carry out
actions, comprising:

setting a pixel in the fusion image to be a first value, when the pixel in the
classification-result image has the first value and the pixel in the segmentation-result image
has the first value;

setting a pixel in the fusion image to be a second value, when the pixel in the
classification-result image has the second value; and

setting a pixel in the fusion image to be the second value, when the pixel in the
segmentation-result image has the second value.

20. The computer readable storage medium of claim 18, wherein the processor is further configured to carry out actions, comprising:
- dilating the first-value block in the classification-result image, before taking the intersection of the classification-result image and the segmentation-result image.
 - the processor configured to carry out the action of taking the intersection of the classification-result image and the segmentation-result image to obtain the binarization fusion image is configured to carry out actions, comprising:
 - taking intersection of the dilated classification-result image and the segmentation-result image to obtain the binarization fusion image.