

INSTANCE SEGMENTATION METHODS AND APPARATUSES, ELECTRONIC DEVICES, PROGRAMS, AND MEDIA

5

CROSS-REFERENCE TO RELATED APPLICATIONS

The present application claims the priority of the Chinese patent application with the application No. CN2018101370447, submitted to Patent Office of the People's Republic of China on 09 February 2018 and entitled "INSTANCE SEGMENTATION METHODS AND APPARATUSES, ELECTRONIC DEVICES, PROGRAMS, AND MEDIA" and the priority of the Chinese patent application with the application No. CN2018101363710, submitted to Patent Office of the People's Republic of China on 09 February 2018, and entitled "IMAGE SEGMENTATION METHODS AND APPARATUSES, ELECTRONIC DEVICES, PROGRAMS, AND MEDIA", which are hereby incorporated by reference in their entirety.

TECHNICAL FIELD

20

The present disclosure relates to computer vision technologies, and in particular, to instance segmentation methods and apparatuses, electronic devices, programs, and media.

25

BACKGROUND

Instance segmentation is an important direction in the field of computer vision; this task combines characteristics of both semantic segmentation and object detection; an independent pixel-level mask is generated for each object of an input image, respectively, and a class corresponding thereto is predicted. The instance segmentation has wide applications in fields of autonomous driving, domestic robots, etc.

35

SUMMARY

Embodiments of the present disclosure provide an instance segmentation solution.

According to an aspect of embodiments of the present disclosure, an instance
5 segmentation method is provided, including:

performing feature extraction on an image via a neural network to output features at at least two different hierarchies;

extracting region features corresponding to at least one instance candidate region in the image from the features at the at least two different hierarchies, and
10 fusing region features corresponding to a same instance candidate region, to obtain a first fusion feature of each instance candidate region; and

performing instance segmentation based on each first fusion feature, to obtain at least one of an instance segmentation result of a corresponding instance candidate region or an instance segmentation result of the image.

15 According to an aspect of embodiments of the present disclosure, an instance segmentation apparatus is provided, including:

a neural network, configured to perform feature extraction on an image to output features at at least two different hierarchies;

an extraction module, configured to extract region features corresponding to at
20 least one instance candidate region in the image from the features at the at least two different hierarchies;

a first fusion module, configured to fuse region features corresponding to a same instance candidate region, to obtain a first fusion feature of each instance candidate region; and

25 a segmentation module, configured to perform instance segmentation based on each first fusion feature, to obtain at least one of an instance segmentation result of the corresponding instance candidate region or an instance segmentation result of the image.

According to another aspect of embodiments of the present disclosure, an
30 electronic device is provided, including:

a memory, configured to store a computer program; and

a processor, configured to execute the computer program stored in the memory, where when the computer program is executed, the method according to any one of

the embodiments of the present disclosure is implemented.

According to a further aspect of the embodiments of the present application, a computer-readable storage medium is provided. A computer program is stored thereon, and when the computer program is executed by a processor, the method according to any one of the embodiments of the present disclosure is implemented.

According to another aspect of the embodiments of the present disclosure, a computer program is provided, including computer instructions, where when the computer instructions are operated in a processor of a device, the method according to any one of the embodiments of the present disclosure is implemented.

Based on the instance segmentation methods and apparatuses, electronic devices, programs, and media provided by the embodiments of the present disclosure, the method includes: performing feature extraction on an image via a neural network to output features at at least two different hierarchies; extracting region features corresponding to at least one instance candidate region in the image from the features at the two different hierarchies, and fusing region features corresponding to a same instance candidate region, to obtain a first fusion feature of each instance candidate region; and performing instance segmentation based on each first fusion feature, to obtain at least one of an instance segmentation result of the corresponding instance candidate region or an instance segmentation result of the image. Embodiments of the present disclosure design a technical solution of instance segmentation based on a deep-learning framework; since the deep-learning has a strong modeling capability, it facilitates to obtain a better instance segmentation result. In addition, performing instance segmentation on the instance candidate region, with respect to directly performing instance segmentation on an entire image, can improve accuracy of instance segmentation, reduce the computing amount and complexity required for the instance segmentation, and improve the instance segmentation efficiency. Moreover, extracting region features corresponding to the instance candidate region from the features at the at least two different hierarchies for fusing and performing the instance segmentation based on the obtained fusion feature enable each instance candidate region to obtain information at different hierarchies at the same time; since the information extracted from the features at different hierarchies are all at different semantic hierarchies, accuracy of the instance segmentation result of each instance candidate region can be improved via context information.

By means of the accompanying drawings and embodiments, the technical solutions of the present disclosure are further described in details.

5

BRIEF DESCRIPTION OF THE DRAWINGS

The drawings constituting a part of the specification describe embodiments of the present disclosure, and are used for explaining the principles of the present disclosure in combination of the description.

10

With reference to the accompanying drawings, according to the detailed description below, the present disclosure can be understood more clearly, where:

FIG. 1 is a flow chart of an embodiment of an instance segmentation method of the present disclosure;

15

FIG. 2 is a feature fusion schematic diagram in embodiments of the present disclosure;

FIG. 3 is a flow chart of another embodiment of the instance segmentation method of the present disclosure;

FIG. 4 is a network structure schematic diagram for two-path mask prediction executed in embodiments of the present disclosure;

20

FIG. 5 is a flow chart of an application embodiment of the instance segmentation method of the present disclosure;

FIG. 6 is a process schematic diagram of the application embodiment shown in FIG. 5;

25

FIG. 7 is a structural schematic diagram of an embodiment of an instance segmentation apparatus of the present disclosure;

FIG. 8 is a structural schematic diagram of another embodiment of the instance segmentation apparatus of the present disclosure;

FIG. 9 is a structural schematic diagram of an embodiment of a segmentation module in the embodiments of the present disclosure; and

30

FIG. 10 is a structural schematic diagram of an embodiment of an electronic device in the embodiments of the present disclosure.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

35

Each exemplary embodiment of the present disclosure is described in detail with reference to the accompany drawings now. It should be noted that: unless otherwise stated specifically, relative arrangement of the components and steps, the numerical expressions, and the values set forth in the embodiments are not intended to
5 limit the scope of the present disclosure.

It should be further understood that in the embodiments of the present disclosure, "a plurality of" may refer to two or more, and "at least one" may refer to one, two, or more.

Persons skilled in the art may understand that terms "first", "second", etc. in
10 the embodiments of the present disclosure are only used for distinguishing different steps, devices, or modules, and do not represent any special technical meanings, and likewise do not represent necessary logic orders therebetween.

It should be further understood that any component, data, or structure mentioned in the embodiments of the present disclosure should be generally
15 understood as one or more under the condition that no explicit definition is provided or no opposite motivation is provided in the context.

It should be further understood that the description of each embodiment of the present disclosure emphasizes differences between the embodiments, and the same or similar points therebetween may be used as reference and omitted for clarity.

20 In addition, it should be understood that, for ease of description, the size of each part shown in the accompanying drawings is not drawn in actual proportion.

The following descriptions of at least one exemplary embodiment are merely illustrative actually, and are not intended to limit the present disclosure and the applications or uses thereof.

25 Technologies, methods and devices known to a person of ordinary skill in the related art may not be discussed in detail, but such technologies, methods and devices should be considered as a part of the specification in appropriate situations.

It should be noted that similar reference numerals and letters in the following accompanying drawings represent similar items. Therefore, once an item is defined in
30 an accompanying drawing, the item does not need to be further discussed in the subsequent accompanying drawings.

In addition, the term "and/or" in the present disclosure only describes an association relation between associated objects, indicating that three relations may

exist, for example, A and/or B may indicate three conditions, i.e., A exists separately, A and B exist at the same time, and B exists separately. In addition, the character "/" in the present disclosure generally represents the preceding and latter associated objects are in an "or" relation.

5 The embodiments of the present disclosure may be applied to electronic devices such as terminal devices, computer systems, servers, which may operate with numerous other general-purpose or special-purpose computing system environments or configurations. Examples of well-known computing systems, environments, and/or configurations suitable for use together with the computer systems/servers include,
10 but are not limited to, personal computer systems, server computer systems, thin clients, thick clients, handheld or laptop devices, microprocessor-based systems, set top boxes, programmable consumer electronics, network personal computers, small computer systems, large computer systems, distributed cloud computing environments that include any one of the foregoing systems.

15 The electronic devices such as terminal devices, computer systems, and servers may be described in the general context of computer system executable instructions (for example, program modules) executed by the computer system. Generally, the program modules may include routines, programs, target programs, components, logics, and data structures, to execute specific tasks or implement
20 specific abstract data types. The computer systems/servers may be practiced in the distributed cloud computing environments in which tasks are executed by remote processing devices that are linked through a communications network. In the distributed computing environments, program modules may be located in local or remote computing system storage media including storage devices.

25 FIG. 1 is a flow chart of an embodiment of an instance segmentation method of the present disclosure. As shown in FIG. 1, the instance segmentation method of the embodiment includes operations 102 to 108.

 In operation 102, feature extraction is performed on an image via a neural network to output features at at least two different hierarchies.

30 Representing forms of the features in the embodiments of the disclosure may include, but not limited to, for instance: a feature diagram, a feature vector, a feature matrix, etc. The at least two different hierarchies refer to two or more network layers located at different depth of the neural network in the neural network. The image may

include, but not limited to, for instance: a static image, a frame image in a video, etc.

In an optional example, operation 102 may be executed by invoking, by a processor, corresponding instructions stored in a memory, and may also be executed by the neural network run by the processor.

5 In operation 104, region features corresponding to at least one instance candidate region in the image are extracted from the features at the at least two different hierarchies.

 The instance may include, but not limited to, for example, a certain specific object, for example, a certain specific person, a certain specific article, etc. One or
10 more instance candidate regions may be obtained by detecting the image via the neural network. The instance candidate region represents a region in the image where the instance may appear.

 In an optional example, operation 104 may be executed by a processor by invoking a corresponding instruction stored in a memory, or may be executed by an
15 extraction module run by the processor.

 In operation 106, region features corresponding to a same instance candidate region is respectively fused, to obtain a first fusion feature of each instance candidate region.

 In embodiments of the preset disclosure, the mode for fusing multiple region
20 features may be, for example, element-wise summing, element-wise max, element-wise averaging, etc.

 In an optional example, the operation 106 is performed by a processor by invoking a corresponding instruction stored in a memory, or is performed by a first fusion module run by the processor.

25 In operation 108, instance segmentation is respectively performed based on each first fusion feature, to obtain at least one of an instance segmentation result of the corresponding instance candidate region or an instance segmentation result of the image.

 In the embodiments of the present disclosure, the instance segmentation result
30 of the instance candidate region may include: elements belonging to a certain instance in the instance candidate region and the class to which the instance belongs, for example, the element belongs to a certain boy in the instance candidate region and the class to which the boy belongs is human.

In an optional example, operation 108 may be executed by a processor by invoking a corresponding instruction stored in a memory, or may be executed by a segmentation module run by the processor.

Based on the instance segmentation methods provided by the embodiments of the present disclosure, the method includes: performing feature extraction on an image via a neural network to output features at at least two different hierarchies; extracting region features corresponding to at least one instance candidate region in the image from the features at the two different hierarchies, and fusing region features corresponding to a same instance candidate region, to obtain a first fusion feature of each instance candidate region; and performing instance segmentation based on each first fusion feature, to obtain at least one of an instance segmentation result of the corresponding instance candidate region or an instance segmentation result of the image. Embodiments of the present disclosure design solving the problem of instance segmentation based on a deep-learning framework; since the deep-learning has a strong modeling capability, it facilitates to obtain a better instance segmentation result. In addition, performing instance segmentation on the instance candidate region, with respect to directly performing instance segmentation on an entire image, can improve accuracy of instance segmentation, reduce the computing amount and complexity required for the instance segmentation, and improve the instance segmentation efficiency. Moreover, extracting region features corresponding to the instance candidate region from the features at the at least two different hierarchies for fusing and performing the instance segmentation based on the obtained fusion feature enable each instance candidate region to obtain information at different hierarchies at the same time; since the information extracted from the features at different hierarchies are all at different semantic hierarchies, accuracy of the instance segmentation result of each instance candidate region can be improved via context information.

In an embodiment of the instance segmentation method embodiments of the present disclosure, operation 102 of the performing feature extraction on the image via the neural network to output the features at the at least two different hierarchies may include: performing feature extraction on the image via the neural network to output the features at the at least two different hierarchies by means of at least two network layers with different network depths in the neural network.

In the embodiments of the present disclosure, the neural network includes two

or more network layers with different network depths; in the network layers included by the neural network, the network layer used for feature extraction may be referred to as a feature layer; after receiving an image by the neural network, feature extraction is performed on the input image by means of a first network layer and the extracted
5 feature is input to a second network layer; starting from the second network layer, feature extraction is performed on the input feature on each network layer in sequence, and the extracted feature is input to a next network layer for feature extraction. The network depth of each network layer in the neural network is from shallow to deep in an order of input and output or in an order of feature extraction; the hierarchies for the
10 features output after feature extraction on each network layer in sequence is from low to high, and resolution ratios thereof are from high to low. As compared with the network layer with a shallower network depth in the same neural network, the network layer with a deeper network depth has a wider vision domain and more structural information of a concerned space, and when the extracted feature is used for
15 instance segmentation, the segmentation result may be more accurate. In the neural network, the network layer generally may include: at least one convolutional layer for feature extraction and an upsampling layer for upsampling the feature extracted from the convolutional layer (e.g., a feature diagram); by upsampling the feature, the size of the feature extracted from the convolutional layer (e.g., the feature diagram) may
20 be reduced.

In an embodiment of the instance segmentation method embodiments of the present disclosure, operation 106 of the respectively fusing region features corresponding to a same instance candidate region may include: respectively perform pixel-level fusion on a plurality of region features corresponding to the same instance
25 candidate region.

For example, in an optional example, the respectively performing pixel-level fusion on a plurality of region features corresponding to the same instance candidate region may be:

respectively performing element-wise max on the plurality of region features
30 corresponding to the same instance candidate region, i.e., taking the maximum value of the features of the element positions in the plurality of region features corresponding to the same instance candidate region;

or respectively performing element-wise averaging on the plurality of region

features corresponding to the same instance candidate region, i.e., taking the averaging value of the features of the element positions in the plurality of region features corresponding to the same instance candidate region;

or respectively performing element-wise summing on the plurality of region features corresponding to the same instance candidate region, i.e., taking the summing value of the features of the element positions in the plurality of region features corresponding to the same instance candidate region.

In the embodiments, when performing pixel-level fusion on the plurality of region features corresponding to the same instance candidate region, as compared with other approaches, the approach of performing element-wise max on the plurality of region features corresponding to the same instance candidate region enables the feature of the instance candidate region more obvious, so that the instance segmentation is more accurate, so as to improve the accuracy of the instance segmentation result.

Optionally, in another embodiment of the instance segmentation method of the present disclosure, before respectively fusing the region features corresponding to the same instance candidate region, the region features corresponding to the same instance candidate region may be adjusted by means of a network layer, for example, a fully convolutional layer or a full connection layer, for example, adjusting a dimension of each region feature corresponding to the same instance candidate region that participates in fusion, and adapting each region feature corresponding to the same instance candidate region that participates in fusion, so that each region feature corresponding to the same instance candidate region that participates in fusion is more suitable for fusion, thereby obtaining a more accurate fusion feature.

In another embodiment of the instance segmentation method of the present disclosure, operation 102, after outputting features at the at least two different hierarchies, may further include, performing at least one retracing fusion on the features at the at least two different hierarchies to obtain a second fusion feature. The one retracing fusion includes: fusing the features at different hierarchies output by the network layers with different network depths respectively based on a network depth direction of the neural network according to two different hierarchy directions in sequence. Accordingly, in this embodiment, operation 104 may include: extracting the region features corresponding to the at least one instance candidate region from the

second fusion feature.

In an embodiment of the embodiments above, the two different hierarchy directions include: a direction from high hierarchy features to low hierarchy features and a direction from the low hierarchy features to the high hierarchy features. Hence, context information is better used for feature fusion, thereby improving the instance segmentation result of each instance candidate region.

In an optional example, the according to the two different hierarchy directions in sequence may include: along the direction from the high hierarchy features to the low hierarchy features (a direction from the feature output by the network layer with a deeper network depth to the feature output by the network layer with a shallower network depth in the neural network) and the direction from the low hierarchy features to the high hierarchy features in sequence (a direction from the feature output by the network layer with a shallower network depth to the feature output by the network layer with a deeper network depth in the neural network); or along the direction from the low hierarchy features to the high hierarchy features and the direction from the high hierarchy features to the low hierarchy features in sequence.

In one of the embodiments of the present disclosure, the fusing the features at the different hierarchies output by the network layers with the different network depths respectively along the direction from the high hierarchy features to the low hierarchy features and the direction from the low hierarchy features to the high hierarchy features in sequence includes:

along a direction from deep to shallow of the network depth of the neural network, after upsampling features at a higher hierarchy output by a network layer with a deeper network depth in the neural network in sequence, fusing the upsampled features with features at a lower hierarchy output by a network layer with a shallower network depth, for example, after upsampling the features at a higher hierarchy, fusing the upsampled features with features at a lower hierarchy, to obtain a third fusion feature. The features at a higher hierarchy may include: features output by the network layer with a deeper network depth in the neural network or features obtained by performing at least one feature extraction on the features output by the network layer with the deeper network depth. For example, in the feature participating in fusion, the features at the highest hierarchy may be the features at the highest hierarchy in the features at the at least two different hierarchies, or may also be

features obtained by performing one or more feature extractions on the features at the highest hierarchy; the third fusion feature may include the features at the highest hierarchy and the fusion feature obtained from each fusion.

Along the direction from the low hierarchy features to the high hierarchy features, after downsampling fusion features at a lower hierarchy in sequence, the downsampled fusion features is fused with fusion features at a higher hierarchy in the third fusion feature. In the fusion feature participating in this fusion, the features at the lowest hierarchy may be the fusion features at the lowest hierarchy in the third fusion features, or may also be features obtained by performing one or more feature extractions on the fusion features at the lowest hierarchy in the third fusion feature; the batch of fusion features obtained by means of feature fusion along the direction from the low hierarchy feature to the high hierarchy feature include the fusion features at the lowest hierarchy in the third fusion feature and the fusion feature obtained from each fusion.

If one retracing fusion is performed on the features at the at least two different hierarchies, the batch of fusion features obtained by means of feature fusion along the direction from the low hierarchy feature to the high hierarchy feature is the second fusion feature; if two or more retracing fusions are performed on the features at the at least two different hierarchies, an operation for fusion along the direction from the high hierarchy features to the low hierarchy features and along the direction from the low hierarchy features to the high hierarchy features can be performed for multiple times, to finally obtain a batch of fusion features, i.e., the second fusion feature.

After upsampling the features at the higher hierarchy output by the network layer with a deeper network depth, when fusing with the upsampled features with the feature at the lower hierarchy output by the network layer with a shallower network depth, after upsampling the features at the higher hierarchy output by the network layer with a deeper network depth (for example, an 80-th network layer along input and output directions of the neural network) in the neural network, the upsampled features can be in sequence fused with the feature at the lower hierarchy output by an adjacent network layer with a shallower network depth (for example, a 79-th network layer along input and output directions of the neural network). In addition, after upsampling the features at the higher hierarchy output by the network layer with the deeper network depth (for example, the 80-th network layer along input and output

directions of the neural network) in the neural network, the upsampled features can be in sequence fused with the feature at the lower hierarchy output by a network layer with a shallower network depth which is not adjacent to the network layer with the deeper network depth (for example, a 50-th network layer along input and output
5 directions of the neural network), i.e., performing over-hierarchy feature fusion.

Similarly, after downsampling the fusion features at a lower hierarchy, when the downsampled features are fused with the fusion features at the higher hierarchy in the third fusion feature, after downsampling the fusion features (e.g., P_2 , where "2" represents the feature hierarchy) at the lower hierarchy, the downsampled features can
10 also be fused with the adjacent fusion features at the higher hierarchy in the third fusion feature (e.g., P_3 , where "3" represents the feature hierarchy). Or after downsampling the fusion features at the lower hierarchy, the downsampled features can be fused with the fusion feature at the higher hierarchy in the third fusion feature which is not adjacent to the feature hierarchy (e.g., P_4 , where "4" represents the
15 feature hierarchy), i.e., performing over-hierarchy fusion feature fusion.

FIG. 2 is a feature fusion schematic diagram in embodiments of the present disclosure. As shown in FIG. 2, a schematic diagram of after downsampling the fusion feature N_i at the lower hierarchy, fusing the downsampled feature N_i with the adjacent feature P_{i+1} at the higher hierarchy to obtain a corresponding fusion feature
20 N_{i+1} is shown, where i is an integer with a value greater than 0.

Based on the embodiment, according to an order from top to bottom (i.e., an order from deep to shallow network depths in the neural network and from high hierarchy features to low hierarchy feature), the features at the high hierarchy and with low resolution ratio are gradually fused with the features at the low hierarchy and
25 with high resolution ratio to obtain a new batch of features; then according to an order from bottom to top (i.e., an order from the low hierarchy features to the high hierarchy features), after downsampling the fusion feature at the lower hierarchy in sequence, the downsampled feature fusions are fused with adjacent features at the higher hierarchy, and the features at the low hierarchy and with high resolution ratio are
30 gradually fused with the features at the high hierarchy and with low resolution ratio to obtain another new batch of features for instance segmentation. In this embodiment, an information path from bottom to top can help low hierarchy information be broadcasted to a high hierarchy network (i.e., the network layer with a deeper network

depth) more easily, and reduce loss of information broadcasting, so that information can be more smoothly transferred inside the neural network; since the low hierarchy information is more sensitive for certain details, it can provide information beneficial to positioning and segmentation, thereby improving the instance segmentation result;

5 by two feature fusions, the high hierarchy network (i.e., the network layer with a deeper network depth) is enabled to obtain the bottom hierarchy information more easily and completely, thereby further improving the instance segmentation result.

In another embodiment of the embodiments of the present disclosure, the fusing the features at the different hierarchies output by the network layers with

10 different network depths respectively along the direction from the low hierarchy features to the high hierarchy features and the direction from the high hierarchy features to the low hierarchy features in sequence includes:

along a direction from shallow to deep of the network depth of the neural network, after downsampling features at a lower hierarchy output by a network layer

15 with a shallower network depth in the neural network in sequence, fusing the downsampled features with features at a higher hierarchy output by a network layer with a deeper network depth, to obtain a fourth fusion feature, where the features at a lower hierarchy, for example, may include: features output by the network layer with a shallower network depth in the neural network or features obtained by performing at

20 least one feature extraction on the features output by the network layer with the shallower network depth. For example, in the feature participating in fusion, the features at the lowest hierarchy may be the features at the lowest hierarchy in the features at the at least two different hierarchies, or may also be features obtained by performing one or more feature extractions on the features at the lowest hierarchy; the

25 fourth fusion feature may include the features at the lowest hierarchy and the fusion feature obtained from each fusion.

Along the direction from the high hierarchy features to the low hierarchy features, after upsampling fusion features at a higher hierarchy in sequence, the upsampled fusion features is fused with fusion features at a lower hierarchy in the

30 fourth fusion feature. In the fusion feature participating in this fusion, the features at the highest hierarchy may be the fusion features at the highest hierarchy in the fourth fusion features, or may also be features obtained by performing one or more feature extractions on the fusion features at the highest hierarchy in the fourth fusion feature;

in the batch of fusion features obtained by means of feature fusion along the direction from the low hierarchy feature to the high hierarchy feature and the direction from the high hierarchy feature to the low hierarchy feature include the fusion features at the highest hierarchy in the fourth fusion feature and the fusion feature obtained from each fusion.

If one retracing fusion is performed on the features at the at least two different hierarchies, the batch of fusion features obtained by means of feature fusion along the direction from the low hierarchy feature to the high hierarchy feature and the direction from the high hierarchy feature to the low hierarchy feature is the second fusion feature; if two or more retracing fusions are performed on the features at the at least two different hierarchies, an operation for fusion along the direction from the low hierarchy features to the high hierarchy features and along the direction from the high hierarchy features to the low hierarchy features can be performed for multiple times, to finally obtain a batch of fusion features, i.e., the second fusion feature.

In an optional example, in the neural network, after downsampling features at the lower hierarchy output by the network layer with the shallower network depth, when fusing the downsampled features with the features at the higher hierarchy output by the network layer with the deeper network depth, after downsampling the features at the lower hierarchy output by the network layer with the shallower network depth in the neural network, the downsampled features can be fused with features at the higher hierarchy output by the network layer with the deeper network depth adjacent to the network layer with a shallower network depth. Or, after downsampling the features at the lower hierarchy output by the network layer with the shallower network depth in the neural network, the downsampled features can also be fused with the feature at the higher hierarchy output by a network layer with a deeper network depth which is not adjacent to the network layer with the shallower network depth, i.e., performing over-hierarchy feature fusion.

Similarly, after upsampling fusion features at the higher hierarchy, when fusing the upsampled fusion features with the fusion features at the lower hierarchy in the fourth fusion feature, after upsampling the fusion features at the higher hierarchy, the upsampled fusion features can also be fused with adjacent fusion features at the lower hierarchy in the fourth fusion feature. Or after upsampling the fusion features at the higher hierarchy, the upsampled features can also be fused with the fusion feature

at the lower hierarchy in the fourth fusion feature which is not adjacent, i.e., performing over-hierarchy fusion feature fusion.

In one of the embodiments of the present disclosure, in operation 108, the performing the instance segmentation based on each first fusion feature, to obtain at least one of the instance segmentation result of the corresponding instance candidate region or the instance segmentation result of the image may include:

based on a first fusion feature, performing instance segmentation on the instance candidate region corresponding to the first fusion feature, to obtain the instance segmentation result of the corresponding instance candidate region, where the first fusion feature is not limited to a specific first fusion feature, and may be a first fusion feature of any instance candidate region; and/or based on each first fusion feature, performing instance segmentation on the image, to obtain the instance segmentation result of the image.

In another embodiment of the embodiments of the present disclosure, in operation 108, the based on the each first fusion feature, performing the instance segmentation to obtain the instance segmentation result of the image may include: respectively performing the instance segmentation on the instance candidate region corresponding to the each first fusion feature based on the each first fusion feature, to obtain the instance segmentation result of each instance candidate region; and obtaining the instance segmentation result of the image based on the instance segmentation result of the each instance candidate region.

FIG. 3 is a flow chart of another embodiment of the instance segmentation method of the present disclosure. As shown in FIG. 3, the instance segmentation method of the embodiment includes operations 302 to 312.

In operation 302, feature extraction is performed on an image via a neural network to output features at the at least two different hierarchies by means of the network layers with the at least two different network depth in the neural network.

In an optional instance, operation 302 may be executed by invoking, by a processor, corresponding instructions stored in a memory, and may also be executed by the neural network run by the processor.

In operation 304, along a direction from deep to shallow of the network depth of the neural network, after upsampling features at a higher hierarchy output by a network layer with a deeper network depth in the neural network in sequence, the

upsampled features are fused with features at a lower hierarchy output by a network layer with a shallower network depth, to obtain a third fusion feature.

The features at a higher hierarchy may include: features output by the network layer with a deeper network depth in the neural network or features obtained by performing at least one feature extraction on the features output by the network layer with the deeper network depth. For example, in the feature participating in fusion, the features at the highest hierarchy may be the features at the highest hierarchy in the features at the at least two different hierarchies, or may also be features obtained by performing one or more feature extractions on the features at the highest hierarchy; the third fusion feature may include the features at the highest hierarchy in the features at the at least two different hierarchies and the fusion feature obtained from each fusion in operation 304.

In operation 306, along a direction from the low hierarchy features to the high hierarchy features, after downsampling fusion features at a lower hierarchy in sequence, the downsampled fusion features is fused with fusion features at a higher hierarchy in the third fusion feature, to obtain the second fusion feature.

In the fusion feature participating in this fusion, the features at the lowest hierarchy may be the fusion features at the lowest hierarchy in the third fusion features, or may also be features obtained by performing one or more feature extractions on the fusion features at the lowest hierarchy in the third fusion feature; the batch of fusion features obtained by means of feature fusion along the direction from the low hierarchy feature to the high hierarchy feature include the fusion features at the lowest hierarchy in the third fusion feature and the fusion feature obtained from each fusion by means of a fusion operation in operation 306.

This embodiment is explained using one retracing fusion as an example: if the features at the at least two different hierarchies are subjected to two or more retracing fusions, operations 304-306 are executed for multiple times, and the batch of fusion features obtained finally is the second fusion feature.

In an optional instance, the operations 304-306 is performed by a processor by invoking a corresponding instruction stored in a memory, or is performed by a second fusion module run by the processor.

In operation 308, the region features corresponding to the at least one instance candidate region are extracted from the second fusion feature respectively according

to each instance candidate region in the image.

In the embodiments of the present disclosure, for example, it may adopt, but not limited to, Region Proposal Network (RPN) to generate each instance candidate region for the image, and to map each instance candidate region on the features in the second fusion region; then for example, it may adopt, but not limited to, a method of
5 Region Of Interest Alignment (ROIAlign) to extract the region features corresponding to each instance candidate region from the second fusion region.

In an optional instance, operation 308 may be executed by a processor by invoking a corresponding instruction stored in a memory, or may be executed by an
10 extraction module run by the processor.

In operation 310, pixel-level fusion is respectively performed on the plurality of region features corresponding to the same instance candidate region, to obtain a fusion feature of each instance candidate region.

In an optional instance, the operation 310 is performed by a processor by
15 invoking a corresponding instruction stored in a memory, or is performed by a first fusion module run by the processor.

In operation 312, instance segmentation is performed respectively based on each first fusion feature, to obtain the instance segmentation result of the corresponding instance candidate region.

In an optional example, operation 312 may be executed by a processor by
20 invoking a corresponding instruction stored in a memory, or may be executed by a segmentation module run by the processor.

In an embodiment of each instance segmentation method embodiment of the present disclosure, based on a first fusion feature, performing instance segmentation
25 on the instance candidate region corresponding to the first fusion feature, to obtain the instance segmentation result of the corresponding instance candidate region may include:

based on the first fusion feature, performing pixel-level example class prediction, to obtain an example class prediction result of the example candidate
30 region corresponding to the first fusion feature; based on the first fusion feature, performing pixel-level foreground and background prediction, to obtain a foreground and background prediction result of the example candidate region corresponding to the first fusion feature, where the first fusion feature is a first fusion feature of any

instance candidate region;

based on the instance class prediction result and the foreground and background prediction result, obtaining the instance segmentation result of the instance object candidate region corresponding to the first fusion feature, where the instance segmentation result includes: an element belonging to a certain instance in the current instance candidate region and class information to which the instance belongs.

Based on this embodiment, based on the first fusion feature, pixel-level instance class prediction and foreground and background prediction are performed at the same time; the pixel-level instance class prediction can perform fine classification and multiple classification on the first fusion feature; better global information can be obtained by means of the foreground and background prediction; since detailed information between multiple instance classes is not required to be concerted, a prediction speed is improved; meanwhile, obtaining the example segmentation result of the instance object candidate region based on the instance class prediction result and the foreground and background prediction result can improve the instance segmentation result of the instance candidate region or image.

In an optional example, based on the first fusion feature, performing pixel-level instance class prediction may include:

performing feature extraction on the first fusion feature by means of a first convolutional network, the first convolutional network including at least one fully convolutional layer; and

performing pixel-level object class prediction by means of a first fully convolutional layer based on a feature output by the first convolutional network.

In an optional example, based on the first fusion feature, performing pixel-level foreground and background prediction includes:

based on the first fusion feature, predicting elements belonging to the foreground and/or elements belonging to the background in the instance candidate region corresponding to the first fusion feature,

where the background and foreground can be set according to requirements. For example, the foreground may include portions corresponding to all instance classes, and the background may include portions other than the portions corresponding to all instance classes; or the background may include portions

corresponding to all instance classes, and the foreground may include portions other than the portions corresponding to all instance classes.

In another optional example, based on the first fusion feature, performing pixel-level foreground and background prediction may include:

5 performing feature extraction on the first fusion feature by means of a second convolutional network, the second convolutional network including at least one fully convolutional layer; and

performing pixel-level foreground and background prediction by means of a full connection layer based on a feature output by the second convolutional network.

10 In an embodiment of the instance segmentation method embodiments of the present disclosure, based on the instance class prediction result and the foreground and background prediction result, obtaining the instance segmentation result of the instance object candidate region corresponding to the first fusion feature includes:

performing pixel-level summing on the object class prediction result of the
15 instance candidate region corresponding to the first fusion feature and the foreground and background prediction result of the instance candidate region corresponding to the first fusion feature, to obtain the instance segmentation result of the instance object candidate region corresponding to the first fusion feature.

In another embodiment approach, after obtaining the foreground and
20 background prediction result of the instance candidate region corresponding to the first fusion feature, it may further include: reshaping the foreground and background prediction result into a foreground and background prediction result with a dimension consistent with that of the instance class prediction result. For example, the foreground and background prediction result is reshaped from a vector to a matrix
25 having consistent dimension with the object class prediction. Accordingly, the performing pixel-level summing on the object class prediction result of the instance candidate region corresponding to the first fusion feature and the foreground and background prediction result of the instance candidate region corresponding to the first fusion feature may include: performing pixel-level summing on the instance class
30 prediction result of the instance candidate region corresponding to the first fusion feature and the reshaped foreground and background prediction result of the instance candidate region corresponding to the first fusion feature.

In the embodiment of the embodiments of the present disclosure, when

instance segmentation is performed respectively based on the first fusion feature of each instance candidate region, to obtain the instance segmentation result of each instance candidate region, since the pixel-level instance class prediction and foreground and background prediction based on the first fusion feature of the instance candidate region at the same time, the part of solution may be referred to as a double-path mask prediction; as shown in FIG. 4, it is a network structure schematic diagram of the double-path mask prediction executed in the embodiment of the present disclosure.

As shown in FIG. 4, the multiple region features corresponding to the instance candidate region are subjected to instance class prediction and foreground and background prediction by means of two branches. The first branch includes: four fully convolutional layers (conv1-conv4), i.e., the first convolution network and a deconvolutional layer (deconv), i.e., the first fully convolutional layer. The other branch includes: a third fully convolutional layer and a fourth convolutional layer (conv3 and conv4) from the first branch, and two fully convolutional layers (conv4-fc and conv5-fc), i.e., the second convolution network; a full connection layer (fc); and a reshaping layer, used for reshaping the foreground and background prediction result into a foreground and background prediction result having a consistent dimension with the instance class prediction result. The pixel-level mask prediction is performed by the first branch on each potential instant class, while the full connection layer performs the mask prediction irrelevant to the instance class (i.e., performing pixel-level foreground and background prediction). Finally, the mask predictions of the two branches are summed to obtain the final instance segmentation result.

FIG. 5 is a flow chart of an application embodiment of the instance segmentation method of the present disclosure. FIG. 6 is a process schematic diagram of the application embodiment shown in FIG. 5. Referring to FIG. 5 and FIG. 6 at the same time, the instance segmentation method of the application embodiment includes operations 502 to 516.

In operation 502, feature extraction is performed on an image via a neural network to output features M_1 - M_4 at four hierarchies output by the network layers with four different network depths in the neural network.

In an optional instance, operation 502 may be executed by invoking, by a processor, corresponding instructions stored in a memory, and may also be executed

by the neural network run by the processor.

In operation 504, in the feature at the four hierarchies, after upsampling the feature M_{i+1} at the higher hierarchy according to an order from the high hierarchy feature M_4 to the low hierarchy feature M_1 (i.e., from top to bottom) in sequence, the
5 upsampled feature is fused with the feature M_i at the lower hierarchy to obtain the first batch of fusion features P_2 - P_5 .

The values of i are integers in 1-3 in sequence. In the features participating in fusion and the first batch of fusion features, the fusion feature P_5 at the highest hierarchy is the feature M_4 at the highest hierarchy in the four different hierarchies or
10 the feature obtained by performing feature extraction on the feature M_4 by means of the fully convolutional layer; the first fusion feature includes the fusion feature at the highest hierarchy in the features at four different hierarchies and fusion features P_2 - P_5 obtained from each fusion.

In operation 506, in the first batch of fusion features, after downsampling the
15 fusion feature P_k at the lower hierarchy according to an order from the low hierarchy feature P_2 to the high hierarchy feature P_5 (i.e., from bottom to top) in sequence, the downsampled feature is fused with the feature P_{k+1} at the adjacent higher hierarchy to obtain the second batch of fusion features N_2 - N_5 .

The values of k are integers in 2-4 in sequence. In the fusion feature
20 participating in this fusion and the second batch of fusion features, the fusion feature N_2 at the lowest hierarchy is the fusion feature P_2 at the lowest hierarchy in the first batch of fusion features or the feature obtained by performing feature extraction on the fusion feature P_2 by means of the fully convolutional layer; the second batch of fusion features includes the feature corresponding to the feature P_2 at the lowest
25 hierarchy in the first batch of fusion features and the fusion feature obtained from each fusion, wherein the feature corresponding to the feature at the lowest hierarchy in the first batch of fusion features is the fusion feature P_2 at the lowest hierarchy in the first batch of fusion features or the feature obtained by performing feature extraction on the fusion feature P_2 by means of the convolutional layer.

30 The present application embodiment takes performing one retracing fusion on the features M_1 - M_4 at the four hierarchies as an example for explanation, and therefore, the second batch of fusion features obtained by means of operation 506 is the second fusion feature in each embodiment of the present disclosure.

In an optional instance, the operations 502-504 is performed by a processor by invoking a corresponding instruction stored in a memory, or is performed by a second fusion module run by the processor.

5 In operation 508, the region features corresponding to the at least one instance candidate region in the image is extracted from the second fusion features N_2 - N_5 .

In the embodiments of the present disclosure, for example, it may adopt, but not limited to, the RPN to generate at least one instance candidate region for the image, and to map each instance candidate region on the features in the second fusion region respectively; then for example, it may adopt, but not limited to, a method of
10 ROIAAlign to extract the region features corresponding to each instance candidate region from the second fusion region respectively.

In an optional instance, operation 508 may be executed by a processor by invoking a corresponding instruction stored in a memory, or may be executed by an extraction module run by the processor.

15 In operation 510, pixel-level fusion is respectively performed on the plurality of region features corresponding to the same instance candidate region, to obtain a first fusion feature of each instance candidate region.

In an optional instance, the operation 510 is performed by a processor by invoking a corresponding instruction stored in a memory, or is performed by a first
20 fusion module run by the processor.

Then, operations 512 and 516 are executed respectively.

In operation 512, instance segmentation is performed based on the first fusion feature of each instance candidate region to obtain the instance segmentation result of each instance candidate region.

25 The instance segmentation result includes an object box or position of each instance and an instance class to which the instance belongs.

In an optional example, operation 512 may be executed by a processor by invoking a corresponding instruction stored in a memory, or may be executed by a first segmentation unit run by the processor.

30 Then, following procedures of this application embodiment are not executed.

In operation 514, pixel-level instance class prediction is respectively performed based on the first fusion feature of each instance candidate region, to obtain the instance class prediction result of each instance candidate region, and pixel-

level foreground and background prediction is respectively performed based on the first fusion feature of each instance candidate region, to obtain the foreground and background prediction result of each instance candidate region.

5 In an optional example, the operation 514 is performed by a processor by invoking a corresponding instruction stored in a memory, or is performed by a first segmentation unit or a first prediction sub-unit and a second prediction sub-unit in the first segmentation unit run by the processor.

10 In operation 516, pixel-level summing is performed on the object class prediction result of the first fusion feature of each instance object candidate region and the foreground and background prediction result, to obtain the instance segmentation result of the instance object candidate region corresponding to the first fusion feature.

15 The instance segmentation result includes: the element belonging to a certain instance in the current instance candidate region and the instance class to which the instance belongs, where the instance claims may be: background or a certain instance class.

20 In an optional example, the operation 516 is performed by a processor by invoking a corresponding instruction stored in a memory, or is performed by a first segmentation unit or an obtaining sub-unit in the first segmentation unit run by the processor.

No specific executing order for executing operation 512 and operations 514-516 exists, i.e., they can be executed at the same time, and can also be executed at any time order.

25 In addition, in the embodiments of the present disclosure, after obtaining the first fusion feature of each instance candidate region, semantic segmentation can further be performed on at least a part of the regions of the image based on the first fusion feature, to obtain the semantic segmentation result.

30 In addition, in the embodiments of the present disclosure, after obtaining the second fusion feature of each instance candidate region, semantic segmentation can further be performed on at least a part of the regions of the image based on the second fusion feature, to obtain the semantic segmentation result.

The semantic segmentation result may include, for example, the class to which each element belongs in the at least a part of the regions in the image.

In embodiments of the present invention, at least a part of regions of the image may be all regions or a partial region (e.g., the candidate region) of the image, i.e., semantic segmentation can be performed on the entire image to obtain the semantic segmentation result of the image and semantic segmentation can also be performed on
5 the partial image (e.g., the candidate region) to obtain the semantic segmentation result of the partial region. The candidate region may be, for example, the instance candidate region in each embodiment, or may also be a candidate region generated by other means.

In an optional example, the operation for performing semantic segmentation
10 on at least a part of regions of the image is performed by a processor by invoking a corresponding instruction stored in a memory, or is executed by a segmentation module or in the segmentation module run by the processor.

Based on the embodiments, the semantic segmentation on at least a part of regions of the image is implemented. In addition, based on the first fusion feature or
15 the second fusion feature, performing semantic segmentation on the at least a part of regions of the image may improve the accuracy of the image semantic segmentation result by using context information.

It should be explained that in the embodiments of the present disclosure, after obtaining the second fusion feature of each instance candidate region, instance
20 segmentation can also be performed based on the second fusion feature, to obtain at least one of the instance segmentation result of the corresponding instance candidate region or the instance segmentation result of the image. The implementation of performing instance segmentation based on the second fusion feature, to obtain at least one of the instance segmentation result of the corresponding instance candidate
25 region or the instance segmentation result of the image can refer to the embodiments of performing the instance segmentation based on the first fusion feature, to obtain at least one of the instance segmentation result of the corresponding instance candidate region or the instance segmentation result of the image; both can be implemented by using similar solutions, and details are not described below again.

30 Any instance segmentation method provided by the embodiments of the present disclosure is executed by any appropriate device having data processing capability, including, but not limited to, a terminal device and a server. Alternatively, any instance segmentation method provided in the embodiments of the present

disclosure is executed by a processor, for example, any instance segmentation method mentioned in the embodiments of the present disclosure is executed by the processor by invoking a corresponding instruction stored in a memory. Details are not described below again.

5 A person of ordinary skill in the art may understand that all or some steps for implementing the foregoing method embodiments are achieved by a program by instructing related hardware; the foregoing program can be stored in a computer-readable storage medium; when the program is executed, steps including the foregoing method embodiments are executed. Moreover, the foregoing storage
10 medium includes various media capable of storing program codes, such as Read-Only Memory (ROM), Random Access Memory (RAM), a magnetic disk, or an optical disk.

FIG. 7 is a structural schematic diagram of an embodiment of an instance segmentation apparatus of the present disclosure. The instance segmentation
15 apparatus in this embodiment can be used for implementing each instance segmentation method embodiment of the present disclosure. As shown in FIG. 7, the apparatus in this embodiment includes: a neural network, an extraction module, a first fusion module, and a segmentation module, where

 the neural network is configured to perform feature extraction on an image to
20 output features at at least two different hierarchies.

 The neural network may include at least two network layers with different network depths for performing feature extraction on the image and outputting features at the at least two hierarchies by means of the at least two network layers with the different network depths.

25 The extraction module is configured to extract region features corresponding to at least one instance candidate region in the image from the features at the at least two different hierarchies.

 The first fusion module is configured to fuse region features corresponding to a same instance candidate region, to obtain a first fusion feature of each instance
30 candidate region.

 The segmentation module is configured to perform instance segmentation based on each first fusion feature, to obtain at least one of an instance segmentation result of the corresponding instance candidate region or an instance segmentation

result of the image.

Based on the instance segmentation apparatuses provided by the embodiments of the present disclosure, the method includes: performing feature extraction on an image via a neural network to output features at at least two different hierarchies; extracting region features corresponding to at least one instance candidate region in the image from the features at the two different hierarchies, and fusing region features corresponding to a same instance candidate region, to obtain a first fusion feature of each instance candidate region; and performing instance segmentation based on each first fusion feature, to obtain at least one of an instance segmentation result of the corresponding instance candidate region or an instance segmentation result of the image. Embodiments of the present disclosure design solving the problem of instance segmentation based on a deep-learning framework; since the deep-learning has a strong modeling capability, it facilitates to obtain a better instance segmentation result. In addition, performing instance segmentation on the instance candidate region, with respect to directly performing instance segmentation on an entire image, can improve accuracy of instance segmentation, reduce the computing amount and complexity required for the instance segmentation, and improve the instance segmentation efficiency. Moreover, extracting region features corresponding to the instance candidate region from the features at the at least two different hierarchies for fusing and performing the instance segmentation based on the obtained fusion feature enable each instance candidate region to obtain information at different hierarchies at the same time; since the information extracted from the features at different hierarchies are all at different semantic hierarchies, accuracy of the instance segmentation result of each instance candidate region can be improved via context information.

FIG. 8 is a structural schematic diagram of another embodiment of the instance segmentation apparatus of the present disclosure. As shown in FIG. 8, upon comparison with the embodiment shown in FIG. 7, the instance segmentation apparatus of this embodiment further includes: a second fusion module, configured to perform at least one retracing fusion on the features at the at least two different hierarchies to obtain a second fusion feature. The one retracing fusion includes: fusing the features at different hierarchies output by the network layers with different network depths respectively based on a network depth direction of the neural network according to two different hierarchy directions in sequence. Accordingly, in this

embodiment, the extraction module is configured to extract the region features corresponding to the at least one instance candidate region from the second fusion feature.

In an embodiment, the two different hierarchy directions may include: a
5 direction from high hierarchy features to low hierarchy features and a direction from the low hierarchy features to the high hierarchy features.

Therefore, the according to the two different hierarchy directions in sequence may include: along the direction from the high hierarchy features to the low hierarchy features and the direction from the low hierarchy features to the high hierarchy
10 features in sequence; or along the direction from the low hierarchy features to the high hierarchy features and the direction from the high hierarchy features to the low hierarchy features in sequence.

In an optional example, the second fusion module, when fusing the features at the different hierarchies output by the network layers with the different network
15 depths respectively along the direction from the high hierarchy features to the low hierarchy features and the direction from the low hierarchy features to the high hierarchy features in sequence, is configured to: along a direction from deep to shallow of the network depth of the neural network, after upsampling features at a higher hierarchy output by a network layer with a deeper network depth in the neural
20 network in sequence, fuse the upsampled features with features at a lower hierarchy output by a network layer with a shallower network depth, to obtain a third fusion feature; and along the direction from the low hierarchy features to the high hierarchy features, after downsampling fusion features at a lower hierarchy in sequence, fuse the downsampled fusion features with fusion features at a higher hierarchy in the third
25 fusion feature.

The features at a higher hierarchy, for example, may include: features output by the network layer with a deeper network depth in the neural network or features obtained by performing at least one feature extraction on the features output by the network layer with the deeper network depth.

In an optional example, the second fusion module, after upsampling features at
30 the higher hierarchy output by the network layer with the deeper network depth in the neural network in sequence, when fusing the upsampled features with the features at the lower hierarchy output by the network layer with the shallower network depth, is

configured to: after upsampling the features at the higher hierarchy output by the network layer with the deeper network depth in the neural network in sequence, fuse the upsampled features with adjacent features at the lower hierarchy output by the network layer with the shallower network depth.

5 In an optional example, the second fusion module, after downsampling fusion features at the lower hierarchy in sequence, when fusing the downsampled fusion features with the fusion features at the higher hierarchy in the third fusion feature, is configured to: after downsampling the fusion features at the lower hierarchy in sequence, fuse the downsampled fusion features with adjacent fusion features at the
10 higher hierarchy in the third fusion feature.

 In an optional example, the second fusion module, when fusing the features at different hierarchies output by the network layers with the different network depths respectively along the direction from the low hierarchy features to the high hierarchy features and the direction from the high hierarchy features to the low hierarchy
15 features in sequence, is configured to: along a direction from shallow to deep of the network depth of the neural network, after downsampling features at a lower hierarchy output by a network layer with a shallower network depth in the neural network in sequence, fuse the downsampled features with features at the higher hierarchy output by the network layer with the deeper network depth, to obtain a
20 fourth fusion feature;

 along the direction from the high hierarchy features to the low hierarchy features, after upsampling fusion features at a higher hierarchy in sequence, fuse the upsampled fusion features with fusion features at a lower hierarchy in the fourth fusion feature.

25 The features at a lower hierarchy, for example, may include: features output by the network layer with a shallower network depth in the neural network or features obtained by performing at least one feature extraction on the features output by the network layer with the shallower network depth.

 In an optional example, the second fusion module, after downsampling
30 features at the lower hierarchy output by the network layer with the shallower network depth in the neural network in sequence, when fusing the downsampled features with the features at the higher hierarchy output by the network layer with the deeper network depth, is configured to: after downsampling the features at the lower

hierarchy output by the network layer with the shallower network depth in the neural network in sequence, fuse the downsampled features with adjacent features at the higher hierarchy output by the network layer with the deeper network depth.

In an optional example, the second fusion module, after upsampling fusion features at the higher hierarchy in sequence, when fusing the upsampled fusion features with the fusion features at the lower hierarchy in the fourth fusion feature, is configured to: after upsampling the fusion features at the higher hierarchy in sequence, fuse the upsampled fusion features with adjacent fusion features at the lower hierarchy in the fourth fusion feature.

In an optional example, the first fusion module, when fusing region features corresponding to a same instance candidate region, is configured to: respectively perform pixel-level fusion on a plurality of region features corresponding to the same instance candidate region.

For example, the first fusion module, when performing the pixel-level fusion on the plurality of region features corresponding to the same instance candidate region, is configured to: respectively perform element-wise max on the plurality of region features corresponding to the same instance candidate region; or respectively perform element-wise averaging on the plurality of region features corresponding to the same instance candidate region; or respectively perform element-wise summing on the plurality of region features corresponding to the same instance candidate region.

In addition, further referring to FIG. 8, in one of the embodiments of the present disclosure, the segmentation module may include:

a first segmentation module, configured to perform the instance segmentation on the instance candidate region corresponding to a first fusion feature based on the first fusion feature, to obtain the instance segmentation result of the corresponding instance candidate region; and/or,

a second segmentation module, configured to perform the instance segmentation on the image based on each first fusion feature, to obtain the instance segmentation result of the image.

FIG. 9 is a structural schematic diagram of an embodiment of a segmentation module in the embodiments of the present disclosure. As shown in FIG. 9, in the embodiments of the present disclosure, the segmentation module may include:

a first segmentation unit, configured to respectively perform the instance

segmentation on the instance candidate region corresponding to the each first fusion feature based on the each first fusion feature, to obtain the instance segmentation result of each instance candidate region; and

an obtaining unit, configured to obtain the instance segmentation result of the image based on the instance segmentation result of the each instance candidate region.

In an embodiment, the first segmentation unit includes:

a first prediction sub-unit, configured to perform pixel-level instance class prediction based on the first fusion feature;

a second prediction sub-unit, configured to perform pixel-level foreground and background prediction based on the first fusion feature, to obtain a foreground and background prediction result of the instance candidate region corresponding to the first fusion feature; and

an obtaining sub-unit, configured to obtain an instance segmentation result of an instance object candidate region corresponding to the first fusion feature based on the instance class prediction result and the foreground and background prediction result.

In an optional example, the second prediction sub-unit is configured to predict an element belonging to a foreground and/or an element belonging to a background in the instance candidate region corresponding to the first fusion feature based on the first fusion feature.

The foreground includes portions corresponding to all instance classes, and the background includes portions other than the portions corresponding to all instance classes; or the background includes portions corresponding to all instance classes, and the foreground includes portions other than the portions corresponding to all instance classes.

In an optional example, the first prediction sub-unit may include: a first convolutional network, configured to perform feature extraction on the first fusion feature; the first convolutional network including at least one fully convolutional layer; and a first fully convolutional layer, configured to perform pixel-level object class prediction based on a feature output by the first convolutional network.

In an optional example, the second prediction sub-unit may include: a second convolutional network, configured to perform feature extraction on the first fusion feature; the second convolutional network including at least one fully convolutional

layer; and a full connection layer, configured to perform pixel-level foreground and background prediction based on a feature output by the second convolutional network.

In an optional example, the obtaining sub-unit is configured to: perform pixel-level summing on the object class prediction result of the instance candidate region
5 corresponding to the first fusion feature and the foreground and background prediction result of the instance candidate region corresponding to the first fusion feature, to obtain the instance segmentation result of the instance object candidate region corresponding to the first fusion feature.

In addition, further referring to FIG. 9, in another embodiment, the first
10 segmentation unit may further include: a reshaping sub-unit, configured to reshape the foreground and background prediction result into a foreground and background prediction result with a dimension consistent with that of the instance class prediction result. Accordingly, in this embodiment, the obtaining sub-unit is configured to perform pixel-level summing on the instance class prediction result of the instance
15 candidate region corresponding to the first fusion feature and the reshaped foreground and background prediction result of the instance candidate region corresponding to the first fusion feature.

In addition, in one embodiment of the embodiments of the present disclosure, the segmentation module may further include: a third segmentation unit, configured to
20 perform semantic segmentation on at least a part of regions of the image based on the first fusion feature to obtain a semantic segmentation result; or, configured to perform semantic segmentation on at least a part of regions of the image based on the second fusion feature to obtain a semantic segmentation result.

In addition, another electronic device provided in the embodiments of the
25 present disclosure includes:

a memory, configured to store a computer program; and

a processor, configured to execute the computer program stored in the memory, where when the computer program is executed, the method according to the embodiments of the present disclosure is implemented.

FIG. 10 is a schematic structural diagram of an application embodiment of an
30 electronic device of the present disclosure. Referring to FIG. 10 below, a schematic structural diagram of an electronic device, which may be a terminal device or a server, suitable for implementing the embodiments of the present disclosure is shown. As

shown in FIG. 10, the electronic device includes one or more processors, a communication part, and the like. The one or more processors are, for example, one or more Central Processing Units (CPUs), and/or one or more Graphic Processing Units (GPUs), and the like. The processor may perform various appropriate actions and processing according to executable instructions stored in a Read-Only Memory (ROM) or executable instructions loaded from a storage section to a Random Access Memory (RAM). The communication part may include, but not limited to, a network card. The network card may include, but not limited to, an Infiniband (IB) network card. The processor may communicate with the ROM and/or the RAM, to execute executable instructions. The processor is connected to the communication part via a bus, and communicates with other target devices via the communication part, thereby implementing corresponding operations in any method provided in the embodiments of the present disclosure, for example, performing feature extraction on an image via a neural network to output features at at least two different hierarchies; extracting region features corresponding to at least one instance candidate region in the image from the features at the at least two different hierarchies, and fusing region features corresponding to a same instance candidate region, to obtain a first fusion feature of each instance candidate region; and performing instance segmentation based on each first fusion feature, to obtain at least one of an instance segmentation result of the corresponding instance candidate region or an instance segmentation result of the image.

In addition, the RAM may further store various programs and data required for operations of an apparatus. The CPU, the ROM, and the RAM are connected to each other via the bus. In the presence of the RAM, the ROM is an optional module. The RAM stores executable instructions, or writes the executable instructions into the ROM during running, where the executable instructions cause the processor to execute corresponding operations of any method of this disclosure. An input/output (I/O) interface is also connected to the bus. The communication part may be integrated, or may be configured to have a plurality of sub-modules (for example, a plurality of IB network cards) connected to the bus.

The following components are connected to the I/O interface: an input section including a keyboard, a mouse and the like; an output section including a Cathode-Ray Tube (CRT), a Liquid Crystal Display (LCD), a speaker and the like; the storage

section including a hard disk and the like; and a communication section of a network interface card including an LAN card, a modem and the like. The communication section performs communication processing via a network such as the Internet. A drive is also connected to the I/O interface according to requirements. A removable
5 medium such as a magnetic disk, an optical disk, a magneto-optical disk, a semiconductor memory or the like is mounted on the drive according to requirements, so that a computer program read from the removable medium may be installed on the storage section according to requirements.

It should be noted that the architecture illustrated in FIG. 10 is merely an
10 optional implementation mode. During specific practice, the number and types of the components in FIG. 10 may be selected, decreased, increased, or replaced according to actual requirements. Different functional components may be separated or integrated or the like. For example, the GPU and the CPU may be separated, or the GPU may be integrated on the CPU, and the communication portion may be separated
15 from or integrated on the CPU or the GPU or the like. These alternative implementations all fall within the scope of protection of this disclosure.

Particularly, a process described above with reference to a flowchart according to the embodiments of the present disclosure is implemented as a computer software program. For example, the embodiments of this disclosure include a computer
20 program product. The computer program product includes a computer program tangibly included in a machine-readable medium. The computer program includes a program code for performing a method shown in the flowchart. The program code may include instructions for performing steps of the face anti-counterfeiting detection method provided in the embodiments of the present disclosure. In such an
25 embodiment, the computer program is downloaded and installed from the network through the communication section, and/or is installed from the removable medium. When the computer program is executed by the CPU, the functions defined in the method according to the present disclosure are executed.

In addition, the embodiment of the present disclosure also provides a computer
30 program, including computer instructions. When the computer instructions run in a processor of a device, the instance segmentation method according to any one of the embodiments of the present disclosure is implemented.

In addition, the embodiment of the present disclosure also provides a

computer-readable storage medium having a computer program stored thereon. When the computer program is executed by a processor, the instance segmentation method according to any one of the embodiments of the present disclosure is implemented.

Embodiments of the present disclosure have wide applications in fields of autonomous driving, domestic robots, maps, etc., for example, the embodiments of the present disclosure may be applied to an automatic driving scene, to accurately identify different traffic participants in the automatic driving scene; the embodiments of the present disclosure may be applied to a street scene, to identify different buildings and objects used with road sign qualities in the street scene, thereby helping construction of an accurate map; the embodiments of the present disclosure may be applied to a domestic robot, for example, when a robot grasps an object, accurate pixel-level positioning for each object is required; using the embodiments of the present disclosure, accurate identification and positioning for the object can be implemented. It should be understood that the above scenes are only exemplary scenes, and should not be understood as limitations on the scope of protection of the present disclosure.

The embodiments in the specification are all described in a progressive manner, for same or similar parts in the embodiments, refer to these embodiments, and each embodiment focuses on a difference from other embodiments. The system embodiments correspond to the method embodiments substantially and therefore are only described briefly, and for the associated part, refer to the descriptions of the method embodiments.

The methods and apparatuses in this disclosure may be implemented in many manners. For example, the methods and apparatuses in this disclosure may be implemented with software, hardware, firmware, or any combination of software, hardware, and firmware. Unless otherwise specially stated, the foregoing sequences of steps of the methods are merely for description, and are not intended to limit the steps of the methods of this disclosure. In addition, in some embodiments, this disclosure may also be implemented as programs recorded in a recording medium. The programs include machine-readable instructions for implementing the methods according to this disclosure. Therefore, the present disclosure further covers the recording medium storing the programs for executing the methods according to the present disclosure.

The descriptions of this disclosure are provided for the purpose of examples

and description, and are not intended to be exhaustive or limit this disclosure to the disclosed form. Many modifications and changes are obvious to a person of ordinary skill in the art. The embodiments are selected and described to better describe a principle and an actual application of the present disclosure, and to make a person of
5 ordinary skill in the art understand the present disclosure, so as to design various embodiments with various modifications applicable to particular use.

CLAIMS

1. An instance segmentation method, comprising:
 - performing feature extraction on an image via a neural network to output
 - 5 features at at least two different hierarchies;
 - extracting region features corresponding to at least one instance candidate region in the image from the features at the at least two different hierarchies, and
 - fusing region features corresponding to a same instance candidate region, to obtain a first fusion feature of each instance candidate region; and
 - 10 performing instance segmentation based on each first fusion feature, to obtain at least one of an instance segmentation result of a corresponding instance candidate region or an instance segmentation result of the image.
2. The method according to claim 1, wherein the performing feature extraction on the
- 15 image via the neural network to output the features at the at least two different hierarchies comprises: performing feature extraction on the image via the neural network to output the features at the at least two different hierarchies via at least two network layers with different network depths in the neural network.
- 20 3. The method according to claim 1 or 2, further comprising: after outputting the features at the at least two different hierarchies, performing at least one retracing fusion on the features at the at least two different hierarchies to obtain a second fusion feature, wherein the one retracing fusion comprises: fusing the features at different hierarchies output by the network layers with the different network depths
- 25 respectively based on a network depth direction of the neural network according to two different hierarchy directions in sequence; and
- the extracting the region features corresponding to the at least one instance candidate region in the image from the features at the at least two different hierarchies comprises: extracting the region features corresponding to the at least one instance
- 30 candidate region from the second fusion feature.

4. The method according to claim 3, wherein the two different hierarchy directions comprise: a direction from high hierarchy features to low hierarchy features and a direction from the low hierarchy features to the high hierarchy features.
- 5 5. The method according to claim 4, wherein the according to the two different hierarchy directions in sequence comprises: along the direction from the high hierarchy features to the low hierarchy features and the direction from the low hierarchy features to the high hierarchy features in sequence; or along the direction from the low hierarchy features to the high hierarchy features and the direction from the high hierarchy features to the low hierarchy features in sequence.
- 10 6. The method according to claim 5, wherein the fusing the features at the different hierarchies output by the network layers with the different network depths respectively along the direction from the high hierarchy features to the low hierarchy features and the direction from the low hierarchy features to the high hierarchy features in sequence comprises:
- 15 along a direction from deep to shallow of the network depth of the neural network, after upsampling features at a higher hierarchy output by a network layer with a deeper network depth in the neural network in sequence, fusing the upsampled features with features at a lower hierarchy output by a network layer with a shallower network depth, to obtain a third fusion feature; and
- 20 along the direction from the low hierarchy features to the high hierarchy features, after downsampling fusion features at a lower hierarchy in sequence, fusing the downsampled fusion features with fusion features at a higher hierarchy in the third fusion feature.
- 25 7. The method according to claim 6, wherein the features at the higher hierarchy comprise: features output by the network layer with a deeper network depth in the neural network or features obtained by performing at least one feature extraction on the features output by the network layer with the deeper network depth.
- 30 8. The method according to claim 6 or 7, wherein the after upsampling features at the higher hierarchy output by the network layer with the deeper network depth in the

neural network in sequence, fusing the upsampled features with the features at the lower hierarchy output by the network layer with the shallower network depth comprises: after upsampling the features at the higher hierarchy output by the network layer with the deeper network depth in the neural network in sequence, fusing the upsampled features with adjacent features at the lower hierarchy output by the network layer with the shallower network depth.

9. The method according to any one of claims 6 to 8, wherein the after downsampling fusion features at the lower hierarchy in sequence, fusing the downsampled fusion features with the fusion features at the higher hierarchy in the third fusion feature comprises: after downsampling the fusion features at the lower hierarchy in sequence, fusing the downsampled fusion features with adjacent fusion features at the higher hierarchy in the third fusion feature.

10. The method according to claim 5, wherein the fusing the features at the different hierarchies output by the network layers with the different network depths respectively along the direction from the low hierarchy features to the high hierarchy features and the direction from the high hierarchy features to the low hierarchy features in sequence comprises:

along a direction from shallow to deep of the network depth of the neural network, after downsampling features at a lower hierarchy output by a network layer with a shallower network depth in the neural network in sequence, fusing the downsampled features with features at the higher hierarchy output by the network layer with the deeper network depth, to obtain a fourth fusion feature; and

along the direction from the high hierarchy features to the low hierarchy features, after upsampling fusion features at the higher hierarchy in sequence, fusing the upsampled fusion features with fusion features at a lower hierarchy in the fourth fusion feature.

11. The method according to claim 10, wherein the features at the lower hierarchy comprise: features output by the network layer with a shallower network depth in the neural network or features obtained by performing at least one feature extraction on the features output by the network layer with the shallower network depth.

12. The method according to claim 10 or 11, wherein the after downsampling features at the lower hierarchy output by the network layer with the shallower network depth in the neural network in sequence, fusing the downsampled features with the features
5 at the higher hierarchy output by the network layer with the deeper network depth comprises: after downsampling the features at the lower hierarchy output by the network layer with the shallower network depth in the neural network in sequence, fusing the downsampled features with adjacent features at the higher hierarchy output by the network layer with the deeper network depth.

10

13. The method according to any one of claims 10 to 12, wherein the after upsampling fusion features at the higher hierarchy in sequence, fusing the upsampled fusion features with the fusion features at the lower hierarchy in the fourth fusion feature comprises: after upsampling the fusion features at the higher hierarchy in sequence,
15 fusing the upsampled fusion features with adjacent fusion features at the lower hierarchy in the fourth fusion feature.

14. The method according to any one of claims 1 to 13, wherein the fusing the region features corresponding to the same instance candidate region comprises: respectively
20 performing pixel-level fusion on a plurality of region features corresponding to the same instance candidate region.

15. The method according to claim 14, wherein the performing the pixel-level fusion on the plurality of region features corresponding to the same instance candidate region
25 comprises: respectively performing element-wise max on the plurality of region features corresponding to the same instance candidate region; or respectively performing element-wise averaging on the plurality of region features corresponding to the same instance candidate region; or respectively performing element-wise summing on the plurality of region features corresponding to the same instance
30 candidate region.

16. The method according to any one of claims 1 to 15, wherein the performing the instance segmentation based on the each first fusion feature, to obtain at least one of

the instance segmentation result of the corresponding instance candidate region or the instance segmentation result of the image comprises at least one of:

performing the instance segmentation on the instance candidate region corresponding to a first fusion feature based on the first fusion feature, to obtain the

5 instance segmentation result of the corresponding instance candidate region; or,

performing the instance segmentation on the image based on the each first fusion feature, to obtain the instance segmentation result of the image.

17. The method according to any one of claims 1 to 16, wherein the performing the
10 instance segmentation based on the each first fusion feature, to obtain the instance segmentation result of the image comprises:

respectively performing the instance segmentation on the instance candidate region corresponding to the each first fusion feature based on the each first fusion feature, to obtain the instance segmentation result of each instance candidate region;

15 and

obtaining the instance segmentation result of the image based on the instance segmentation result of the each instance candidate region.

18. The method according to claim 16 or 17, wherein the performing the instance
20 segmentation on the instance candidate region corresponding to a first fusion feature based on the first fusion feature, to obtain the instance segmentation result of the corresponding instance candidate region comprises:

performing pixel-level instance class prediction based on the first fusion feature, to obtain an instance class prediction result of the instance candidate region
25 corresponding to the first fusion feature; performing pixel-level foreground and background prediction based on the first fusion feature, to obtain a foreground and background prediction result of the instance candidate region corresponding to the first fusion feature; and

obtaining the instance segmentation result of an instance object candidate
30 region corresponding to the first fusion feature based on the instance class prediction result and the foreground and background prediction result.

19. The method according to claim 18, wherein the performing pixel-level foreground and background prediction based on the first fusion feature comprises at least one of: predicting pixels belonging to a foreground in the instance candidate region corresponding to the first fusion feature based on the first fusion feature; or,
5 predicting pixels belonging to a background in the instance candidate region corresponding to the first fusion feature based on the first fusion feature.

20. The method according to claim 19, wherein the foreground comprises portions corresponding to all instance classes, and the background comprises portions other
10 than the portions corresponding to all instance classes; or the background comprises portions corresponding to all instance classes, and the foreground comprises portions other than the portions corresponding to all instance classes.

21. The method according to any one of claims 18 to 20, wherein the performing
15 pixel-level instance class prediction based on the first fusion feature comprises:
 performing feature extraction on the first fusion feature via a first convolutional network, the first convolutional network comprising at least one fully convolutional layer; and
 performing pixel-level object class prediction via a first fully convolutional
20 layer based on a feature output by the first convolutional network.

22. The method according to any one of claims 18 to 21, wherein the performing pixel-level foreground and background prediction based on the first fusion feature comprises:
25 performing feature extraction on the first fusion feature via a second convolutional network, the second convolutional network comprising at least one fully convolutional layer; and
 performing pixel-level foreground and background prediction via a full connection layer based on a feature output by the second convolutional network.

30 23. The method according to any one of claims 18 to 22, wherein the obtaining an instance segmentation result of an instance object candidate region corresponding to the first fusion feature based on the instance class prediction result and the foreground

and background prediction result comprises: performing pixel-level summing on the object class prediction result of the instance candidate region corresponding to the first fusion feature and the foreground and background prediction result of the instance candidate region corresponding to the first fusion feature, to obtain the instance segmentation result of the instance object candidate region corresponding to the first fusion feature.

24. The method according to claim 23, after obtaining the foreground and background prediction result of the instance candidate region corresponding to the first fusion feature, further comprising: reshaping the foreground and background prediction result into a foreground and background prediction result with a dimension consistent with that of the instance class prediction result; and

the performing pixel-level summing on the object class prediction result of the instance candidate region corresponding to the first fusion feature and the foreground and background prediction result of the instance candidate region corresponding to the first fusion feature comprising: performing pixel-level summing on the instance class prediction result of the instance candidate region corresponding to the first fusion feature and the reshaped foreground and background prediction result of the instance candidate region corresponding to the first fusion feature.

25. The method according to any one of claims 1 to 15, further comprising: after obtaining the first fusion feature of the each instance candidate region, performing semantic segmentation on at least a part of regions of the image based on the first fusion feature to obtain a semantic segmentation result.

26. The method according to any one of claims 3 to 15, after obtaining the second fusion feature of each instance candidate region, further comprising: performing semantic segmentation on at least a part of regions of the image based on the second fusion feature to obtain a semantic segmentation result.

27. An instance segmentation apparatus, comprising:
a neural network, configured to perform feature extraction on an image to output features at at least two different hierarchies;

an extraction module, configured to extract region features corresponding to at least one instance candidate region in the image from the features at the at least two different hierarchies;

5 a first fusion module, configured to fuse region features corresponding to a same instance candidate region, to obtain a first fusion feature of each instance candidate region; and

a segmentation module, configured to perform instance segmentation based on each first fusion feature, to obtain at least one of an instance segmentation result of a corresponding instance candidate region or an instance segmentation result of the
10 image.

28. The apparatus according to claim 27, wherein the neural network comprises at least two network layers with different network depths, configured to perform feature extraction on the image to output features at at least two different hierarchies via the
15 at least two network layers with the different network depths.

29. The apparatus according to claim 27 or 28, further comprising:

a second fusion module, configured to perform at least one retracing fusion on the features at the at least two different hierarchies to obtain a second fusion feature,
20 wherein the one retracing fusion comprises: fusing the features at the different hierarchies output by the network layers with the different network depths respectively based on a network depth direction of the neural network according to two different hierarchy directions in sequence; and the extraction module, configured to extract the region features corresponding to the at least one instance candidate
25 region from the second fusion feature.

30. The apparatus according to claim 29, wherein the two different hierarchy directions comprise: a direction from high hierarchy features to low hierarchy features and a direction from the low hierarchy features to the high hierarchy features.

31. The apparatus according to claim 30, wherein the according to the two different hierarchy directions in sequence comprises: along the direction from the high hierarchy features to the low hierarchy features and the direction from the low

hierarchy features to the high hierarchy features in sequence; or along the direction from the low hierarchy features to the high hierarchy features and the direction from the high hierarchy features to the low hierarchy features in sequence.

5 32. The apparatus according to claim 31, wherein the second fusion module, when
fusing the features at different hierarchies output by the network layers with the
different network depths respectively along the direction from the high hierarchy
features to the low hierarchy features and the direction from the low hierarchy
features to the high hierarchy features in sequence, is configured to: along a direction
10 from deep to shallow of the network depth of the neural network, after upsampling
features at a higher hierarchy output by a network layer with a deeper network depth
in the neural network in sequence, fuse the upsampled features with features at a
lower hierarchy output by a network layer with a shallower network depth, to obtain a
third fusion feature; and along the direction from the low hierarchy features to the
15 high hierarchy features, after downsampling fusion features at a lower hierarchy in
sequence, fuse the downsampled fusion features with fusion features at a higher
hierarchy in the third fusion feature.

20 33. The apparatus according to claim 32, wherein the features at the higher hierarchy
comprise:

features output by the network layer with a deeper network depth in the neural
network or features obtained by performing at least one feature extraction on the
features output by the network layer with the deeper network depth.

25 34. The apparatus according to claim 32 or 33, wherein the second fusion module,
after upsampling features at the higher hierarchy output by the network layer with the
deeper network depth in the neural network in sequence, when fusing the upsampled
features with the features at the lower hierarchy output by the network layer with the
shallower network depth, is configured to: after upsampling the features at the higher
30 hierarchy output by the network layer with the deeper network depth in the neural
network in sequence, fuse the upsampled features with adjacent features at the lower
hierarchy output by the network layer with the shallower network depth.

35. The apparatus according to any one of claims 32 to 34, wherein the second fusion module, after downsampling fusion features at the lower hierarchy in sequence, when fusing the downsampled fusion features with the fusion features at the higher hierarchy in the third fusion feature, is configured to: after downsampling the fusion
5 features at the lower hierarchy in sequence, fuse the downsampled fusion features with adjacent fusion features at the higher hierarchy in the third fusion feature.

36. The apparatus according to claim 31, wherein the second fusion module, when fusing the features at the different hierarchies output by the network layers with the
10 different network depths respectively along the direction from the low hierarchy features to the high hierarchy features and the direction from the high hierarchy features to the low hierarchy features in sequence, is configured to: along a direction from shallow to deep of the network depth of the neural network, after downsampling features at a lower hierarchy output by a network layer with a shallower network
15 depth in the neural network in sequence, fuse the downsampled features with features at the higher hierarchy output by the network layer with the deeper network depth, to obtain a fourth fusion feature; and along the direction from the high hierarchy features to the low hierarchy features, after upsampling fusion features at a high hierarchy in sequence, fuse the upsampled fusion features with fusion features at a lower hierarchy
20 in the fourth fusion feature.

37. The apparatus according to claim 36, wherein the features at the lower hierarchy comprise: features output by the network layer with a shallower network depth in the neural network or features obtained by performing at least one feature extraction on
25 the features output by the network layer with the shallower network depth.

38. The apparatus according to claim 36 or 37, wherein the second fusion module, after downsampling features at the lower hierarchy output by the network layer with the shallower network depth in the neural network in sequence, when fusing the
30 downsampled features with the features at the higher hierarchy output by the network layer with the deeper network depth, is configured to: after downsampling the features at the lower hierarchy output by the network layer with the shallower network depth in the neural network in sequence, fuse the downsampled features with adjacent

features at the higher hierarchy output by the network layer with the deeper network depth.

39. The apparatus according to any one of claims 36 to 38, wherein the second fusion
5 module, after upsampling fusion features at the higher hierarchy in sequence, when
fusing the upsampled fusion features with the fusion features at the lower hierarchy in
the fourth fusion feature, is configured to: after upsampling the fusion features at the
higher hierarchy in sequence, fuse the upsampled fusion features with adjacent fusion
features at the lower hierarchy in the fourth fusion feature.

10

40. The apparatus according to any one of claims 27 to 39, wherein the first fusion
module, when fusing region features corresponding to a same instance candidate
region, is configured to: respectively perform pixel-level fusion on a plurality of
region features corresponding to the same instance candidate region.

15

41. The apparatus according to claim 40, wherein the first fusion module, when
performing the pixel-level fusion on the plurality of region features corresponding to
the same instance candidate region, is configured to: respectively perform element-
wise max on the plurality of region features corresponding to the same instance
20 candidate region; or respectively perform element-wise averaging on the plurality of
region features corresponding to the same instance candidate region; or respectively
perform element-wise summing on the plurality of region features corresponding to
the same instance candidate region.

25 42. The apparatus according to any one of claims 27 to 41, wherein the segmentation
module comprises at least one of:

a first segmentation module, configured to perform the instance segmentation
on a instance candidate region corresponding to a first fusion region based on the first
fusion feature, to obtain the instance segmentation result of the corresponding
30 instance candidate region; or,

a second segmentation module, configured to perform the instance
segmentation on the image based on the each first fusion feature, to obtain the
instance segmentation result of the image.

43. The apparatus according to any one of claims 27 to 41, wherein the segmentation module comprises:

5 a first segmentation unit, configured to respectively perform the instance segmentation on the instance candidate region corresponding to the each first fusion feature based on the each first fusion feature, to obtain the instance segmentation result of the each instance candidate region; and

an obtaining unit, configured to obtain the instance segmentation result of the image based on the instance segmentation result of the each instance candidate region.

10

44. The apparatus according to claim 42 or 43, wherein the first segmentation unit comprises:

15 a first prediction sub-unit, configured to perform pixel-level instance class prediction based on the first fusion feature, to obtain an instance class prediction result of the instance candidate region corresponding to the first fusion feature;

a second prediction sub-unit, configured to perform pixel-level foreground and background prediction based on the first fusion feature, to obtain a foreground and background prediction result of the instance candidate region corresponding to the first fusion feature; and

20 an obtaining sub-unit, configured to obtain the instance segmentation result of an instance object candidate region corresponding to the first fusion feature based on the instance class prediction result and the foreground and background prediction result.

25 45. The apparatus according to claim 44, wherein the second prediction sub-unit is configured to perform at least one of: predicting pixels belonging to a foreground in the instance candidate region corresponding to the first fusion feature based on the first fusion feature; or,

30 predicting pixels belonging to a background in the instance candidate region corresponding to the first fusion feature based on the first fusion feature.

46. The apparatus according to claim 45, wherein the foreground comprises portions corresponding to all instance classes, and the background comprises portions other than the portions corresponding to all instance classes; or

the background comprises portions corresponding to all instance classes, and
5 the foreground comprises portions other than the portions corresponding to all instance classes.

47. The apparatus according to any one of claims 44 to 46, wherein the first prediction sub-unit comprises:

10 a first convolutional network, configured to perform feature extraction on the first fusion feature, the first convolutional network comprising at least one fully convolutional layer; and

a first fully convolutional layer, configured to perform pixel-level object class prediction based on a feature output by the first convolutional network.

15

48. The apparatus according to any one of claims 44 to 47, wherein the second prediction sub-unit comprises:

a second convolutional network, configured to perform feature extraction on the first fusion feature, the second convolutional network comprising at least one fully
20 convolutional layer; and

a full connection layer, configured to perform pixel-level foreground and background prediction based on a feature output by the second convolutional network.

49. The apparatus according to any one of claims 44 to 48, wherein the obtaining sub-
25 unit is configured to perform pixel-level summing on the object class prediction result of the instance candidate region corresponding to the first fusion feature and the foreground and background prediction result of the instance candidate region corresponding to the first fusion feature, to obtain the instance segmentation result of the instance object candidate region corresponding to the first fusion feature.

30

50. The apparatus according to claim 49, wherein the first segmentation unit further comprises:

a reshaping sub-unit, configured to reshape the foreground and background prediction result into a foreground and background prediction result with a dimension consistent with that of the instance class prediction result; and

5 the obtaining sub-unit, configured to perform pixel-level summing on the instance class prediction result of the instance candidate region corresponding to the first fusion feature and the reshaped foreground and background prediction result of the instance candidate region corresponding to the first fusion feature.

10 51. The apparatus according to any one of claims 27 to 50, wherein the segmentation module further comprises: a third segmentation unit, configured to perform semantic segmentation on at least a part of regions of the image based on the first fusion feature to obtain a semantic segmentation result.

15 52. The apparatus according to any one of claims 29 to 50, wherein the segmentation module further comprises: a third segmentation unit, configured to perform semantic segmentation on at least a part of regions of the image based on the second fusion feature to obtain a semantic segmentation result.

20 53. An electronic device, comprising:
a memory, configured to store a computer program; and
a processor, configured to execute the computer program stored in the memory, wherein when the computer program is executed, the method according to any one of claims 1 to 26 is implemented.

25 54. A computer-readable storage medium having a computer program stored thereon, wherein when the computer program is executed by a processor, the method according to any one of claims 1 to 26 is implemented.