(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
16 April 2020 (16.04.2020)

WIPO | PCT

(10) International Publication Number
WO 2020/073310 A1

(51) International Patent Classification:
G06T 7/246 (2017.01)

(21) International Application Number:
PCT/CN2018/110023

(22) International Filing Date:
12 October 2018 (12.10.2018)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicant: **NOKIA TECHNOLOGIES OY** [FI/FI]; Karaportti 3, Espoo, 02610 (FI).

(71) Applicant *(for LC only)*: **NOKIA TECHNOLOGIES (BEIJING) CO., LTD.** [CN/CN]; Unit 18, Room 1001, Level 10, East Tower, World Financial Center, No. 1, East Third Ring Middle Road, Chaoyang District, Beijing 100020 (CN).

(72) Inventor: **NIE, Jing**; Tianjin Tiandatz Technology Co. Ltd., Room 6-4-501, Xinyuan Cun, Songshan Road, Nankai District, Tianjin 300072 (CN).

(74) Agent: **KING & WOOD MALLESONS**; 20th Floor, East Tower, World Financial Centre, No. 1 Dongsanhuan Zhonglu, Chaoyang District, Beijing 100020 (CN).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report (Art. 21(3))*

(54) **Title:** METHOD AND APPARATUS FOR CONTEXT-EMBEDDING AND REGION-BASED OBJECT DETECTION

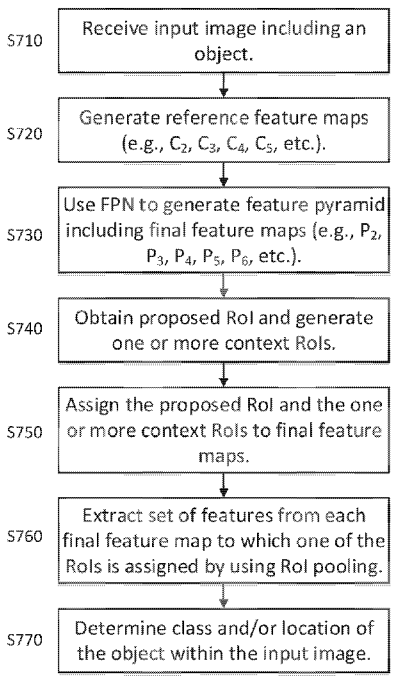

FIG. 7

(57) **Abstract:** A method of detecting an object in an image using a convolutional neural network (CNN) includes generating, based on the image, a plurality of reference feature maps and a corresponding feature pyramid including a plurality of final feature maps; obtaining a proposed region of interest (ROI); generating at least a first context ROI having an area larger than an area of the proposed ROI; assigning the proposed ROI and the first context ROI to a first and second final feature maps having different sizes; extracting, by performing ROI pooling, a first set of features from the first final feature map using the proposed ROI and a second set of features from the second final feature map using the first context ROI; and determining, based on the first and second sets of extracted features, at least one of a location of the object and a class of the object.

WO 2020/073310 A1

METHOD AND APPARATUS FOR CONTEXT-EMBEDDING AND REGION-BASED OBJECT DETECTION

BACKGROUND

1. Field

[0001] Various example embodiments relate generally to methods and apparatuses for performing region-based object detection.

2. Related Art

[0002] Object detection is a task in the area of computer vision that is aimed at localizing and recognizing object instances with a bounding box. Convolutional neural network (CNN)-based object detection can be utilized in the areas of visual surveillance, Advanced Driver Assistant Systems (ADAS), and human-machine interaction (HMI).

[0003] Current object detection frameworks could be grouped into two main streams: the region-based methods and the region-free methods. Examples of region-based detectors are discussed in, for example, *Y. S. Cao, X. Niu, and Y. Dou, "Region-based convolutional neural networks for object detection in very high resolution remote sensing images," In International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery, 2016; R. Girshick, "Fast r-cnn," Computer Science, 2015*; and *S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," in International Conference on Neural Information Processing Systems, 2015, pp. 91–99.* Generally, region-based methods divide the object detection into two steps. In the first step, a region proposal network (RPN) generates high-quality proposals. Then, in the second step, the proposals are further classified and regressed by a region-wise subnet. Generally, the region-free methods detect objects by regular and dense sampling over locations, scales and aspect ratios.

SUMMARY

[0004] According to at least some example embodiments, a method of detecting an object in an image using a convolutional neural network (CNN) includes generating, by the CNN, a plurality of reference feature maps based on the image; generating a feature pyramid including a plurality of final feature maps corresponding, respectively, to the plurality of reference feature maps; obtaining a proposed region of interest (ROI); generating at least a first context ROI based on the proposed ROI such that an area of the first context ROI is larger than an area of the proposed ROI; assigning the proposed ROI to a first final feature map from among the plurality of final feature maps; assigning the first context ROI to a second final feature map from among the plurality of final feature maps, a size of the first final feature map being different than a size of the second final feature map; extracting a first set of features from the first final feature map by performing an ROI pooling operation on the first final feature map using the proposed ROI; extracting a second set of features from the second final feature map by performing an ROI pooling operation on the second final feature map using the first context ROI; and determining, based on the first and second sets of extracted features, at least one of a location of the object with respect to the image and a class of the object.

[0005] The feature pyramid may be generated based on the plurality of reference feature maps in accordance with a feature pyramid network (FPN) architecture.

[0006] The area of the first context ROI may be 2^2 times the area of the proposed ROI.

[0007] The method may further include concatenating the first and second sets of extracted features, wherein the determining includes determining, based on the concatenated sets of extracted features, at least one of a location of the object with respect to the image and a class of the object.

[0008] The method may further include applying the concatenated sets of extracted features to a squeeze-and-excitation block (SEB), wherein the at least one of a location of the object with respect to the image and a class of the object is determined based on an output of the SEB.

[0009] The method may further include generating a second context ROI based on the proposed ROI such that an area of the second context ROI is larger than an area of the first context ROI; assigning the second context ROI to a third final feature map from among the plurality of final feature maps, a size of the third final feature map being different than the sizes of the first and second final feature maps; and extracting a third set of features from the first final feature map by performing ROI pooling on the first final feature map using the second context ROI, wherein the determining includes determining, based on the first, second and third sets of extracted features, at least one of the location of the object with respect to the image and the class of the object.

[0010] The feature pyramid may be generated based on the plurality of reference feature maps in accordance with a feature pyramid network (FPN) architecture.

[0011] The area of the first context ROI may be 2^2 times the area of the proposed ROI, and the area of the second context ROI may be 4^2 times the area of the area of the proposed ROI.

[0012] The method may further include concatenating the first, second and third sets of extracted features, wherein the determining includes determining, based on the concatenated sets of extracted features, at least one of a location of the object with respect to the image and a class of the object.

[0013] The method may further include applying the concatenated sets of extracted features to a squeeze-and-excitation block (SEB), wherein the at least one of a location of the

object with respect to the image and a class of the object is determined based on an output of the SEB.

[0014] According to at least some example embodiments, a computer-readable medium includes program instructions for causing an apparatus to perform at least generating, by a convolutional neural network (CNN), a plurality of reference feature maps based on an image that includes an object; generating a feature pyramid including a plurality of final feature maps corresponding, respectively, to the plurality of reference feature maps; obtaining a proposed region of interest (ROI); generating at least a first context ROI based on the proposed ROI such that an area of the first context ROI is larger than an area of the proposed ROI; assigning the proposed ROI to a first final feature map from among the plurality of final feature maps; assigning the first context ROI to a second final feature map from among the plurality of final feature maps, a size of the first final feature map being different than a size of the second final feature map; extracting a first set of features from the first final feature map by performing an ROI pooling operation on the first final feature map using the proposed ROI; extracting a second set of features from the second final feature map by performing an ROI pooling operation on the second final feature map using the first context ROI; and determining, based on the first and second sets of extracted features, at least one of a location of the object with respect to the image and a class of the object.

[0015] The feature pyramid may be generated based on the plurality of reference feature maps in accordance with a feature pyramid network (FPN) architecture.

[0016] The area of the first context ROI may be 2^2 times the area of the proposed ROI.

[0017] The computer-readable medium may further include program instructions for causing an apparatus to perform 1t least concatenating the first and second sets of extracted features, wherein the determining includes determining, based on the concatenated sets of

extracted features, at least one of a location of the object with respect to the image and a class of the object.

[0018] The computer-readable medium of claim 14 may further include program instructions for causing an apparatus to perform at least applying the concatenated sets of extracted features to a squeeze-and-excitation block (SEB), wherein the at least one of a location of the object with respect to the image and a class of the object is determined based on an output of the SEB.

[0019] According to at least some example embodiments, an apparatus includes at least one processor; and at least one memory including computer program code, the at least one memory and the computer program code configured to, with the at least one processor, cause the apparatus at least to perform, generating, by a convolutional neural network (CNN), a plurality of reference feature maps based on an image that includes an object; generating a feature pyramid including a plurality of final feature maps corresponding, respectively, to the plurality of reference feature maps; obtaining a proposed region of interest (ROI); generating at least a first context ROI based on the proposed ROI such that an area of the first context ROI is larger than an area of the proposed ROI; assigning the proposed ROI to a first final feature map from among the plurality of final feature maps; assigning the first context ROI to a second final feature map from among the plurality of final feature maps, a size of the first final feature map being different than a size of the second final feature map; extracting a first set of features from the first final feature map by performing an ROI pooling operation on the first final feature map using the proposed ROI; extracting a second set of features from the second final feature map by performing an ROI pooling operation on the second final feature map using the first context ROI; and determining, based on the first and second sets of extracted features, at least one of a location of the object with respect to the image and a class of the object.

[0020] The feature pyramid may be generated based on the plurality of reference feature maps in accordance with a feature pyramid network (FPN) architecture.

[0021] The area of the first context ROI may be twice the area of the proposed ROI.

[0022] The at least one memory and the computer program code may be further configured to, with the at least one processor, cause the apparatus at least to perform concatenating the first and second sets of extracted features, wherein the determining includes determining, based on the concatenated sets of extracted features, at least one of a location of the object with respect to the image and a class of the object.

[0023] The at least one memory and the computer program code may be further configured to, with the at least one processor, cause the apparatus at least to perform applying the concatenated sets of extracted features to a squeeze-and-excitation block (SEB), wherein the at least one of a location of the object with respect to the image and a class of the object is determined based on an output of the SEB.

BRIEF DESCRIPTION OF THE DRAWINGS

[0024] At least some example embodiments will become more fully understood from the detailed description provided below and the accompanying drawings, wherein like elements are represented by like reference numerals, which are given by way of illustration only and thus are not limiting of example embodiments and wherein:

[0025] FIG. 1 is a diagram of a surveillance network 10 according to at least some example embodiments.

[0026] FIG. 2 is a diagram illustrating an example structure of an object detection device according to at least some example embodiments.

[0027] FIG. 3 illustrates an object detection sub-network of a multi-scale convolutional neural network (MS-CNN) detector.

[0028] FIG. 4 illustrates a portion of a backbone convolutional neural network (CNN) according to at least some example embodiments.

[0029] FIG. 5 illustrates a feature pyramid network (FPN) according to at least some example embodiments.

[0030] FIG. 6 illustrates a diagram of a portion of a context-embedding, region-based objection detection network 600 according to at least some example embodiments.

[0031] FIG. 7 is a flow chart illustrating an example algorithm for performing the context-embedding, region-based object detection method according to at least some example embodiments.

DETAILED DESCRIPTION OF EXAMPLE EMBODIMENTS

[0032] Various example embodiments will now be described more fully with reference to the accompanying drawings in which some example embodiments are shown.

[0033] Detailed illustrative embodiments are disclosed herein. However, specific structural and functional details disclosed herein are merely representative for purposes of describing at least some example embodiments. Example embodiments may, however, be embodied in many alternate forms and should not be construed as limited to only the embodiments set forth herein.

[0034] Accordingly, while example embodiments are capable of various modifications and alternative forms, embodiments thereof are shown by way of example in the drawings and will herein be described in detail. It should be understood, however, that there is no intent to limit example embodiments to the particular forms disclosed, but on the contrary, example embodiments are to cover all modifications, equivalents, and alternatives falling within the scope of example embodiments. Like numbers refer to like elements throughout

the description of the figures. As used herein, the term "and/or" includes any and all combinations of one or more of the associated listed items.

[0035] It will be understood that when an element is referred to as being "connected" or "coupled" to another element, it can be directly connected or coupled to the other element or intervening elements may be present. In contrast, when an element is referred to as being "directly connected" or "directly coupled" to another element, there are no intervening elements present. Other words used to describe the relationship between elements should be interpreted in a like fashion (e.g., "between" versus "directly between", "adjacent" versus "directly adjacent", etc.).

[0036] The terminology used herein is for the purpose of describing particular embodiments only and is not intended to be limiting of example embodiments. As used herein, the singular forms "a", "an" and "the" are intended to include the plural forms as well, unless the context clearly indicates otherwise. It will be further understood that the terms "comprises", "comprising,", "includes" and/or "including", when used herein, specify the presence of stated features, integers, steps, operations, elements, and/or components, but do not preclude the presence or addition of one or more other features, integers, steps, operations, elements, components, and/or groups thereof.

[0037] It should also be noted that in some alternative implementations, the functions/acts noted may occur out of the order noted in the figures. For example, two figures shown in succession may in fact be executed substantially concurrently or may sometimes be executed in the reverse order, depending upon the functionality/acts involved.

[0038] Exemplary embodiments are discussed herein as being implemented in a suitable computing environment. Although not required, exemplary embodiments will be described in the general context of computer-executable instructions (e.g., program code), such as program modules or functional processes, being executed by one or more computer

processors or CPUs. Generally, program modules or functional processes include routines, programs, objects, components, data structures, etc. that performs particular tasks or implement particular abstract data types.

[0039] In the following description, illustrative embodiments will be described with reference to acts and symbolic representations of operations (e.g., in the form of flowcharts) that are performed by one or more processors, unless indicated otherwise. As such, it will be understood that such acts and operations, which are at times referred to as being computer-executed, include the manipulation by the processor of electrical signals representing data in a structured form. This manipulation transforms the data or maintains it at locations in the memory system of the computer, which reconfigures or otherwise alters the operation of the computer in a manner well understood by those skilled in the art.

I. Overview

[0040] As is discussed in greater detail below, a context-embedding, region-based object detection method according to at least some example embodiments is based on region-based object detection methods and includes embedding a context branch in order to obtain rich context information thereby resulting in improved object detection. According to at least some example embodiments, the context information is beneficial for detecting small, blurred and occluded objects. Further, as is also discussed in greater detail below, the context-embedding, region-based object detection method according to at least some example embodiments employs a squeeze-and-excitation block in conjunction with the context branch to reduce or, alternatively, avoid noise information. The context-embedding, region-based object detection method according to at least some example embodiments can be applied in several different ways including, for example, visual surveillance.

[0041] Example structures of a surveillance network and object detection device 100 which may utilize the context-embedding, region-based object detection method according to at least some example embodiments will be discussed below in section II of the present disclosure. Next, examples of using feature pyramids and context embedding to perform object detection will be discussed in section III of the present disclosure. Next, examples of a convolutional neural network (CNN) architecture and algorithm for performing the context-embedding, region-based object detection method according to at least some example embodiments will be discussed in section IV of the present disclosure. Further, methods of training the CNN architecture will be discussed in section V of the present disclosure.

II. Example structures for implementing the context-embedding, region-based object detection method according to at least some example embodiments

[0042] For example, FIG. 1 illustrates a diagram of a surveillance network 10 according to at least some example embodiments. As is shown in FIG. 1, the surveillance network 10 may include an object detection device 100 and a surveillance system 150.

[0043] The surveillance system 150 may include one or more cameras each capturing image data representing a scene in a vicinity of the location of the camera. For example, as is illustrated in FIG. 1, the surveillance system 150 includes the camera 152, which captures surveillance scene 154. The camera 152 may capture surveillance scene 154 by, for example, continuously capturing a plurality of temporally-adjacent images (i.e., capturing video or moving image data) of the surveillance scene 154. According to at least some example embodiments, the camera 152 transmits image data 120 corresponding to the captured surveillance scene 154 to the object detection device 100. An example structure of the object detection device 100 will now be discussed in greater detail below with reference to FIG. 2.

[0044] FIG. 2 is a diagram illustrating an example structure of the object detection device 100 according to at least some example embodiments.

[0045] Referring to FIG. 2, the object detection device 100 may include, for example, a data bus 259, a transmitting unit 252, a receiving unit 254, a memory unit 256, and a processing unit 258.

[0046] The transmitting unit 252, receiving unit 254, memory unit 256, and processing unit 258 may send data to and/or receive data from one another using the data bus 259.

[0047] The transmitting unit 252 is a device that includes hardware and any necessary software for transmitting signals including, for example, control signals or data signals via one or more wired and/or wireless connections to one or more other network elements in a wireless communications network.

[0048] The receiving unit 254 is a device that includes hardware and any necessary software for receiving wireless signals including, for example, control signals or data signals via one or more wired and/or wireless connections to one or more other network elements in a wireless communications network.

[0049] The memory unit 256 may be any device capable of storing data including magnetic storage, flash storage, etc. Further, though not illustrated, the memory unit 256 may further include one or more of a port, dock, drive (e.g., optical drive), or opening for receiving and/or mounting removable storage media (e.g., one or more of a USB flash drive, an SD card, an embedded MultiMediaCard (eMMC), a CD, a DVD, and a Blue-ray disc).

[0050] The processing unit 258 may be any device capable of processing data including, for example, a processor.

[0051] According to at least one example embodiment, any operations described herein, for example with reference to any of FIGS. 1-7, as being performed by a an object detection device may be performed by an electronic device having the structure of the object detection

device 100 illustrated in FIG. 2. For example, according to at least one example embodiment, the object detection device 100 may be programmed, in terms of software and/or hardware, to perform any or all of the functions described herein as being performed by an object detection device. Consequently, the object detection device 100 may be embodied as a special purpose computer through software and/or hardware programming.

[0052] Examples of the object detection device 100 being programmed, in terms of software, to perform any or all of the functions described herein as being performed by an object detection device will now be discussed below. For example, the memory unit 256 may store a program that includes executable instructions (e.g., program code) corresponding to any or all of the operations described herein as being performed by an object detection device. According to at least one example embodiment, additionally or alternatively to being stored in the memory unit 256, the executable instructions (e.g., program code) may be stored in a computer-readable medium including, for example, an optical disc, flash drive, SD card, etc., and the object detection device 100 may include hardware for reading data stored on the computer readable-medium. Further, the processing unit 258 may be a processor configured to perform any or all of the operations described herein with reference to FIGS. 1-4 as being performed by an object detection device, for example, by reading and executing the executable instructions (e.g., program code) stored in at least one of the memory unit 256 and a computer readable storage medium loaded into hardware included in the object detection device 100 for reading computer-readable mediums.

[0053] Examples of the object detection device 100 being programmed, in terms of hardware, to perform any or all of the functions described herein as being performed by an object detection device will now be discussed below. Additionally or alternatively to executable instructions (e.g., program code) corresponding to the functions described with reference to FIGS. 1-7 as being performed by an object detection device being stored in a

memory unit or a computer-readable medium as is discussed above, the processing unit 258 may include a circuit (e.g., an integrated circuit) that has a structural design dedicated to performing any or all of the operations described herein with reference to FIGS. 1-6 as being performed by an object detection device. For example, the above-referenced circuit included in the processing unit 258 may be a FPGA or ASIC physically programmed, through specific circuit design, to perform any or all of the operations described with reference to FIGS. 1-7 as being performed by an object detection device.

[0054] According to at least some example embodiments, the object detection device 100 performs region-based object detection using context embedding which results improving object detection performance with respect to small, blurred and occluded objects with reference to other object detection methods, while also being able to detect objects at multiple scales. Two features used by some other object detection methods, feature pyramids and embedding context will now be discussed in greater detail below in section III.

III. Feature pyramids and embedding context

[0055] For example, some object detection methods utilize feature pyramids, which include feature maps of multiple levels (i.e., multiple scales). For example, the region-based detector, multi-scale CNN (MS-CNN), uses convolutional layers of different spatial resolutions to generate region proposals of different scales. However, the different layers of the MS-CNN detector may have an inconsistent semantic. An example of the MS-CNN is discussed, for example, in *Z. Cai and Q. Fan, R. S Feris and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," European Conference on Computer Vision. Springer, Cham, 2016.*

[0056] Further, in addition to using feature pyramids to generate region proposals the MS-CNN detector also includes an object detection sub-network that utilizes context

embedding. FIG. 3 illustrates an object detection sub-network 300 of an MS-CNN detector. As is illustrated in FIG. 3, the MS-CNN object detection sub-network 300 includes trunk CNN layers 310, a first feature map 320 corresponding to a conv4-3 convolutional layer, and a second feature map 330 corresponding to a conv4-3-2x convolutional layer resulting from performing a deconvolution operation on first feature map 320 such that the second feature map 330 is an enlarged version of the first feature map 320. For the example depicted in FIG. 3, the first feature map 320 has the dimensions H/8 × W/8 × 512, and the second feature map 330 has the dimensions H/4 × W/4 × 512, where H is the height of the input image initially input to the MS-CNN detector and W is the width of the input image.

[0057] As is illustrated in FIG. 3, within the second feature map 330, there is a first region 334A (i.e., the innermost cube illustrated within the second feature map 330) and a second region 332A (i.e., the cube illustrated within the second feature map 330 as encompassing the first region 334A). The second region 332A is an enlarged version of the first region 334A and is 1.5 times as large as the first region 334A. Further, as is also illustrated in FIG. 3, features of the second feature map 330 corresponding to a first region 334A are reduced, by ROI pooling, to a first fixed-dimension feature map 334B having the dimensions 7 × 7 × 512. Further, features of the second feature map 330 corresponding to a second region 332A are reduced, by ROI pooling, to a second fixed-dimension feature map 332B, which also has the dimensions 7 × 7 × 512. As is illustrated in FIG. 3, the MS-CNN object detection sub-network 300 concatenates the first and second fixed-dimension feature maps 334B and 332B, reduces the resulting feature map to a third fixed-dimension feature map 340B having the dimensions 5 × 5 × 512, and feeds the features of the third fixed-dimension feature map 340B to a fully connected layer 350 for determination of a class probability 370 and a bounding box 360. By using the enlarged second region 332A in conjunction with the first region 334A, the MS-CNN detector attempts to embed context

information of a high level of the feature pyramid included in the MS-CNN detector. However, the richness of the context information corresponding to the enlarged second region 332A may be limited because the enlarged second region 332A and the first region 334A are both mapped to the same level of the feature pyramid (i.e., the conv4-3-2x layer).

[0058] In contrast, as is explained below with reference to FIGS. 4-6, a context-embedding, region-based object detection method according to at least some example embodiments disclosed herein includes embedding a context branch such that features corresponding to a proposed region of interest (RoI) and context information corresponding to one or more enlarged RoIs are extracted from multiple levels of the feature pyramid. Consequently, the richness of the extracted context information may be improved relative to context information of the MS-CNN detector, and thus, the object detection performance of the context-embedding, region-based object detection method according to at least some example embodiments may also be improved.

[0059] Examples of a convolutional neural network (CNN) architecture and an algorithm for performing the context-embedding, region-based object detection method according to at least some example embodiments will now be discussed in section IV of the present disclosure.

IV. Example CNN architecture and algorithm for implementing the context-embedding, region-based object detection method according to at least some example embodiments

[0060] According to at least some example embodiments, the CNN structures and algorithms discussed below with reference to FIGS. 4-7 may be implemented by the object detection device 100 discussed above with reference to FIGS. 1 and 2. Thus, any or all operations discussed below with reference to FIGS. 4-7 may be executed or controlled by the object detection device 100 (i.e., the processing unit 258).

[0061] According to at least some example embodiments, a CNN architecture for implementing the context-embedding, region-based object detection method may include a backbone CNN and a feature pyramid network (FPN) which may be used together to implement one or both of a region proposal network (RPN) and a context-embedding, region-based object detection network.

[0062] For example, FIG. 4 illustrates a portion of a backbone CNN 400 according to at least some example embodiments. Further, one type of CNN that may serve as the backbone CNN 400 is the residual network CNN (i.e., ResNet), examples of which (including ResNet36 and ResNet50) are discussed, for example, in *K He, X Zhang, S Ren, J Sun, "Deep Residual Learning for Image Recognition," Proc. IEEE Computer Vision and Pattern Recognition, 2016.* For the purpose of simplicity, the structure of the backbone CNN 400 illustrated in FIG. 4 is the structure of the ResNet36 CNN. However, according to at least some example embodiments, the backbone CNN 400 is implemented by the ResNet50 CNN. Further, the backbone CNN 400 is not limited to the ResNet36 CNN and the ResNet50 CNN. According to at least some example embodiments, the backbone CNN 400 may be implemented by any CNN that generates multiple feature maps having different scales.

[0063] As is shown in FIG. 4, when the backbone CNN 400 is implemented by a ResNet, the backbone CNN 400 may include a plurality of convolution layers which output a plurality of reference feature maps, respectively. For example, the backbone CNN 400 illustrated in FIG. 4 includes a first convolutional layer conv1_x (not illustrated), a second convolutional layer conv2_x that outputs a second reference feature map C_2, a third convolutional layer conv3_x that outputs a third reference feature map C_3, a fourth convolutional layer conv4_x that outputs a fourth reference feature map C_4, and a fifth convolutional layer conv5_x that outputs a fifth reference feature map C_5. As will be discussed in greater detail below with

reference to FIG. 5, the reference feature maps C_2, C_3, C_4, and C_5 may form the basis of an FPN.

[0064] FIG. 5 illustrates an FPN 500 according to at least some example embodiments. The FPN 500 may be constructed based on the reference feature maps (e.g., 2nd through fifth reference feature maps C_2, - C_5) of the backbone CNN 400. For example, examples of FPNs are discussed in *T. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection," Proc. IEEE Computer Vision and Pattern Recognition, 2017*; *T. Kong, F. Sun, A. Yao, H. Liu, M. Lu, and Y. Chen, "Ron: Reverse connection with objectness prior networks for object detection," Proc. IEEE Computer Vision and Pattern Recognition, 2017;* and *Lin T Y, Goyal P, Girshick R, et al., "Focal Loss for Dense Object Detection," Proc. IEEE Computer Vision and Pattern Recogn*ition. In contrast to the multi-scale feature maps of the MS-CNN detector discussed above with reference to FIG. 4, the FPN 500 employs a top-down architecture to create a feature pyramid that includes high-level semantic feature maps at all scales. For example, the FPN 500 creates final feature maps P_{k0+2}, P_{k0+1}, P_{k0}, P_{k0-1}, P_{k0-2} corresponding to reference feature maps C_{k0+2}, C_{k0+1}, C_{k0}, C_{k0-1}, C_{k0-2}, respectively, where k_0 is a constant, the value of which may be set, for example, in accordance with the preferences of a designer and/or user of the object detection device 100. The constant k_0 will be discussed in greater detail below with reference to equation 1 and FIGS. 6 and 7. Further, as is discussed in greater detail below with reference to FIGS. 6 and 7, the final feature maps P generated by the FPN 500 can be used for one or both of region proposal and context-embedding, region-based object detection.

[0065] FIG. 6 illustrates a diagram of a portion of a context-embedding, region-based objection detection network 600 according to at least some example embodiments. FIG. 7 is a flow chart illustrating an example algorithm for performing the context-embedding, region-

based object detection method according to at least some example embodiments. An example algorithm for performing the context-embedding, region-based object detection method according to at least some example embodiments will now be discussed with reference to FIGS. 4-7, with respect to an example scenario in which the algorithm is performed by the object detection device 100, and the object detection device 100 implements (i.e., embodies) the backbone CNN 400, FPN 500, and objection detection network 600. Thus, operations described with respect to FIGS. 4-7 as being performed by the backbone CNN 400, FPN 500, or objection detection network 600, or an element thereof, may be performed by the object detection device 100 (e.g., by the processing unit 258 of the object detection device 100 executing computer-readable program code corresponding to the operations of the backbone CNN 400, FPN 500, and objection detection network 600).

[0066] Further, for the purpose of simplicity and ease of description, FIG. 7 will be explained with reference to detecting a single object included in an input image. However, the algorithm for performing the context-embedding, region-based object detection method according to at least some example embodiments is not limited to receiving an image including only one object, nor is the algorithm limited to detecting only one object. The input image can include several objects, and the algorithm is capable of detecting several objects of varying classes, locations and scales, concurrently.

[0067] Referring to FIG. 7, in step S710, the object detection device 100 receives an input image including an object. According to at least one example embodiment of the inventive concepts, the object detection device 100 may receive the input image as part of image data 120 received from the surveillance system 150, as is discussed above with reference to FIG. 1. After receiving the input image, the object detection device 100 may apply the received image as input to the backbone CNN 400. After step S710, the object detection device 100 proceeds to step S720.

[0068] In step S720, the object detection device 100 may generate reference feature maps. For example, the object detection device 100 may generate, using the backbone CNN 400, a plurality of reference feature maps based on the input image received in step S710.

[0069] For example, in step S720, the second to fifth convolutional layers {conv2_x, conv3_x, conv4_x, conv5_x} of the backbone CNN 400 may generate the second to fifth reference feature maps $\{C_2, C_3, C_4, C_5\}$, respectively. The reference feature maps $\{C_2, C_3, C_4, C_5\}$ may each have different sizes/scales which decrease from the second reference feature map C_2 to the fifth reference feature map C_5. After step S720, the object detection device 100 proceeds to step S730.

[0070] In step S730, the object detection device 100 may use an FPN to generate a feature pyramid including final feature maps. For example, the object detection device 100 may generate a feature pyramid including a plurality of final feature maps corresponding, respectively, to the plurality of reference feature maps generated in step S720.

[0071] For example, as is discussed above with reference to the FPN 500 illustrated in FIG. 5, in step S720, the FPN 500 may generate first to fifth final feature maps and, optionally, an additional sixth final feature map $\{P_2, P_3, P_4, P_5, P_6\}$. The first to fifth final feature maps $\{P_2, P_3, P_4, P_5\}$ may correspond, respectively, to the first to fifth reference feature maps $\{C_2, C_3, C_4, C_5\}$ generated in step S720. The sixth final feature map P_6 may be generated by the FPN 500 based on the fifth final feature map P_5, for example, by performing a stride 2 subsampling of the fifth final feature map P_5, as is discussed, for example, in *T. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection," Proc. IEEE Computer Vision and Pattern Recognition, 2017.* The final feature maps $\{P_2, P_3, P_4, P_5, P_6\}$ may each have different sizes/scales which decrease from the second final feature map P_2 to the sixth final feature map P_6. After step S730, the object detection device 100 proceeds to step S740.

[0072] In step S740, the object detection device 100 obtains a proposed region of interest (RoI or ROI), and generates one or more context RoIs.

[0073] For example, according to at least some example embodiments, the object detection device 100 may obtain a proposed RoI from an external source. Alternatively, the object detection device 100 may obtain the proposed RoI by implementing a region proposal network (RPN) based on the FPN 500, and using the FPN-based RPN to generate a proposed RoI.

[0074] For example, according to at least some example embodiments, the final feature maps P_{k0+2}, P_{k0+1}, P_{k0}, P_{k0-1}, P_{k0-2} generated by the FPN 500 as is illustrated in FIG. 5 can be used to implement the FPN-based RPN. One of ordinary skill in the art will recognize that example methods of implementing an FPN-based RPN are discussed in *T. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection," Proc. IEEE Computer Vision and Pattern Recognition, 2017.* For example, when $k_0 = 4$, the FPN 500 generates 2^{nd} through 6^{th} final feature maps P_2, P_3, P_4, P_5, and P_6. The 6^{th} final feature map P_6 may be generated based on the 5^{th} final feature map P_5 in the same manner discussed above with reference to step S730. Further, in order to generate region proposals, the FPN-based RPN may use anchors of three different aspect ratios {1:2, 1:1, 2:1} for each of the second through sixth final feature maps P_2- P_6 such that the anchors used on the 5 different final feature maps $\{P_2, P_3, P_4, P_5, P_6\}$ have 5 different areas $\{32^2, 64^2, 128^2, 256^2, 512^2\}$, respectively.

[0075] Thus, in step S740, the object detection device 100 may obtain the proposed RoI by either one of receiving the proposed RoI and generating the proposed RoI.

[0076] Further, in step S740, based on the obtained proposed RoI, the object detection device 100 may obtain one or more context RoIs by enlarging the proposed RoI. For example, FIG. 6 illustrates an input image 605, a proposed RoI 610, and first and second

context RoIs, 615A and 615B. According to at least some example embodiments of the inventive concepts, the object detection network 600 generates the first context RoI 615A by enlarging the area (i.e., w×h) of the proposed RoI 610 by a factor s1 and the object detection network 600 generates the second context RoI 615B by enlarging the area (i.e., w×h) of the proposed RoI 610 by a factor s2, where 'w' is the width of the input image 605, 'h' is the height of the input image 605, and s1 and s2 are both positive numbers greater than 1. In the example illustrated in FIG. 6, $s1 = 2^2$ and $s2 = 4^2$. Further, according to at least some example embodiments, the object detection network 600 may determine coordinates for context RoIs, which are generated by enlarging a proposed RoI, in such a manner that the context RoIs are concentric with the proposed RoI.

[0077] Further, step S740 is described as obtaining "a proposed RoI" for the purpose of simplicity and ease of description. However, the algorithm for performing the context-embedding, region-based object detection method according to at least some example embodiments is not limited to obtaining just one RoI or just one RoI at a time. For example, the object detection device 100 is capable of obtaining several RoI's of varying locations, scales and aspect ratios, concurrently, in step S740.

[0078] Further, though step S740 is described above with reference to an example scenario in which two context RoIs (i.e, two enlarged version of the proposed RoI 610) are generated, according to at least some example embodiments, any number of context RoIs (e.g., 1, 3, 5, etc.) may be generated by enlarging the proposed RoI 610. After step S740, the object detection device 100 proceeds to step S750.

[0079] In step S750, the object detection device 100 assigns the proposed RoI and the one or more context RoIs to final feature maps. For example, in step S750, the object detection device may assign the proposed RoI 610, the first context RoI 615A, and the second

context RoI 615B to final feature maps, e.g., from among the final feature maps {P_2, P_3, P_4, P_5, P_6} generated in step S730.

[0080] For example, to perform the above-reference assigning, the object detection device 100 may use the following equation:

$$k = \lfloor k_0 + \log_2(\sqrt{wh}/224) \rfloor. \qquad (1)$$

In Equation 1, 'w' represents width, 'h' represents height, and k_0 is a constant, the value of which may be set, for example, in accordance with the preferences of a designer and/or user of the object detection device 100. Additional details for setting k_0 are discussed in document [6]. In the example scenario illustrated in FIG. 6, $k_0 = 4$. It means that k_0 is corresponding to the area of 224^2, (i.e. $w \times h = 224^2$). Equation 1 is discussed, for example, in *T. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection," Proc. IEEE Computer Vision and Pattern Recognition, 2017.*

[0081] For each of the proposed RoI 610, the first context RoI 615A, and the second context RoI 615B, the object detection device 100 may apply the width 'w' and height 'h' of the RoI to Equation 1, above, to obtain the output k, and assign the RoI to the k^{th} final feature map P_k. For example, in the example scenario illustrated in FIG. 6, when the width w and height h of the proposed RoI 610 are applied to Equation 1, k = 3. Accordingly, the object detection network 600 assigns the proposed RoI 610 to the 3^{rd} final feature map P_3, as is illustrated in FIG. 6. Similarly, when the widths w and heights h of the first and second context RoIs 615A and 615B are applied to Equation 1, k = 4 and 5, respectively. Accordingly, the object detection network 600 assigns the first and second context RoIs 615A and 615B to the 4^{th} and 5^{th} final feature maps P_4 and P_5, respectively, as is illustrated in FIG. 6. After step S750, the object detection device 100 proceeds to step S760.

[0082] In step S760, the object detection device 100 extracts a set of features from each final feature map to which one of the RoIs is assigned, using RoI pooling. For example, in

step S760, the object detection network 600 embodied by the object detection device 100 may perform RoI pooling with respect to the proposed RoI 610 and the final feature map to which the proposed RoI 610 is assigned. Specifically, with respect to the proposed RoI 610, the object detection network 600 performs RoI pooling on the final feature map to which the proposed RoI 610 is assigned (i.e., the 3rd final feature map P_3) such that the features of the 3rd final feature map P_3 which fall within the proposed RoI 610 are pooled, by an RoI pooling operation, to generate a fixed-size original feature map 620. Thus, the fixed-size original feature map 620 is a set of features extracted from the 3rd final feature map P_3 based on the RoI that was originally proposed, proposed RoI 610.

[0083] Further, in step S760, the object detection network 600 forms a context branch 630 by performing RoI pooling on the first context RoI 615A and the second context RoI 615B and the final feature maps to which the first context RoI 615A and the second context RoI 615B are assigned. Specifically, with respect to the first and second context RoIs 615A and 615B, the object detection network 600 performs RoI pooling on the final feature maps to which the first and second context RoIs 615A and 615B are respectively assigned (i.e., the 4th and 5th final feature maps P_4 and P_5) such that the features of the 4th final feature map P_4 which fall within the first context RoI 615A are pooled, by an RoI pooling operation, to generate a first fixed-size context feature map 632, and the features of the 5th final feature map P_5 which fall within the second context RoI 615B are pooled, by an RoI pooling operation, to generate a second fixed-size context feature map 634. Thus, the first fixed-size context feature map 632 is a set of features extracted from the 4th final feature map P_4 based on the first context RoI 615A and the second fixed-size context feature map 634 is a set of features extracted from the 5th final feature map P_5 based on the second context RoI 615B.

[0084] According to at least some example embodiments, the RoI pooling operations discussed above with reference to step S750 may be performed by using the operations of the

RoI pooling layer discussed in document *R. Girshick, "Fast r-cnn," Computer Science, 2015.* Alternatively, according to at least some example embodiments, the RoI pooling operations discussed above with reference to step S750 may be performed by using the operations of the *RoIAlign* layer. Examples of the *RoIAlign* layer are discussed, for example, in *K. He, G. Gkioxari, P. Dollar, R. Girshick, "Mask R-CNN," In ICCV 2018.* After step S760, the object detection device 100 then proceeds to step S770.

[0085] In step S770, the object detection device 100 determines a class and/or location of the object included in the image. For example, in step S770, the object detection network 600 may perform context embedding by concatenating the first and second fixed-size context feature maps 632 and 634 to the fixed-size original feature map 620, thereby forming the concatenated feature map 625, as is shown in FIG. 6.

[0086] Further, in contrast the MS-CNN object detection sub-network 300 discussed above with respect to FIG. 3, the object detection network 600 may obtain richer context features and improved object detection results because the features included in the concatenated feature map 625 were not all extracted from the same convolutional layer or the same layer of the feature pyramid $\{P_2, P_3, P_4, P_5, P_6\}$.

[0087] As is also shown in FIG. 6, the object detection network 600 includes a squeeze and excitation (SE) block 640 and may apply the concatenated feature map 625 to the SE block 640 in order to reduce or, alternatively, eliminate noise information, for example, by recalibrating channel-wise feature responses. The SE block 640 contains two steps: squeeze and excitation. The first step is to squeeze global spatial information into a channel descriptor. This is achieved by using global average pooling to generate the channel-wise statistics. The second step is adaptive recalibration. For example, the SE block 640 may include a fully connected layer fc1 followed by a rectifier linear unit (ReLU), whose output has the dimensions $1\times1\times C'$. Further, the SE block 640 may include another fully connected layer fc2

followed by a sigmoid, the output of which has the dimensions 1×1×C (where, generally, C' =C/16) and is used to rescale the initial features of the concatenated feature map 625, for example, via channel-wise multiplication, as is shown in FIG. 6. Example structures and methods for constructing and using SE blocks are described, for example, in *Hu, Jie, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," arXiv:1709.01507, 2017.*

[0088] Next, a class and bounding box (i.e., location) of the object included in the input image 605 are determined by using the output of the SE block 640 is to generate a class probability values 660 and bounding box values 670. For example, the output of the SE block 640 may be applied to another fully connected layer 650 in order to generate class probability values (or class labels) 660 and bounding box values 670.

[0089] Object detection utilizes bounding boxes to accurately locate where objects are and assign the objects correct class labels. When image patches or frames of video are used as the input image in step S710, the class probability values 660 and bounding box values 670 are object detection results of the context-embedding, region-based object detection method discussed above with reference to FIGS. 4-7.

[0090] At least some example embodiments of the context-embedding, region-based object detection method discussed above with reference to FIGS. 4-7 can be applied to a wide variety of functions including autonomous driving system and video surveillance, as is discussed above with respect to FIG. 1. For example, referring to FIG. 1, when the camera 152 of the surveillance network 10 is placed at the entrance of a subway station, the object detection device 100 implementing the context-embedding, region-based object detection method discussed above with reference to FIGS. 4-7 can help count the pedestrian flow through the subway. In addition, when the camera 152 of the surveillance network 10 is placed in a market, the object detection device 100 implementing the context-embedding, region-based object detection method according to at least some example embodiments can

help count the number of customers in the market thereby enabling an owner or operator of the market to control a number of customers, for example, for safety reasons.

[0091] Further, the context-embedding, region-based object detection method according to at least some example embodiments includes enlarging the size of the original RoI (e.g., proposed RoI 610) in order to obtain more context information using the enlarged RoIs (e.g., first and second context RoIs 615A and 615B). Further, the enlarged RoIs are mapped to a different feature map than the original RoIs, thereby boosting the representation power of the context information obtained via the enlarged RoIs. Thus, the obtained context information is beneficial for the task of detecting small and occluded objects in the input image.

[0092] Example methods of training a CNN architecture to perform the context-embedding, region-based object detection method discussed above with reference to FIGS. 4-7 will now be discussed below in section V.

V. Example training methods

[0093] The CNN architecture for performing the context-embedding, region-based object detection method discussed above with reference to FIGS. 4-7 can be trained in accordance with known CNN training techniques, for example, to set the various values of the filters used in the various convolutional layers (e.g., the filters of the first through fifth convolutional layers conv1_x - conv5_x of the backbone CNN 400 illustrated in FIG 4.).

[0094] To begin the training stage, a proper loss function is designed. For the task of object detection, a multi-task loss function, may be used. An example of a multi-task loss function is discussed, for example, in *Lin T Y, Goyal P, Girshick R, et al., "Focal Loss for Dense Object Detection," Proc. IEEE Computer Vision and Pattern Recognition, 2017.* Further, according to at least some example embodiments, training may be performed by using the Common Object in Context (COCO) train and val-minus-minival data sets as

training data. With the technique of back-propagation, the parameters of the above-referenced filters are iteratively updated until convergence by the stochastic gradient descent (SGD) algorithm.

[0095] Example embodiments being thus described, it will be obvious that embodiments may be varied in many ways. Such variations are not to be regarded as a departure from example embodiments, and all such modifications are intended to be included within the scope of example embodiments.

WHAT IS CLAIMED IS:

1. A method of detecting an object in an image using a convolutional neural network (CNN), the method comprising:

generating, by the CNN, a plurality of reference feature maps based on the image;

generating a feature pyramid including a plurality of final feature maps corresponding, respectively, to the plurality of reference feature maps;

obtaining a proposed region of interest (ROI);

generating at least a first context ROI based on the proposed ROI such that an area of the first context ROI is larger than an area of the proposed ROI;

assigning the proposed ROI to a first final feature map from among the plurality of final feature maps;

assigning the first context ROI to a second final feature map from among the plurality of final feature maps, a size of the first final feature map being different than a size of the second final feature map;

extracting a first set of features from the first final feature map by performing an ROI pooling operation on the first final feature map using the proposed ROI;

extracting a second set of features from the second final feature map by performing an ROI pooling operation on the second final feature map using the first context ROI; and

determining, based on the first and second sets of extracted features, at least one of a location of the object with respect to the image and a class of the object.

2. The method of claim 1, wherein the feature pyramid is generated based on the plurality of reference feature maps in accordance with a feature pyramid network (FPN) architecture.

3. The method of claim 1, wherein the area of the first context ROI is 2^2 times the area of the proposed ROI.

4. The method of claim 1, further comprising:

concatenating the first and second sets of extracted features,

wherein the determining includes determining, based on the concatenated sets of extracted features, at least one of a location of the object with respect to the image and a class of the object.

5. The method of claim 4, further comprising:

applying the concatenated sets of extracted features to a squeeze-and-excitation block (SEB),

wherein the at least one of a location of the object with respect to the image and a class of the object is determined based on an output of the SEB.

6. The method of claim 1, further comprising:

generating a second context ROI based on the proposed ROI such that an area of the second context ROI is larger than an area of the first context ROI;

assigning the second context ROI to a third final feature map from among the plurality of final feature maps, a size of the third final feature map being different than the sizes of the first and second final feature maps; and

extracting a third set of features from the first final feature map by performing ROI pooling on the first final feature map using the second context ROI,

wherein the determining includes determining, based on the first, second and third sets of extracted features, at least one of the location of the object with respect to the image and the class of the object.

7. The method of claim 6, wherein the feature pyramid is generated based on the plurality of reference feature maps in accordance with a feature pyramid network (FPN) architecture.

8. The method of claim 6, wherein the area of the first context ROI is 2^2 times the area of the proposed ROI, and the area of the second context ROI is 4^2 times the area of the area of the proposed ROI.

9. The method of claim 6, further comprising:

concatenating the first, second and third sets of extracted features,

wherein the determining includes determining, based on the concatenated sets of extracted features, at least one of a location of the object with respect to the image and a class of the object.

10. The method of claim 9, further comprising:

applying the concatenated sets of extracted features to a squeeze-and-excitation block (SEB),

wherein the at least one of a location of the object with respect to the image and a class of the object is determined based on an output of the SEB.

11. A computer-readable medium comprising program instructions for causing an apparatus to perform at least the following:

generating, by a convolutional neural network (CNN), a plurality of reference feature maps based on an image that includes an object;

generating a feature pyramid including a plurality of final feature maps corresponding, respectively, to the plurality of reference feature maps;

obtaining a proposed region of interest (ROI);

generating at least a first context ROI based on the proposed ROI such that an area of the first context ROI is larger than an area of the proposed ROI;

assigning the proposed ROI to a first final feature map from among the plurality of final feature maps;

assigning the first context ROI to a second final feature map from among the plurality of final feature maps, a size of the first final feature map being different than a size of the second final feature map;

extracting a first set of features from the first final feature map by performing an ROI pooling operation on the first final feature map using the proposed ROI;

extracting a second set of features from the second final feature map by performing an ROI pooling operation on the second final feature map using the first context ROI; and

determining, based on the first and second sets of extracted features, at least one of a location of the object with respect to the image and a class of the object.

12. The computer-readable medium of claim 11, wherein the feature pyramid is generated based on the plurality of reference feature maps in accordance with a feature pyramid network (FPN) architecture.

13. The computer-readable medium of claim 11, wherein the area of the first context ROI is 2^2 times the area of the proposed ROI.

14. The computer-readable medium of claim 11, further comprising program instructions for causing an apparatus to perform at least the following:

concatenating the first and second sets of extracted features,

wherein the determining includes determining, based on the concatenated sets of extracted features, at least one of a location of the object with respect to the image and a class of the object.

15. The computer-readable medium of claim 14, further comprising program instructions for causing an apparatus to perform at least the following:

applying the concatenated sets of extracted features to a squeeze-and-excitation block (SEB),

wherein the at least one of a location of the object with respect to the image and a class of the object is determined based on an output of the SEB.

16. An apparatus comprising:

at least one processor; and

at least one memory including computer program code,

the at least one memory and the computer program code configured to, with the at least one processor, cause the apparatus at least to perform,

generating, by a convolutional neural network (CNN), a plurality of reference feature maps based on an image that includes an object;

generating a feature pyramid including a plurality of final feature maps corresponding, respectively, to the plurality of reference feature maps;

obtaining a proposed region of interest (ROI);

generating at least a first context ROI based on the proposed ROI such that an area of the first context ROI is larger than an area of the proposed ROI;

assigning the proposed ROI to a first final feature map from among the plurality of final feature maps;

assigning the first context ROI to a second final feature map from among the plurality of final feature maps, a size of the first final feature map being different than a size of the second final feature map;

extracting a first set of features from the first final feature map by performing an ROI pooling operation on the first final feature map using the proposed ROI;

extracting a second set of features from the second final feature map by performing an ROI pooling operation on the second final feature map using the first context ROI; and

determining, based on the first and second sets of extracted features, at least one of a location of the object with respect to the image and a class of the object.

17. The apparatus of claim 16, wherein the feature pyramid is generated based on the plurality of reference feature maps in accordance with a feature pyramid network (FPN) architecture.

18. The apparatus of claim 16, wherein the area of the first context ROI is twice the area of the proposed ROI.

19. The apparatus of claim 16, wherein the at least one memory and the computer program code are further configured to, with the at least one processor, cause the apparatus at least to perform:

concatenating the first and second sets of extracted features,

wherein the determining includes determining, based on the concatenated sets of extracted features, at least one of a location of the object with respect to the image and a class of the object.

20. The apparatus of claim 19, wherein the at least one memory and the computer program code are further configured to, with the at least one processor, cause the apparatus at least to perform:

applying the concatenated sets of extracted features to a squeeze-and-excitation block (SEB),

wherein the at least one of a location of the object with respect to the image and a class of the object is determined based on an output of the SEB.

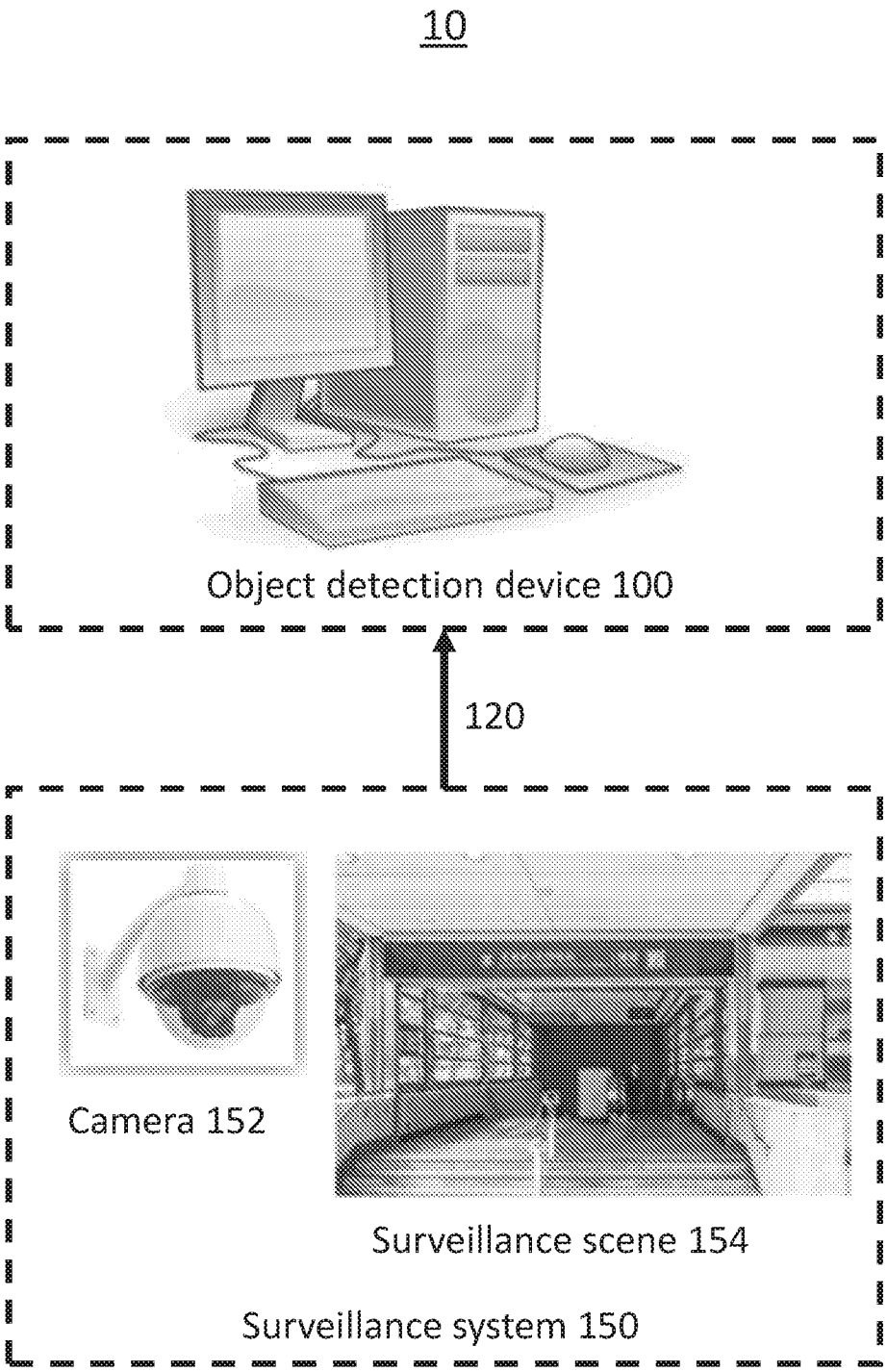
10
Object detection device 100
120
Camera 152
Surveillance scene 154
Surveillance system 150

FIG. 1

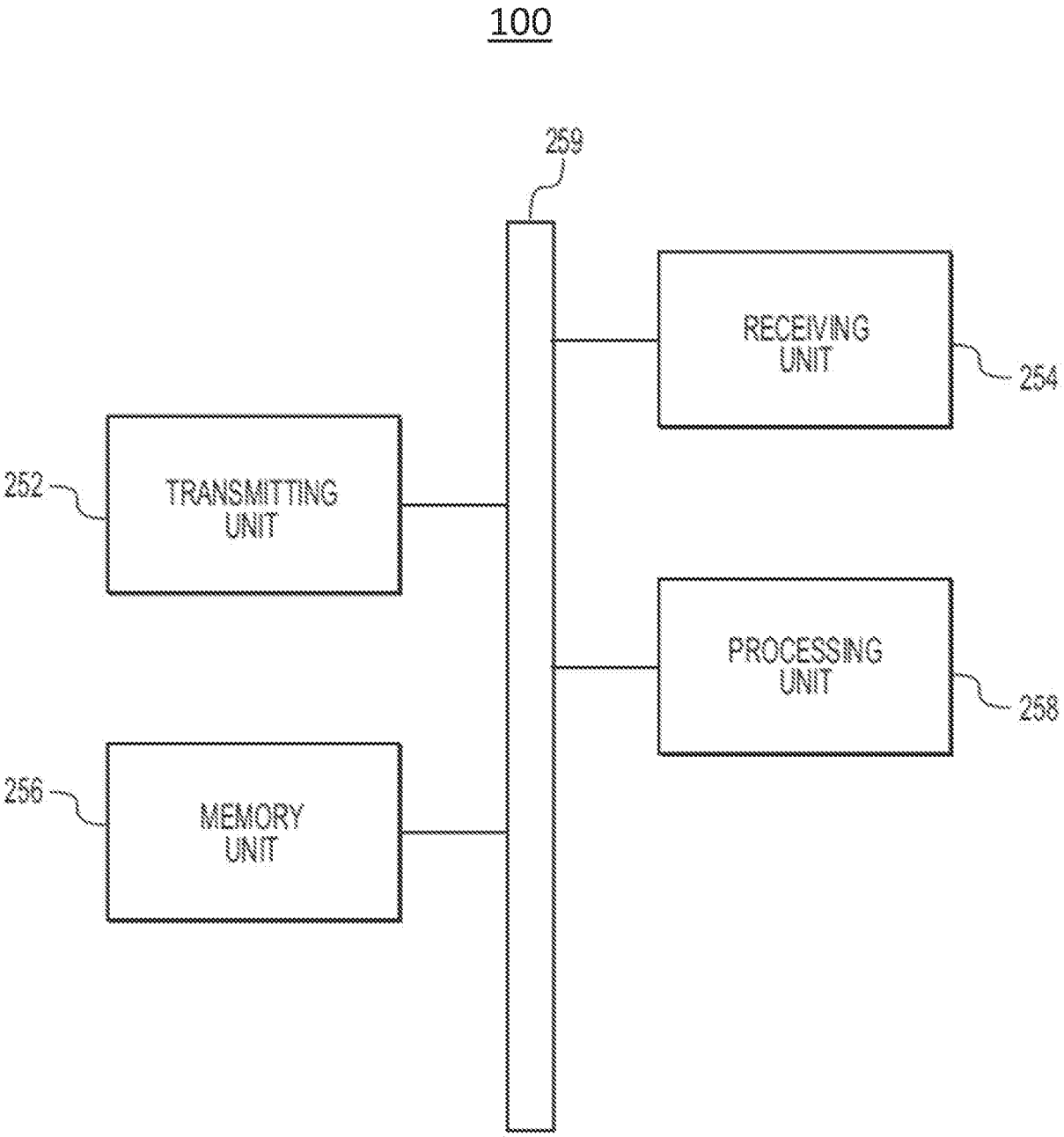
100
259
RECEIVING UNIT
254
TRANSMITTING UNIT
252
PROCESSING UNIT
258
MEMORY UNIT
256

FIG. 2

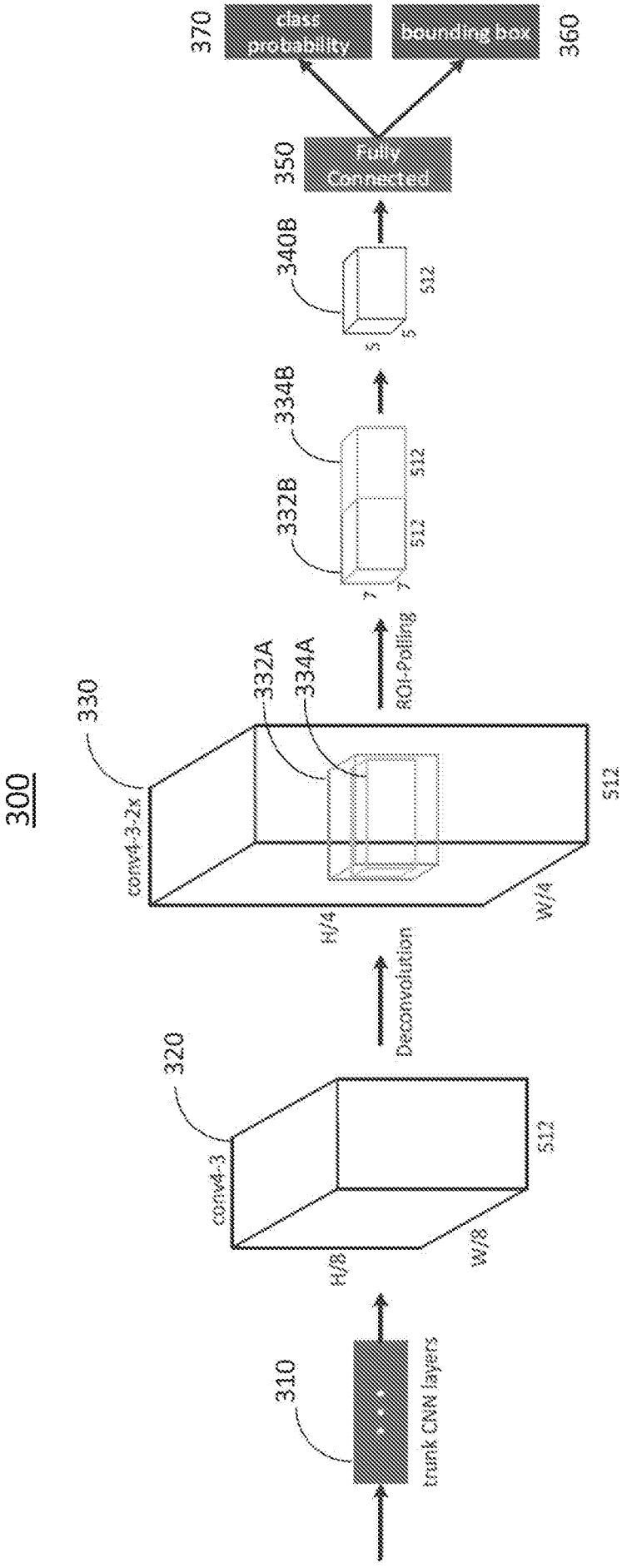
300
310
trunk CNN layers
320
conv4-3
H/8
W/8
512
Deconvolution
330
conv4-3-2x
H/4
W/4
512
332A
334A
ROI-Pooling
332B
334B
7
7
512
512
340B
5
5
512
350
Fully Connected
370
class probability
360
bounding box

FIG. 3

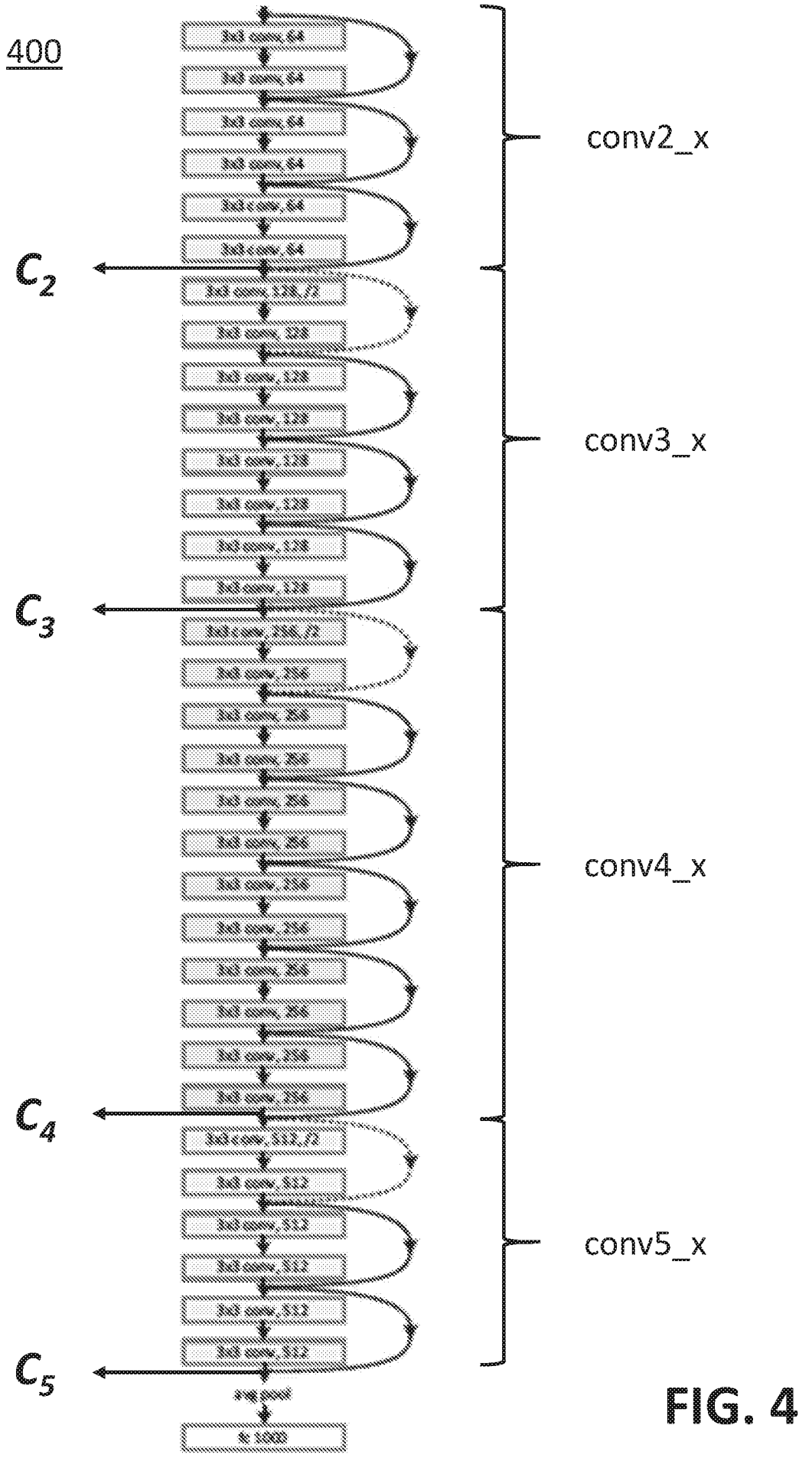

FIG. 4

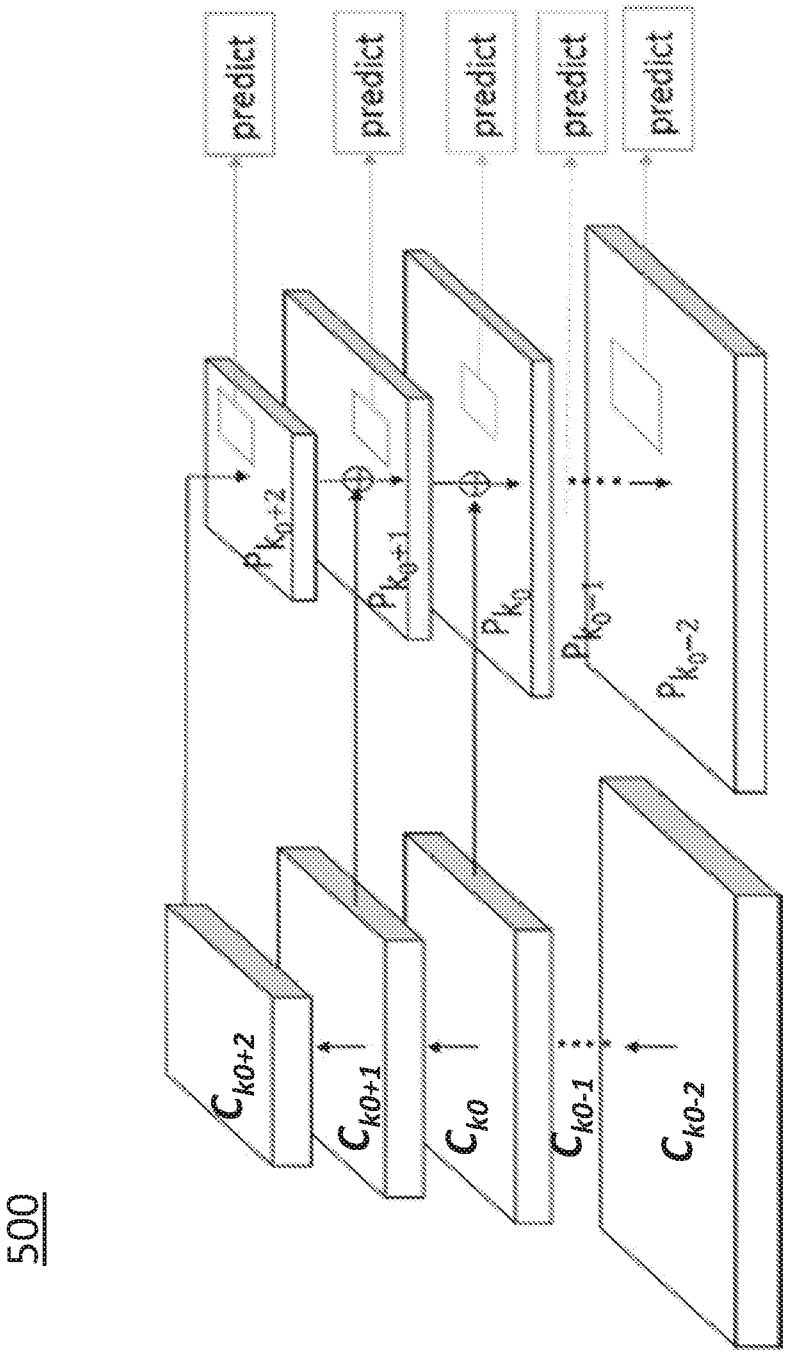
500
predict
predict
predict
predict
predict
C_{k0+2}
C_{k0+1}
C_{k0}
C_{k0-1}
C_{k0-2}
P_{k_0+2}
P_{k_0+1}
P_{k_0}
P_{k_0-1}
P_{k_0-2}

FIG. 5

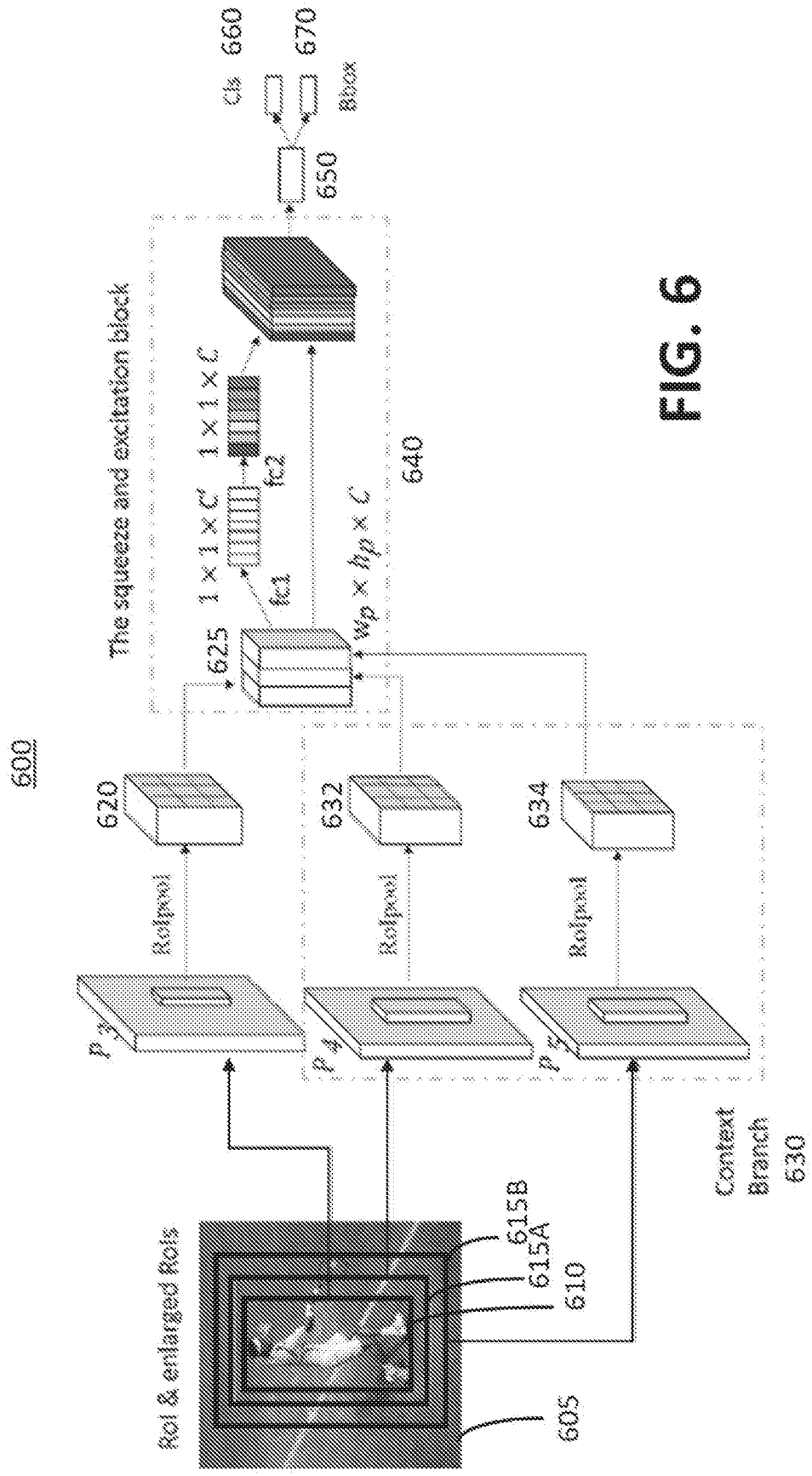
600
RoI & enlarged RoIs
605
610
615A
615B
Context Branch
630
P_3
P_4
P_5
RoIpool
620
632
634
625
The squeeze and excitation block
640
$w_p \times h_p \times C$
fc1
fc2
$1 \times 1 \times C'$
$1 \times 1 \times C$
650
Cls
660
Box
670

FIG. 6

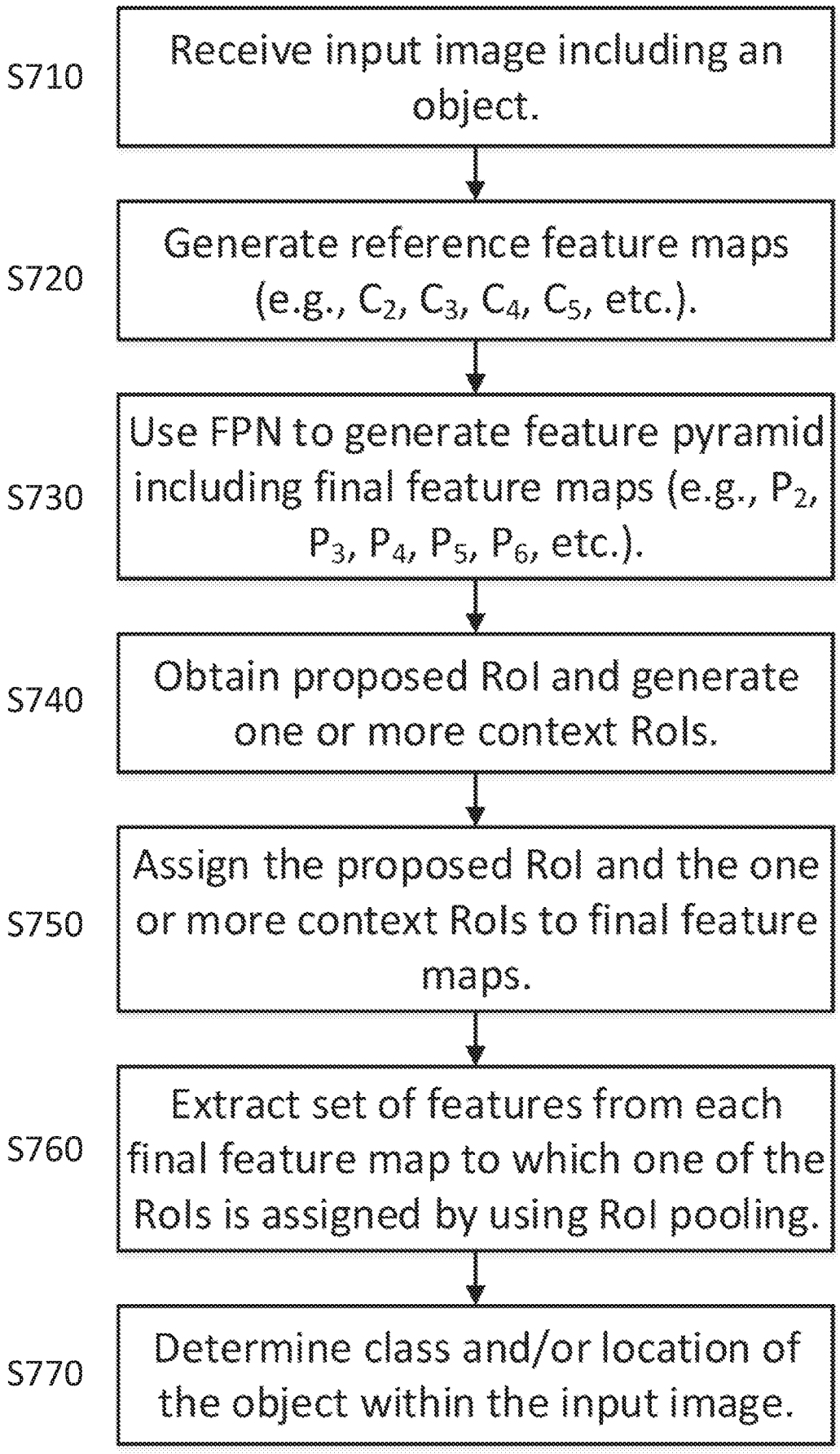
S710
Receive input image including an object.
S720
Generate reference feature maps (e.g., C_2, C_3, C_4, C_5, etc.).
S730
Use FPN to generate feature pyramid including final feature maps (e.g., P_2, P_3, P_4, P_5, P_6, etc.).
S740
Obtain proposed RoI and generate one or more context RoIs.
S750
Assign the proposed RoI and the one or more context RoIs to final feature maps.
S760
Extract set of features from each final feature map to which one of the RoIs is assigned by using RoI pooling.
S770
Determine class and/or location of the object within the input image.

FIG. 7

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2018/110023

A. CLASSIFICATION OF SUBJECT MATTER

G06T 7/246(2017.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06T

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNPAT,CNKI,WPI,EPODOC,GOOGLE: image, object, convolutional neural network, CNN, feature map, region of interest, ROI, detection

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	WO 2017015947 A1 (WANG, XIAOGANG) 02 February 2017 (2017-02-02) description, paragraphs [0005]-[0011], and figures 1-7	1-20
A	US 2018150956 A1 (INDUSTRIAL TECHNOLOGY RESEARCH INSTITUTE) 31 May 2018 (2018-05-31) the whole document	1-20
A	CN 106339680 A (BEIJING XIAOMI MOBILE SOFTWARE CO., LTD.) 18 January 2017 (2017-01-18) the whole document	1-20
A	CN 106203432 A (HANGZHOU JIANPEI TECHNOLOGY CO., LTD.) 07 December 2016 (2016-12-07) the whole document	1-20

☐ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
01 July 2019	**10 July 2019**

Name and mailing address of the ISA/CN	Authorized officer
National Intellectual Property Administration, PRC **6, Xitucheng Rd., Jimen Bridge, Haidian District, Beijing 100088** **China**	**BAI,Lushuang**
Facsimile No. **(86-10)62019451**	Telephone No. **86-(10)53961390**

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2018/110023

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
WO	2017015947	A1	02 February 2017	US	2018165548	A1	14 June 2018
				CN	108027972	A	11 May 2018
US	2018150956	A1	31 May 2018	TW	201820203	A	01 June 2018
				CN	108108732	A	01 June 2018
CN	106339680	A	18 January 2017		None		
CN	106203432	A	07 December 2016		None		