

SYSTEMS AND METHODS FOR OBJECT DETECTION**CROSS-REFERENCE TO RELATED APPLICATIONS**

[0001] This application claims priority of Chinese Patent Application No. 202110875131.4, filed on July 30, 2021, the entire contents of which are hereby incorporated by reference.

TECHNICAL FIELD

[0002] The present disclosure generally relates to image processing, and more particularly, relates to systems and methods for object detection based on a group of images of different modalities.

BACKGROUND

[0003] With the rapid development of machine learning technology and the rapid improvement of computing power, computer vision algorithms based on deep learning are widely used in tasks such as video surveillance and intelligent driving, which greatly improve the camera's ability to perceive the environment. However, in real scenes, various video capture devices may face various complex environments, especially rain, snow, night, haze, and other environments. The object identifiability on RGB images may be low and the characteristics are not obvious, and various sensors are greatly affected. At this time, the deep learning model trained from image data with specific characteristics will not be able to identify low-recognition objects well, and the camera will be severely "blind", which will bring great potential safety hazards.

[0004] Therefore, it is desirable to develop systems and methods for improving the accuracy of object detection in complex environments based on the fusion of multimodal image data.

SUMMARY

[0005] According to an aspect of the present disclosure, a method for object detection is provided. The method may include acquiring a group of images of an object. The group of images may include at least three images of different modalities. The method may further include determining a recognition result of the object based on the group of images according to an object recognition model. The recognition result may include at least one of a position of the object or a category of the object.

[0006] In some embodiments, the object recognition model may include a feature extraction submodel and an object recognition submodel. In some embodiments, the determining a recognition result of the object based on the group of images according to an object recognition model may include obtaining a fusion feature image based on the group of images according to the fusion feature extraction submodel and determining the recognition result based on the fusion feature image according to the object recognition submodel. In some embodiments, the fusion feature image may include at least one fusion feature of the group of images.

[0007] In some embodiments, the obtaining a fusion feature image may include obtaining a fusion image by fusing the group of images and determining the fusion feature image by extracting features from the fusion image.

[0008] In some embodiments, the obtaining a fusion feature image may include obtaining a plurality of feature images of the group of images by extracting image features from each image of the group of images and determining the fusion feature image by fusing the plurality of feature images.

[0009] In some embodiments, the obtaining a fusion feature image may include obtaining a plurality of feature images of the group of images by extracting image features from each image of the group of images, obtaining a preliminary fusion feature image by fusing the plurality of feature images, and determining the fusion feature image by extracting features from the preliminary fusion feature image.

[0010] In some embodiments, the fusion feature extraction submodel may include a plurality of convolution blocks connected in series. In some embodiments, each of a portion of the plurality of convolution blocks may include at least one convolution layer and a pooling layer. In some embodiments, the last convolution layer may include at least one convolution layer.

[0011] In some embodiments, the fusion feature extraction submodel may include a first convolution block, a second convolution block, a third convolution block, a fourth convolution block, and a fifth convolution block. In some embodiments, the first convolution block may include at least one convolution layer and a pooling layer. In some embodiments, the second convolution block may include at least one convolution layer and a pooling layer. In some embodiments, the third convolution block may include at least one convolution layer and a pooling layer. In some embodiments, the fourth convolution block may include at least one convolution layer and a pooling layer. In some embodiments, the fifth convolution block may include at least one convolution layer.

[0012] In some embodiments, the obtaining a fusion feature image may include obtaining a stacked matrix by concatenating matrixes corresponding to the group of images using a matrix concatenation function, determining a fused matrix by reducing a dimension of the stacked matrix, and determining the fusion feature image based on the fused matrix. In some embodiments, the matrixes corresponding to the group of images may include matrixes of a plurality of feature images of the group of images. In some embodiments, the matrixes corresponding to the group of images may include matrixes of images in the group of images.

[0013] In some embodiments, the object recognition model may include a plurality of object recognition submodels, each of the plurality of object recognition submodels corresponding to an image of a modality. In some embodiments, the obtaining a recognition result of the object based on the group of images according to an object recognition model may include obtaining a plurality of candidate recognition results and confidence scores of the plurality of candidate recognition results according to the plurality of object recognition

submodels based on the group of images and determining the recognition result based on the plurality of candidate recognition results and the confidence scores. In some embodiments, each of the plurality of candidate recognition results may correspond to one of the plurality of recognition submodels.

[0014] In some embodiments, the determining the recognition result based on the plurality of candidate recognition results and the confidence scores may include determining a weight of each of the plurality of candidate recognition results based on environment information and determining the recognition result based on the weight of each of the plurality of candidate recognition results, the plurality of candidate recognition results, and the confidence scores.

[0015] In some embodiments, the group of images may include a color image, an infrared image, and a polarization image.

[0016] According to another aspect of the present disclosure, a system for object detection is provided. The system may include at least one storage device including a set of instructions and at least one processor configured to communicate with the at least one storage device. When the set of instructions are executed, the at least one processor is configured to instruct the system to perform operations. The operations may include acquiring a group of images of an object. The group of images may include at least three images of different modalities. The operations may further include determining a recognition result of the object based on the group of images according to an object recognition model. The recognition result may include at least one of a position of the object or a category of the object.

[0017] According to yet another aspect of the present disclosure, a non-transitory computer-readable storage medium is provided, storing computer instructions, wherein when a computer reads the computer instructions in the storage medium, the computer executes a method for object detection. The method may include acquiring a group of images of an object. The group of images may include at least three images of different modalities. The method may further include determining a recognition result of the object based on the group of images according to an object recognition model. The recognition result may include at least one of a position of the object or a category of the object.

[0018] Additional features will be set forth in part in the description which follows, and in part will become apparent to those skilled in the art upon examination of the following and the accompanying drawings or may be learned by production or operation of the examples. The features of the present disclosure may be realized and attained by practice or use of various aspects of the methodologies, instrumentalities, and combinations set forth in the detailed examples discussed below.

BRIEF DESCRIPTION OF THE DRAWINGS

[0019] The present disclosure is further described in terms of exemplary embodiments. These exemplary embodiments are described in detail with reference to the drawings. The drawings are not to scale. These

embodiments are non-limiting exemplary embodiments, in which like reference numerals represent similar structures throughout the several views of the drawings, and wherein:

[0020] FIG. 1 is a schematic diagram illustrating an exemplary detection system according to some embodiments of the present disclosure;

[0021] FIG. 2 is a schematic diagram of an exemplary computing device according to some embodiments of the present disclosure;

[0022] FIG. 3 is a schematic diagram illustrating an exemplary image acquisition device according to some embodiments of the present disclosure;

[0023] FIG. 4 is a block diagram illustrating an exemplary processing device according to some embodiments of the present disclosure;

[0024] FIG. 5 is a flowchart illustrating an exemplary process for determining a recognition result of the object according to some embodiments of the present disclosure;

[0025] FIG. 6 is a block diagram illustrating an exemplary object recognition model according to some embodiments of the present disclosure;

[0026] FIG. 7 is a flowchart illustrating an exemplary process for determining the recognition result based on the fusion feature image according to some embodiments of the present disclosure;

[0027] FIG. 8 is a flowchart illustrating an exemplary process for determining the fusion feature image according to some embodiments of the present disclosure;

[0028] FIG. 9 is a schematic diagram illustrating an exemplary three-modal object recognition model according to some embodiments of the present disclosure;

[0029] FIG. 10 is a flowchart illustrating an exemplary process for determining the fusion feature image according to some embodiments of the present disclosure;

[0030] FIG. 11 is a schematic diagram illustrating an exemplary fusion feature extraction submodel according to some embodiments of the present disclosure;

[0031] FIG. 12 is a flowchart illustrating an exemplary process for determining the fusion feature image according to some embodiments of the present disclosure;

[0032] FIG. 13 is a schematic diagram illustrating an exemplary three-modal object recognition model according to some embodiments of the present disclosure;

[0033] FIG. 14 is a schematic diagram illustrating an exemplary three-modal object recognition model according to some embodiments of the present disclosure; and

[0034] FIG. 15 is a schematic flowchart illustrating an exemplary process for determining the recognition result based on the plurality of candidate recognition results and the confidence scores according to some embodiments of the present disclosure.

DETAILED DESCRIPTION

[0035] In the following detailed description, numerous specific details are set forth by way of examples in order to provide a thorough understanding of the relevant disclosure. However, it should be apparent to those skilled in the art that the present disclosure may be practiced without such details. In other instances, well-known methods, procedures, systems, components, and/or circuitry have been described at a relatively high-level, without detail, in order to avoid unnecessarily obscuring aspects of the present disclosure. Various modifications to the disclosed embodiments will be readily apparent to those skilled in the art, and the general principles defined herein may be applied to other embodiments and applications without departing from the spirit and scope of the present disclosure. Thus, the present disclosure is not limited to the embodiments shown, but to be accorded the widest scope consistent with the claims.

[0036] The terminology used herein is for the purpose of describing particular example embodiments only and is not intended to be limiting. As used herein, the singular forms “a,” “an,” and “the” may be intended to include the plural forms as well, unless the context clearly indicates otherwise. It will be further understood that the terms “comprise,” “comprises,” and/or “comprising,” “include,” “includes,” and/or “including,” when used in this specification, 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 will be understood that the term “system,” “engine,” “unit,” “module,” and/or “block” used herein are one method to distinguish different components, elements, parts, sections, or assembly of different levels in ascending order. However, the terms may be displaced by another expression if they achieve the same purpose.

[0038] Generally, the word “module,” “unit,” or “block,” as used herein, refers to logic embodied in hardware or firmware, or to a collection of software instructions. A module, a unit, or a block described herein may be implemented as software and/or hardware and may be stored in any type of non-transitory computer-readable medium or another storage device. In some embodiments, a software module/unit/block may be compiled and linked into an executable program. It will be appreciated that software modules can be callable from other modules/units/blocks or from themselves, and/or may be invoked in response to detected events or interrupts. Software modules/units/blocks configured for execution on computing devices (e.g., processor 210 as illustrated in FIG. 2) may be provided on a computer-readable medium, such as a compact disc, a digital video disc, a flash drive, a magnetic disc, or any other tangible medium, or as a digital download (and can be originally stored in a compressed or installable format that needs installation, decompression, or decryption prior to execution). Such software code may be stored, partially or fully, on a storage device of the executing computing device, for execution by the computing device. Software instructions may be

embedded in firmware, such as an EPROM. It will be further appreciated that hardware modules/units/blocks may be included in connected logic components, such as gates and flip-flops, and/or can be included in programmable units, such as programmable gate arrays or processors. The modules/units/blocks or computing device functionality described herein may be implemented as software modules/units/blocks but may be represented in hardware or firmware. In general, the modules/units/blocks described herein refer to logical modules/units/blocks that may be combined with other modules/units/blocks or divided into sub-modules/sub-units/sub-blocks despite their physical organization or storage. The description may be applicable to a system, an engine, or a portion thereof.

[0039] It will be understood that when a unit, engine, module, or block is referred to as being “on,” “connected to,” or “coupled to,” another unit, engine, module, or block, it may be directly on, connected or coupled to, or communicate with the other unit, engine, module, or block, or an intervening unit, engine, module, or block may be present unless the context clearly indicates otherwise. As used herein, the term “and/or” includes any and all combinations of one or more of the associated listed items.

[0040] These and other features, and characteristics of the present disclosure, as well as the methods of operation and functions of the related elements of structure and the combination of parts and economies of manufacture, may become more apparent upon consideration of the following description with reference to the accompanying drawings, all of which form a part of this disclosure. It is to be expressly understood, however, that the drawings are for the purpose of illustration and description only and are not intended to limit the scope of the present disclosure. It is understood that the drawings are not to scale.

[0041] The present disclosure provides mechanisms (which can include methods, systems, a computer-readable medium, etc.) for object detection. The methods provided in the present disclosure may include acquiring a group of images of an object. The group of images may include at least three images of different modalities. The method may further include determining a recognition result of the object based on the group of images according to an object recognition model. The recognition result may include at least one of a position of the object or a category of the object.

[0042] Due to using at least three images of different modalities to detect an object, the systems and methods of the present disclosure may fuse image data from different modalities or fuse object recognition results of the at least three images of different modalities to detect an object, thereby improving the accuracy of object detection, especially in complex environments.

[0043] In some embodiments, the at least three images of different modalities may include a color image, an infrared image, and a polarization image. The infrared image may be generated according to the temperature difference of the environment and not affected by lighting conditions. The polarization image may be generated by a polarization camera with a polarizing lens that can filter out bright spots and/or flares formed

on the polarization image due to polarized light and improve the image definition at such bright spots. In many complex environments, different characteristics of the infrared image and the polarization image may provide information reseeded on the color image.

[0044] In some embodiments, features in the images of different modalities may be extracted and fused through a multi-channel deep convolution neural network, which can effectively improve the detection and recognition performance (e.g., accuracy and efficiency) in complex environments. The problem of missing detection and false detection of low identified targets in a complex environment may be solved.

[0045] FIG. 1 is a schematic diagram illustrating an exemplary object detection system 100 according to some embodiments of the present disclosure. As shown, the object detection system 100 may include a server 110, a network 120, an image acquisition device 130, a user device 140, and a storage device 150.

[0046] The object detection system 100 may be applied to a variety of application scenarios for object detection. For example, the object detection system 100 may detect objects around a driving vehicle. As another example, the object detection system 100 may detect objects near an intersection or a building. The detected objects may include a vehicle, a pedestrian, a bicycle, an animal, or the like.

[0047] The server 110 may be configured to manage resources and processing data and/or information from at least one component or external data source of the object detection system 100. In some embodiments, the server 110 may be a single server or a server group. The server group may be centralized or distributed (e.g., the server 110 may be a distributed system). In some embodiments, the server 110 may be local or remote. For example, the server 110 may access information and/or data stored in the image acquisition device 130, the user device 140, and/or the storage device 150 via the network 120. As another example, the server 110 may be directly connected to the image acquisition device 130, the user device 140, and/or the storage device 150 to access stored information and/or data. In some embodiments, the server 110 may be implemented on a cloud platform. Merely by way of example, the cloud platform may include a private cloud, a public cloud, a hybrid cloud, a community cloud, a distributed cloud, an inter-cloud, a multi-cloud, or the like, or any combination thereof. In some embodiments, the server 110 may be implemented on a computing device 200 including one or more components illustrated in FIG. 2 of the present disclosure.

[0048] In some embodiments, the server 110 may include a processing device 112. The processing device 112 may process information and/or data relating to monitoring to perform one or more functions described in the present disclosure. For example, the processing device 112 may obtain a group of images of an object. The group of images may include at least three images of different modalities, for example, a color image (e.g., an RGB image), an infrared image, a polarization image (e.g., a level polarization image, a vertical polarization image, a 45° polarization image, and so on), a monochromatic light image (e.g., 324nm monochromatic light image, 525nm monochromatic light image, and so on), or the like. As another example,

the processing device 112 may determine a recognition result of the object based on the group of images according to an object recognition model. The recognition result may include at least one of a position of the object or a category of the object. As still another example, the processing device 112 may obtain a fusion feature image based on the group of images according to a fusion feature extraction submodel. The fusion feature image may include at least one fusion feature of the group of images. As still another example, the processing device 112 may determine the recognition result based on the fusion feature image according to an object recognition submodel. In some embodiments, the processing device 112 may include one or more processing devices (e.g., single-core processing device(s) or multi-core processor(s)).

[0049] In some embodiment, the server 110 may be unnecessary and all or part of the functions of the server 110 may be implemented by other components (e.g., the image acquisition device 130, the user device 140) of the object detection system 100. For example, the processing device 112 may be integrated into the image acquisition device 130 and the functions (e.g., determining the detection result of the subject 160) of the processing device 112 may be implemented by the image acquisition device 130.

[0050] The network 120 may facilitate the exchange of information and/or data for the object detection system 100. In some embodiments, one or more components (e.g., the server 110, the image acquisition device 130, the user device 140, the storage device 150) of the object detection system 100 may transmit information and/or data to other component(s) of the object detection system 100 via the network 120. For example, the server 110 may obtain images (e.g., the group of images of different modalities) from the image acquisition device 130 via the network 120. As another example, the server 110 may transmit the detection result of an object to the user device 140 via the network 120. In some embodiments, the network 120 may be any type of wired or wireless network, or a combination thereof.

[0051] In some embodiments, the network 120 may be configured to connect to each component of the object detection system 100 and/or connect the object detection system 100 and an external resource portion. The network 120 may be configured to implement communication between components of the object detection system 100 and/or between each component of the object detection system 100 and an external resource portion. In some embodiments, the network 120 may include a wired network, a wireless network, or a combination thereof. For example, the network 120 may include a cable network, a fiber network, a telecommunication network, Internet, a local area network (LAN), a wide area network (WAN), a wireless local area network (WLAN), a metropolitan area network (MAN), a public switched telephone network (PSTN), a Bluetooth network, Zigbee, near field communication (NFC), an intra-device bus, an intra-device line, a cable connection, or the like, or any combination thereof. In some embodiments, the network 120 may include a point-to-point topology structure, a shared topology structure, a centralized topology structure, or the like, or any combination thereof. In some embodiments, the network 120 may include one or more network

access points. For example, the network 120 may include a wired or wireless network access point, such as base station and/or network exchange points 120-1, 120-2, etc. One or more components of the object detection system 100 may be connected to the network 120 to exchange data and/or information through these network access points.

[0052] The image acquisition device 130 may capture a group of images of a scene. The scene may include one or more objects. The image acquisition device 130 may include at least three types of imaging apparatuses configured to obtain a group of images of different modalities. The group of images may include at least three images of different modalities, for example, a color image, an infrared image, a polarization image, a monochromatic light image, or the like. The color image may be an RGB image, an HSB image, etc. The color image may be captured by a camera, a video recorder, an image sensor, etc. The infrared image may be captured by an infrared camera, a thermal imaging sensor, an infrared image recorder, etc. The polarization image may be a level polarization image, a vertical polarization image, a 45° polarization image, etc. The polarization image may be captured by a polarization camera, a polarization video recorder, a polarization image sensor, etc. The monochromatic light image may be a 324nm monochromatic light image, a 525nm monochromatic light image, a 660nm monochromatic light image, an 880nm monochromatic light image, etc. The monochromatic light image may be captured by a monochromatic light sensor, a monochromatic light camera, a monochromatic light video recorder, etc.

[0053] In some embodiments, the group of images may be registered. The group of registered images may have the same viewing angle and overlapping area. More descriptions regarding the image acquisition device 130 may be found elsewhere in the present disclosure, e.g., FIG. 3.

[0054] In some embodiments, the image acquisition device 130 may transmit the acquired image data to one or more components (e.g., the server 110, the user device 140, the storage device 150) of the object detection system 100 via the network 120. In some embodiments, the image(s) generated by the image acquisition device 130 may be stored in the storage device 150, and/or sent to the server 110 via the network 120. In some embodiments, the image acquisition device 130 may be connected with the server 110 via the network 120. In some embodiments, the image acquisition device 130 may be connected with the server 110 directly as indicated by the dashed bidirectional arrow linking the image acquisition device 130 and the server 110 illustrated in FIG. 1.

[0055] The user device 140 may be configured to receive information and/or data from the server 110, the image acquisition device 130, and/or the storage device 150, via the network 120. For example, the user device 140 may receive the group of images from the image acquisition device 130. As another example, the user device 140 may receive the detection result of an object from the server 110. In some embodiments, the user device 140 may process information and/or data received from the server 110, the image acquisition

device 130, and/or the storage device 150, via the network 120. In some embodiments, the user device 140 may include an input device, an output device, etc. The input device may include alphanumeric and other keys that may be input via a keyboard, a touch screen (for example, with haptics or tactile feedback), a speech input, an eye-tracking input, a brain monitoring system, or any other comparable input mechanism. The input information received through the input device may be transmitted to the server 110 for further processing. Other types of input device may include a cursor control device, such as a mouse, a trackball, or cursor direction keys, etc. The output device may include a display, a speaker, a printer, or the like, or a combination thereof. In some embodiments, the user device 140 may include a display that can display information in a human-readable form, such as text, image, audio, video, graph, animation, or the like, or any combination thereof. The display of the user device 140 may include a cathode ray tube (CRT) display, a liquid crystal display (LCD), a light-emitting diode (LED) display, a plasma display panel (PDP), a three-dimensional (3D) display, or the like, or a combination thereof. In some embodiments, the user device 140 may be part of the processing device 112. In some embodiments, the user device 140 may include a mobile phone, a computer, a smart vehicle, a wearable device, or the like, or any combination thereof.

[0056] The storage device 150 may be configured to store data and/or instructions. The data and/or instructions may be obtained from, for example, the server 110, the image acquisition device 130, and/or any other component of the object detection system 100. In some embodiments, the storage device 150 may store data and/or instructions that the server 110 may execute or use to perform exemplary methods described in the present disclosure. In some embodiments, the storage device 150 may include a mass storage, a removable storage, a volatile read-and-write memory, a read-only memory (ROM), or the like, or any combination thereof. In some embodiments, the storage device 150 may be implemented on a cloud platform. Merely by way of example, the cloud platform may include a private cloud, a public cloud, a hybrid cloud, a community cloud, a distributed cloud, an inter-cloud, a multi-cloud, or the like, or any combination thereof.

[0057] In some embodiments, the storage device 150 may be connected to the network 120 to communicate with one or more components (e.g., the server 110, the image acquisition device 130, the user device 140) of the object detection system 100. One or more components of the object detection system 100 may access the data or instructions stored in the storage device 150 via the network 120. In some embodiments, the storage device 150 may be directly connected to or communicate with one or more components (e.g., the server 110, the image acquisition device 130, the user device 140) of the object detection system 100. In some embodiments, the storage device 150 may be part of other components of the object detection system 100, such as the server 110, the image acquisition device 130, or the user device 140.

[0058] It should be noted that the above description is merely provided for the purposes of illustration, and not intended to limit the scope of the present disclosure. For persons having ordinary skills in the art,

multiple variations and modifications may be made under the teachings of the present disclosure. However, those variations and modifications do not depart from the scope of the present disclosure.

[0059] FIG. 2 is a schematic diagram of an exemplary computing device according to some embodiments of the present disclosure. One or more components of the object detection system 100 (e.g., server 110, user device 140) may be implemented in computing device 200, which may be configured to perform one or more functions of the object detection system 100 (e.g., one or more functions of server 110) disclosed in this disclosure. Computing device 200 may include a bus 210, a processor 220, a read only memory (ROM) 230, a random access memory (RAM) 240, a storage device 250, an input/output port 260, and a communication interface 270.

[0060] The image acquisition device 130 may include a visible light camera, an infrared camera, a polarization camera, a monochromatic light camera, or the like, or a combination thereof. In some embodiments, the computing device 200 may be a single device. Alternatively, the computing device 200 may include a plurality of devices. One or more components of the computing device 200 may be implemented by one or more independent devices. For example, the processor 220 and the storage device 250 may be implemented in a same device. Alternatively, the processor 220 and the storage device 250 may be implemented in different devices, and the processor 220 may access the storage device 250 through wired or wireless connection (via, for example, the network 120).

[0061] Bus 210 may couple various components of computing device 200 and facilitate the transfer of data between them. Bus 210 can be any bus structure, including, for example, a memory bus or memory controller, a peripheral bus, and a local bus using any of a variety of bus architectures.

[0062] I/O port 260 may be configured to allow the transfer of data between computing device 200 and other components of the object detection system 100 (e.g., the image acquisition device 130). I/O port 260 may include a Universal Serial Bus (USB) port, a Component Object Mode (COM) port, PS/2 port, High Definition Multimedia Interface (HDMI) port, Video Graphics Array (VGA) port, or the like. Communication interface 270 may allow transfer of data between network 120 and computing device 200. Communication interface 270 may be a network interface card (NIC).

[0063] Processor 220 may include any general-purpose processor configured to perform one or more functions of the computing device 200 disclosed in this disclosure. The processor 220 may contain multiple cores or processors, cache, etc. A multicore processor can be symmetric or asymmetric. The processor 220 may essentially be a completely independent computing system with a similar structure as computing device 200. The processor 220 may receive the group of images of different modalities from the image acquisition device 130. The processor 220 may determine a recognition result of the object based on the group of images according to an object recognition model.

[0064] ROM 230, RAM 240, and storage device 250 may be configured to store data, e.g., data 252. ROM 230 may store a basic input/output (BIOS) which may provide the basic routine that helps to transfer information between devices/components within computing device 200, such as during initializing of a computer operating system. Storage device 250 may provide nonvolatile storage for data 252. Storage device 250 may connect to bus 210 through a drive interface. Storage device 250 may include a hard disk, a solid state disk (SSD), a flash memory card, a magnetic disk drive, an optical disk drive, tape drive, or the like.

[0065] ROM 230, RAM 240, and/or storage device 250 may store computer readable instructions that can be executed by processor 220 to perform one or more functions disclosed in this disclosure (e.g., the functions of server 110, image acquisition device 130, user device 140). Computer readable instructions may be packaged as a software or firmware. Data structures may include a tree structure, a linked list, a neural network, a graph structure, or the like, or their variants, or the combination thereof. Temporary data may be data generated by processor 220 when processor 220 performs computer readable instructions.

[0066] Data 252 may include raw imaging data or code implementing computer readable instructions, data structures, images, temporary data, and others. Data 252 may be transferred through bus 210 to RAM 240 before being processed by processor 220.

[0067] FIG. 3 is a schematic diagram illustrating an exemplary image acquisition device according to some embodiments of the present disclosure.

[0068] As illustrated in FIG. 3, the image acquisition device 300 may include a supporting assembly 340. The supporting assembly 340 may be configured to support and/or fix imaging apparatuses. The supporting assembly 340 may be made of aluminum alloy, titanium alloy, steel, carbon fiber, or the like, or any combination thereof. The frame fixing device 340 may include at least three fixing positions 310, 320, and 330. At least three imaging apparatuses of different types may be fixed on the three fixing positions 310, 320, and 330, respectively. The imaging apparatuses may include a visible light camera, an infrared camera, a polarization camera, a monochromatic light camera, or the like, or a combination thereof. In some embodiments, a **visible light** camera may be fixed on the fixing position 310, an infrared camera may be fixed on the fixing position 320, and a polarization camera may be fixed on the fixing position 330. In some embodiments, a **visible light** camera may be fixed on the fixing position 320, an infrared camera may be fixed on the fixing position 310, and a polarization camera may be fixed on the fixing position 330. In some embodiments, a **visible light** camera may be fixed on the fixing position 310, an infrared camera may be fixed on the fixing position 320, and a monochromatic light camera may be fixed on the fixing position 330. The types and the fixing position of the imaging apparatuses may be various and are not limited to the embodiments above.

[0069] In some embodiments, the imaging apparatuses fixed on the positions 310, 320, and 330 may be positioned according to their respective shooting axis to ensure that their shooting axes are in the same vertical plane. The shooting axis may be a straight line through a lens. In some embodiments, images captured by the imaging apparatuses may be registered, so that the imaging apparatuses can acquire three images of different modalities with the same visual angle and overlapping area in real-time. The visual angle may be an angle between a lens center point and both ends of the diagonal of the imaging plane. The overlapping area may be the same field of view. In some embodiments, the pixels in each of the images captured by the imaging apparatuses correspond to each other one by one.

[0070] The front of the supporting assembly 340 may be the shape as shown in FIG. 3 or other shapes, for example, a W shape, an H shape, an inverted A shape, etc., which is not limited here. The number of the fixing positions may be three as shown in FIG. 3, or any of the numbers greater than three, for example, four, five, seven, ten, etc. The fixing positions of the fixing device 340 may be the positions 310, 320, and 330 as shown in FIG. 3 or other positions, which are not limited here.

[0071] FIG. 4 is a block diagram illustrating an exemplary processing device according to some embodiments of the present disclosure. The processing device 400 may be exemplary processing device 112 as described in connection with FIG. 1.

[0072] As illustrated in FIG. 4, the processing device 400 may include an obtaining module 410, a determination module 420, and a model training module 430. In some embodiments, the obtaining module 410, the determination module 420, and the model training module 430 may be implemented on a processing device. In some embodiments, the obtaining module 410 and the determination module 420 may be implemented on a first processing device, while the model training module 430 may be implemented on a second processing device. The determination module 420 may obtain an object recognition model from the second processing device.

[0073] The obtaining module 410 may be configured to acquire a group of images of an object. The group of images may include at least three images of different modalities. The group of images may share the same angle and overlapping area. In some embodiments, the group of images may be obtained from the image acquisition device 130, the storage device 150, or any other storage device. In some embodiments, the group of images may include a color image, an infrared image, a polarization image, a monochromatic light image, or the like.

[0074] In some embodiments, the imaging apparatuses may be fixed on a supporting assembly (e.g., the supporting assembly 340 as shown in FIG. 3). In some embodiments, the imaging apparatuses may be positioned according to their respective shooting axis to ensure that their shooting axes are in the same vertical plane. In some embodiments, the obtaining module 410 may register images captured by the imaging

apparatuses to obtain the group of images, so that the at least three images of different modalities may have the same angle and overlapping area in real-time.

[0075] The determination module 420 may be configured to determine a recognition result of the object based on the group of images according to an object recognition model. The recognition result may indicate one or more characteristics of the object. For example, the recognition result may include at least one of a position of the object or a category of the object. In some embodiments, the determination module 420 may obtain the object recognition model from the model training module 430. In some embodiments, the determination module 420 may obtain the object recognition model from the storage device 150 or 250. In some embodiments, the determination module 420 may input the group of images into the object recognition model. The object recognition model may output the recognition result of the object.

[0076] The model training module 430 may be configured to determine the object recognition model.

[0077] In some embodiments, the object recognition model may be a trained machine learning model. In some embodiments, the object recognition model may be constructed based on a neural network model. The neural network model may include a convolutional neural network (CNN) model, a deep convolutional neural network (DCNN) model, a recurrent neural network (RNN) model, a backpropagation (BP) neural network model, a radial basis function (RBF) neural network model, a residual neural network model, etc.

[0078] In some embodiments, the object recognition model may be generated by training a preliminary object recognition model based on training samples. In some embodiments, the model training module 430 may be configured to train the preliminary object recognition model based on the training samples.

[0079] In some embodiments, the object recognition model may be constructed based on a neural network model. In some embodiments, the training of the preliminary object recognition model (e.g., the neural network model) may be performed based on the back-propagation algorithm. The preliminary object recognition model (e.g., the neural network model) may be regarded as a calculation diagram composed of nodes, and the parameters are updated layer by layer from back to front.

[0080] It should be noted that the above description is merely provided for the purposes of illustration, and not intended to limit the scope of the present disclosure. For persons having ordinary skills in the art, multiple variations or modifications may be made under the teachings of the present disclosure. However, those variations and modifications do not depart from the scope of the present disclosure. In some embodiments, the processing device 400 may include one or more additional modules, such as a storage module (not shown) for storing data.

[0081] FIG. 5 is a flowchart illustrating an exemplary process for determining a recognition result of the object according to some embodiments of the present disclosure. In some embodiments, a process 500 may be implemented as a set of instructions (e.g., an application) stored in the storage device 150 or storage device

250. The processing device 400, the processor 220, and/or the processing device 112 may execute the set of instructions, and when executing the instructions, the processing device 400, the processor 220, and/or the processing device 112 may be configured to perform the process 500. The operations of the illustrated process presented below are intended to be illustrative. In some embodiments, the process 500 may be accomplished with one or more additional operations not described, and/or without one or more of the operations discussed. Additionally, the order of the operations of the process 500 illustrated in FIG. 5 and described below is not intended to be limiting.

[0082] In 510, the processing device 400 (e.g., the obtaining module 410) may acquire a group of images of an object. The group of images may include at least three images of different modalities. The group of images may share the same angle and overlapping area. In some embodiments, the group of images may be obtained from the image acquisition device 130, the storage device 150, or any other storage device.

[0083] In some embodiments, the group of images may include a color image, an infrared image, a polarization image, a monochromatic light image, or the like. The color image may be an RGB image, an HSB image, etc. The color image may be captured by a camera, a video recorder, an image sensor, etc. The infrared image may be captured by an infrared camera, a thermal imaging sensor, an infrared image recorder, etc. The polarization image may be a level polarization image, a vertical polarization image, a 45° polarization image, etc. The polarization image may be captured by a polarization camera, a polarization video recorder, a polarization image sensor, etc. The monochromatic light image may be a 324nm monochromatic light image, a 525nm monochromatic light image, a 660nm monochromatic light image, an 880nm monochromatic light image, etc. The monochromatic light image may be captured by a monochromatic light sensor, a monochromatic light camera, a monochromatic light video recorder, etc.

[0084] In some embodiments, the group of images may include a color image, an infrared image, and a polarization image. In some embodiments, the group of images may include a color image, an infrared image, a polarization image, and a monochromatic light image. In some embodiments, the group of images may include a color image, an infrared image, and a monochromatic light image.

[0085] In some embodiments, the processing device 400 (e.g., the obtaining module 410) may acquire the group of images indicating an environment in a physical area. In some embodiments, the environment may be an environment around a driving vehicle, an environment near an intersection or a building, etc. The physical area may include one or more objects, for example, a vehicle, a pedestrian, a bicycle, an animal, etc.

[0086] In some embodiments, the imaging apparatuses may be fixed on a supporting assembly (e.g., the supporting assembly 340 as shown in FIG. 3). In some embodiments, the imaging apparatuses may be positioned according to their respective shooting axis to ensure that their shooting axes are in the same vertical plane. In some embodiments, the obtaining module 410 may register images captured by the imaging

apparatuses to obtain the group of images, so that the at least three images of different modalities may have the same angle and overlapping area in real-time.

[0087] In 520, the processing device 400 (e.g., the determination module 420) may determine a recognition result of the object based on the group of images according to an object recognition model. The recognition result may indicate one or more characteristics of the object. For example, the recognition result may include at least one of a position of the object or a category of the object.

[0088] In some embodiments, the processing device 400 may obtain the object recognition model from the model training module 430. In some embodiments, the processing device 400 may obtain the object recognition model from the storage device 150 or 250.

[0089] In some embodiments, the processing device 400 may input the group of images into the object recognition model. The object recognition model may output the recognition result of the object. In some embodiments, the recognition result of the object may include a bounding box corresponding to the object and/or the category of the object presented in one of the group of images. The bounding box corresponding to the object may enclose the object and indicate the position of the object. In some embodiments, the recognition result of the object may include a highlighted object and/or the category of the object presented in one of the group of images. In some embodiments, the recognition result of the object may include coordinates of the object and/or the category of the object. In some embodiments, the recognition result of the object may further include at least one of the moving direction of the object, the velocity of the object, the distance between the object and the imaging apparatuses, and the acceleration of the object. More descriptions regarding determining a recognition result of the object may be found elsewhere in the present disclosure, e.g., FIGs. 7-14.

[0090] In some embodiments, the object recognition model may be a trained machine learning model. In some embodiments, the object recognition model may be constructed based on a neural network model. The neural network model may include a convolutional neural network (CNN) model, a deep convolutional neural network (DCNN) model, a recurrent neural network (RNN) model, a backpropagation (BP) neural network model, a radial basis function (RBF) neural network model, a residual neural network model, etc.

[0091] In some embodiments, the object recognition model may be generated by training a preliminary object recognition model based on training samples. In some embodiments, the model training module 430 may be configured to train the preliminary object recognition model based on the training samples.

[0092] In some embodiments, the training samples may include a plurality of groups of sample images. Each group of the plurality of groups of sample images may include at least three sample images of different modalities. The group of sample images may have the same angle and overlapping area. In some embodiments, each sample image may include one or more objects. The training samples may further

include one or more labels corresponding to the one or more objects in each sample image. Each of the one or more labels may include the position and the category of each of the one or more objects. In some embodiments, the one or more labels may be annotated by a user. In some embodiments, the training samples may further include environment information. In some embodiments, the environment information may be determined based on the group of sample images. In some embodiments, the environment information may be determined based on inputted information, for example, time information, weather information, etc., inputted by the obtaining module 410. In some embodiments, the plurality of groups of sample images may be obtained from the imaging apparatuses (e.g., the image acquisition device 130), the storage device 150, or any other storage device.

[0093] In some embodiments, the model training module 430 may obtain the training samples and the preliminary object recognition model. The model training module 430 may input the training samples into the preliminary object recognition model. In some embodiments, each of the plurality of groups of sample images may be used as an input, and each of the one or more labels may be used as a desired output of one of the training samples, and the object recognition model may be generated by training the preliminary object recognition model accordingly.

[0094] In some embodiments, the object recognition model may be constructed based on a neural network model. In some embodiments, the training of the preliminary object recognition model (e.g., the neural network model) may be performed based on the back-propagation algorithm. The preliminary object recognition model (e.g., the neural network model) may be regarded as a calculation diagram composed of nodes, and the parameters are updated layer by layer from back to front.

[0095] It should be noted that the above description regarding the process 500 is merely provided for the purposes of illustration, and not intended to limit the scope of the present disclosure. For persons having ordinary skills in the art, multiple variations or modifications may be made under the teachings of the present disclosure. However, those variations and modifications do not depart from the scope of the present disclosure. In some embodiments, one or more operations may be omitted and/or one or more additional operations may be added. For example, the process 500 may further include an operation to display the **recognition result of the object on the interface of the user device 140**. The process 500 may further include transmitting the **recognition result of the object** to a public security bureau data center via the network 120.

[0096] FIG. 6 is a block diagram illustrating an exemplary object recognition model according to some embodiments of the present disclosure. As illustrated in FIG. 6, the object recognition model 600 may include a fusion feature extraction submodel 610 and an object recognition submodel 620.

[0097] The fusion feature extraction submodel 610 may be configured to determine a fusion feature image based on the group of images. The fusion feature image may include at least one fusion feature of the group of images.

[0098] In some embodiments, the processing device 400 (e.g., the determination module 420) may input the group of images into the fusion feature extraction submodel 610. The fusion feature extraction submodel 610 may output the fusion feature image. In some embodiments, the fusion feature extraction submodel may include the feature extraction part of a visual geometry group (VGG) network, a residual network (ResNet), an Inception network, etc.

[0099] In some embodiments, the fusion feature image may be obtained by extracting features from a fusion image. The fusion image may be obtained by fusing the group of images. In some embodiments, the fusion feature image may be obtained by extracting features from a preliminary fusion feature image. The preliminary fusion feature image may be obtained by fusing a plurality of feature images that are obtained by extracting image features from each image of the group of images. In some embodiments, the fusion feature image may be obtained by fusing a plurality of feature images. The plurality of feature images may be obtained by extracting image features from each image of the group of images.

[0100] The object recognition submodel 620 may be configured to determine the recognition result based on the fusion feature image. In some embodiments, the processing device 400 (e.g., the determination module 420) may input the fusion feature image to the object recognition submodel. The object recognition submodel may output the recognition result. In some embodiments, the object recognition submodel may include object recognition networks that is at least a part of the Faster region-convolutional neural network (R-CNN) or You Only Look Once (YOLO) network.

[0101] FIG. 7 is a flowchart illustrating an exemplary process for determining the recognition result based on the fusion feature image according to some embodiments of the present disclosure. In some embodiments, a process 700 may be implemented as a set of instructions (e.g., an application) stored in the storage device 150 or storage device 250. The processing device 400, the processor 220, and/or the processing device 112 may execute the set of instructions, and when executing the instructions, the processing device 400, the processor 220, and/or the processing device 112 may be configured to perform the process 700. The operations of the illustrated process presented below are intended to be illustrative. In some embodiments, the process 700 may be accomplished with one or more additional operations not described, and/or without one or more of the operations discussed. Additionally, the order of the operations of the process 700 illustrated in FIG. 7 and described below is not intended to be limiting.

[0102] In 710, the processing device 400 (e.g., the determination module 420) may obtain a fusion feature image based on the group of images according to a fusion feature extraction submodel 610. The fusion

feature image may include at least one fusion feature of the group of images. Detailed description regarding the group of images please refer to elsewhere in the present disclosure, for example, the description of operation 510 in FIG. 5.

[0103] In some embodiments, the processing device 400 (e.g., the determination module 420) may input the group of images into the fusion feature extraction submodel 610. The fusion feature extraction submodel 610 may output the fusion feature image. In some embodiments, the fusion feature extraction submodel 610 may include the feature extraction part of a visual geometry group (VGG) network, a residual network (ResNet), an Inception network, etc.

[0104] In some embodiments, the fusion feature image may be obtained by extracting features from a fusion image. The fusion image may be obtained by fusing the group of images. In some embodiments, the fusion feature image may be obtained by extracting features from a preliminary fusion feature image. The preliminary fusion feature image may be obtained by fusing a plurality of feature images that are obtained by extracting image features from each image of the group of images. In some embodiments, the fusion feature image may be obtained by fusing a plurality of feature images. The plurality of feature images may be obtained by extracting image features from each image of the group of images.

[0105] In some embodiments, the fusion feature extraction submodel 610 may include a plurality of convolution blocks. Merely by way of example, the fusion feature extraction submodel 610 may include five convolution blocks. In some embodiments, each of the plurality of convolutional blocks may include at least one convolution layer and a pooling layer. In some embodiments, each of a portion of the plurality of convolution blocks may include at least one convolution layer and a pooling layer, and the last convolution layer may merely include at least one convolution layer. In some embodiments, one of the groups of images processed by each convolution layer may be processed by a rectified linear unit (ReLU) activation function. The image processed by the at least one convolution layer may be inputted into the pooling layer to reduce the size of the image. In some embodiments, one of the plurality of convolution blocks may include two or three convolution layers. A convolution layer may use a plurality of convolution kernels to convolute an image inputted into the convolution layer. In some embodiments, the size of the convolution kernel may be $3*3*d$, wherein d indicates a depth of the inputted image. In some embodiments, the size of the convolution kernel may be $5*5*d$, wherein d indicates a depth of the inputted image. In some embodiments, the size of the convolution kernel may be $2*2*d$, wherein d indicates a depth of the inputted image. In some embodiments, the size of the pooling layer may be $2*2$. An image of size $m*m*n$ may be inputted into the pooling layer, and the pooling layer may output an image of size $m/2*m/2*n$ after processing the image. The size of the convolution kernel and pooling layer may be various and are not limited to the embodiments above.

[0106] In some embodiments, the fusion feature extraction submodel 610 may further include a fusion unit. The fusion unit may include a concatenation layer and a convolution layer. An image may be represented as a matrix. The concatenation layer may be used to concatenate a plurality of matrixes corresponding to the group of images to obtain a stacked matrix. The convolution layer may be used to reduce the depth of the stacked matrix.

[0107] In some embodiments, the concatenation layer may include a matrix concatenation function. The concatenation layer may concatenate the plurality of matrixes corresponding to the group of images using the matrix concatenation function to obtain the stacked matrix. In some embodiments, the concatenation layer may concatenate the plurality of matrixes of the plurality of feature images of the group of images. At least one of the plurality of feature images of the group of images may be generated by any one of the plurality of convolution blocks. In some embodiments, at least one of the plurality of feature images of the group of images may be generated by the fourth block of the plurality of convolution blocks. In some embodiments, at least one of the plurality of feature images of the group of images may be generated by the fifth block of the plurality of convolution blocks. In some embodiments, the concatenation layer may concatenate the plurality of matrixes of the group of images.

[0108] In some embodiments, the convolution layer may include a convolution kernel of size 1×1 . The convolution layer may be used to reduce the dimension of the stacked matrix while the length and width of the stacked matrix are fixed. Merely by way of example, an image of size $m \times m \times n$ may be inputted into the convolution layer, and the convolution layer may output an image of size $m \times m \times n/3$ after processing the image of size $m \times m \times n$.

[0109] In some embodiments, the environment information may be inputted into the fusion unit. The fusion unit may fuse the feature images based on the environment information. In some embodiments, the fusion unit may determine a weight for each of the group of images or feature images extracted from the group of images. Merely by way of example, if the weather is good (e.g., a sunny day, a cloudy day), the weight for the color image may exceed the weight for the polarization image, and the weight for the polarization image may exceed the weight for the infrared image; if the weather is bad (e.g., rainy, snowy, foggy days) or in the night, the weight for the infrared image may exceed the weight for the color image, and the weight for the color image may exceed the weight for the polarization image. In some embodiments, the weight may be determined in the training process and may be determined based on the model itself. In some embodiments, the weight factor may be set up by the user.

[0110] In some embodiments, the fusion unit may be positioned before the plurality of convolution blocks. The group of images may be inputted into the fusion unit. The fusion unit may fuse the group of images to obtain a fusion image. Then the fusion image may be inputted into the plurality of convolution blocks.

connected in series. The plurality of convolution blocks may extract features from the fusion image to obtain a fusion feature image.

[0111] In some embodiments, the fusion unit may be positioned among the plurality of convolution blocks. For example, the fusion unit may be positioned between the fourth convolution block and the fifth convolution block. As another example, the fusion unit may be positioned between the first convolution block and the second convolution block. In some embodiments, the plurality of convolution blocks may include a first portion of convolution blocks and a second portion of convolution blocks. The first portion of convolution blocks may include a plurality of convolution networks in parallel. In some embodiments, a convolution network may include one or more convolution blocks. In some embodiments, there may be two or more convolution blocks connected in series in the convolution network. Each of the plurality of convolution networks may correspond to each image of the group of images. Merely by way of example, the group of images may include three images of different modalities, and the plurality of convolution networks may include three convolution networks in parallel. Each image of the group of images may be inputted into one of the plurality of convolution networks corresponding to the image. The one of the plurality of convolution networks may extract image features from the image and output a feature image of the image. The plurality of feature images of the group of images may be inputted into the fusion unit. The fusion unit may fuse the plurality of feature images to obtain a preliminary fusion feature image. Then the preliminary fusion feature image may be inputted into the second portion of convolution blocks. The second portion of convolution blocks may extract features from the preliminary fusion feature image and output the fusion feature image. In some embodiments, the second portion of convolution blocks may include one or more convolution blocks connected in series.

[0112] In some embodiments, the fusion unit may be positioned after the plurality of convolution blocks. The plurality of convolution blocks may include a plurality of convolution subnetworks in parallel. Each of the plurality of convolution subnetworks may include one or more convolution blocks connected in series if there are two or more convolution blocks. Each of the plurality of convolution subnetworks may correspond to each image of the group of images. Merely by way of example, the group of images may include three images of different modalities, and the plurality of convolution blocks may include three convolution subnetworks in parallel. Each image of the group of images may be inputted into one of the plurality of convolution subnetworks corresponding to the image. Each of the plurality of convolution subnetworks may extract image features from one of the group of images and output a feature image of the image. The plurality of feature images of the group of images may be inputted into the fusion unit. The fusion unit may fuse the plurality of feature images and output the fusion feature image.

[0113] In some embodiments, each of the plurality of convolution subnetworks before the fusion unit may be trained independently. In some embodiments, the plurality of convolution subnetworks before the fusion unit may be jointly trained with the rest part of the object recognition model.

[0114] In 720, the processing device 400 (e.g., the determination module 420) may determine the recognition result based on the fusion feature image according to the object recognition submodel.

[0115] In some embodiments, the processing device 400 (e.g., the determination module 420) may input the fusion feature image to the object recognition submodel. The object recognition submodel may output the recognition result. In some embodiments, the object recognition submodel may include object recognition networks that is at least a part of the Faster region-convolutional neural network (R-CNN) or You Only Look Once (YOLO) network.

[0116] In some embodiments, the object recognition submodel may include an RPN (region proposal networks) layer, an ROI (region of interest) pooling layer, and a classification layer. The RPN layer may include a softmax function, a bounding box regression, and a proposal layer. The RPN layer may use a plurality of anchor boxes to generate a plurality of region proposals in the fusion feature image. The plurality of anchor boxes may be of different sizes and/or shapes. The softmax function may be used to judge whether the plurality of proposal regions are positive or negative. The bounding box regression may be used to obtain the offsets of the plurality of proposal regions and correct the plurality of proposal regions based on the offsets. The proposal layer may be used to obtain one or more final proposal regions based on positive proposal regions and the corresponding offsets of the positive proposal regions of the plurality of proposal regions, while proposal regions that are too small and beyond a bounding box among the plurality of proposal regions are eliminated. The bounding box corresponding to the proposal region may be used to enclose the proposal region. The ROI pooling layer may be used to extract one or more proposal feature maps from the fusion feature image based on the one or more final proposal regions. The one or more proposal feature maps may be of the same size. The ROI pooling layer may change the size of the one or more final proposal regions to generate the one or more proposal feature maps of the same size. The classification layer may include two full connection layers and a softmax function. In some embodiments, the image processed by each full connection layer may be processed by the ReLU activation function. The two full connection layers and the softmax function may be used to calculate the specific category (e.g., a person, a dog, a vehicle, a TV, etc.) of the one or more proposed feature maps and output probability vectors corresponding to the one or more proposed feature maps. In some embodiments, the classification layer may further include a bounding box regression. The bounding box regression may be used to obtain the position offset of the one or more proposal feature maps to regress a more accurate bounding box for positioning the object.

[0117] In some embodiments, the object recognition submodel may include two full connection layers. In some embodiments, one of the group of images processed by each full connection layer may be processed by the ReLU activation function. The two full connection layers may output one or more object bounding boxes and the corresponding probability of the category based on the fusion feature image.

[0118] In some embodiments, the setting of a configuration file of object recognition model (i.e., the network) may be described as follows: the learning rate of training may be 0.001, and the learning rate may be set to 0.0001 after 50000th steps of iteration. The aspect ratio of the anchor of the RPN network may be [1, 2, 0.5], and the scale may be [8, 16, 32]. An input image may need to be standardized for model training. The pixel mean value of the color image may be [85.38, 107.37, 103.21], the pixel mean value of the infrared image may be [99.82, 53.63, 164.85], and the pixel mean value of the polarization image may be [79.68, 88.75, 94.55]. The optimizer used for model training may be a momentum optimizer, and the momentum super parameter may be set to 0.9. The model training iteration may be 100000 steps.

[0119] It should be noted that the above description regarding the process 700 is merely provided for the purposes of illustration, and not intended to limit the scope of the present disclosure. For persons having ordinary skills in the art, multiple variations or modifications may be made under the teachings of the present disclosure. However, those variations and modifications do not depart from the scope of the present disclosure.

[0120] FIG. 8 is a flowchart illustrating an exemplary process for determining the fusion feature image according to some embodiments of the present disclosure. In some embodiments, a process 800 may be implemented as a set of instructions (e.g., an application) stored in the storage device 150 or storage device 250. The processing device 400, the processor 220, and/or the processing device 112 may execute the set of instructions, and when executing the instructions, the processing device 400, the processor 220, and/or the processing device 112 may be configured to perform the process 800. The operations of the illustrated process presented below are intended to be illustrative. In some embodiments, the process 800 may be accomplished with one or more additional operations not described, and/or without one or more of the operations discussed. Additionally, the order of the operations of the process 800 illustrated in FIG. 8 and described below is not intended to be limiting.

[0121] In 810, the processing device 400 (e.g., the determination module 420) may obtain a fusion image by fusing the group of images. Detailed description regarding the group of images please refer to elsewhere in the present disclosure, for example, the description of 510 in FIG. 5.

[0122] In some embodiments, the group of images may be inputted into a fusion unit. The fusion unit may fuse the group of images to obtain the fusion image. In some embodiments, the fusion unit may include a concatenation layer and a convolution layer. The concatenation layer may be used to concatenate a plurality

of matrixes of the group of images to obtain a stacked matrix. The convolution layer may be used to reduce the depth of the stacked matrix to output the fusion image.

[0123] In some embodiments, the concatenation layer may include a matrix concatenation function. The concatenation layer may concatenate a plurality of matrixes of the group of images using the matrix concatenation function to obtain the stacked matrix. In some embodiments, the convolution layer may further include a convolution kernel of size 1×1 . The convolution layer may be used to reduce the dimension of the stacked matrix while the length and width of the stacked matrix are fixed. Merely by way of example, an image of size $m \times m \times n$ may be inputted into the convolution layer, and the convolution layer may output an image of size $m \times m \times n/3$ after processing the image of size $m \times m \times n$.

[0124] In 820, the processing device 400 (e.g., the determination module 420) may determine the fusion feature image by extracting features from the fusion image.

[0125] In some embodiments, the fusion image may be inputted into an image feature extracted network, which may include the feature extraction part of a VGG network, a ResNet, or an Inception network.

[0126] In some embodiments, the fusion image may be inputted into a plurality of convolution blocks connected in series. The plurality of convolution blocks may extract features from the fusion image to obtain the fusion feature image. Merely by way of example, the plurality of convolution blocks may include five convolution blocks. In some embodiments, each of the plurality of convolutional blocks may include at least one convolution layer and a pooling layer. In some embodiments, each of a portion of the plurality of convolution blocks may include at least one convolution layer and a pooling layer, and the last convolution layer may merely include at least one convolution layer. An image processed by the at least one convolution layer may be inputted into the pooling layer to reduce the size of the image. In some embodiments, the convolution block may include two or three convolution layers. The convolution layer may use a plurality of convolution kernels to convolute an image inputted into the convolution layer. In some embodiments, the size of the convolution kernel may be $3 \times 3 \times d$, wherein d indicates a depth of the inputted image. In some embodiments, the size of the convolution kernel may be $5 \times 5 \times d$, wherein d indicates a depth of the inputted image. In some embodiments, the size of the convolution kernel may be $2 \times 2 \times d$, wherein d indicates a depth of the inputted image. In some embodiments, the size of the pooling layer may be 2×2 . An image of size $m \times m \times n$ may be inputted into the pooling layer, and the pooling layer may output an image of size $m/2 \times m/2 \times n$ after processing the image of size $m \times m \times n$. The size of the convolution kernel and pooling layer may be various and are not limited to the embodiments above.

[0127] Merely by way of example, the process 800 may be executed by the object recognition model 900 as shown in FIG. 9. The object recognition model 900 may include a fusion unit, a CNN feature extraction network, and a Faster-CNN classification regression.

[0128] The group of images may include three images of different modalities and may be inputted into the fusion unit. The group of images may include a color image, an infrared image, and a polarization image. The fusion unit may include a fusion module and a convolution layer. The fusion module may include a matrix concatenation function. The convolution layer may include a convolution kernel of size 1×1 . The fusion module may concatenate three matrixes of the group of images in the depth dimension to obtain a stacked matrix. The convolution layer may reduce the depth of the stacked matrix to output the fusion image. Merely by way of example, each image of the group of images may have a size of $m \times m \times n$, the stacked matrix may have a size of $m \times m \times 3n$, and the fusion image may have a size of $m \times m \times n$.

[0129] The fusion image may be inputted into the CNN feature extraction network and the fusion feature image may be outputted. The CNN feature extraction network may include VGG-16 as a backbone feature extraction network. The CNN feature extraction network may include five convolution blocks. The first convolution block may include two convolution layers and a pooling layer. The second convolution block may include two convolution layers and a pooling layer. The third convolution block may include three convolution layers and a pooling layer. The fourth convolution block may include three convolution layers and a pooling layer. The fifth convolution block may include three convolution layers. In some embodiments, the two convolution layers in the first convolution block may have convolution kernels of $64 \times 3 \times 3 \times 3$ and $64 \times 3 \times 3 \times 64$. In some embodiments, the two convolution layers in the second convolution block may have convolution kernels of $128 \times 3 \times 3 \times 64$ and $128 \times 3 \times 3 \times 128$. In some embodiments, the three convolution layers in the third convolution block may have convolution kernels of $256 \times 3 \times 3 \times 128$, $256 \times 3 \times 3 \times 256$, and $256 \times 3 \times 3 \times 256$. In some embodiments, the three convolution layers in the fourth convolution block may have convolution kernels of $512 \times 3 \times 3 \times 256$, $512 \times 3 \times 3 \times 512$, and $512 \times 3 \times 3 \times 512$. In some embodiments, each of the three convolution layers in the fifth convolution block may have 512 convolution kernels of size $3 \times 3 \times 512$. The pooling layer in four convolution blocks may be 2×2 . The fusion image may have a size of $224 \times 224 \times 3$. The fusion feature image outputted by the first convolution block may have a size of $112 \times 112 \times 64$ and is inputted into the second convolution block, the fusion feature image outputted by the second convolution block may have a size of $56 \times 56 \times 128$ and is inputted into the third convolution block, the fusion feature image outputted by the third convolution block may have a size of $28 \times 28 \times 256$ and is inputted into the fourth convolution block, the fusion feature image outputted by the fourth convolution block may have a size of $14 \times 14 \times 512$ and is inputted into the fifth convolution block, and the fusion feature image outputted by the fifth convolution block may have a size of $14 \times 14 \times 512$. The group of images may include one or more objects with low resolution and less pixel information. In order to improve the detection of such objects, the last pooling layer of the feature extraction network may be removed, and thus

the resolution of high-level image features may be improved, and more image details may be retained to prevent the loss of small object features caused by oversampling.

[0130] The fusion feature image may be inputted into the Faster-CNN classification regression. The coordinate box regression and the classification may be outputted. The Faster-CNN classification regression may include an RPN layer, an ROI pooling layer, and a classification layer. The classification layer may include two full connection layers. Detailed description regarding the RPN layer, the ROI pooling layer, and the classification layer please refer to elsewhere in the present disclosure, for example, the description of 720 in FIG. 7. The coordinate box regression may include the bounding box of the object presented in an image. The classification may be the category of the object. The category of the object may be one category of a plurality of categories that have been recognized in the training process.

[0131] FIG. 10 is a flowchart illustrating an exemplary process for determining the fusion feature image according to some embodiments of the present disclosure. In some embodiments, a process 1000 may be implemented as a set of instructions (e.g., an application) stored in the storage device 150 or storage device 250. The processing device 400, the processor 220, and/or the processing device 112 may execute the set of instructions, and when executing the instructions, the processing device 400, the processor 220, and/or the processing device 112 may be configured to perform the process 1000. The operations of the illustrated process presented below are intended to be illustrative. In some embodiments, the process 1000 may be accomplished with one or more additional operations not described, and/or without one or more of the operations discussed. Additionally, the order of the operations of the process 1000 illustrated in FIG. 10 and described below is not intended to be limiting.

[0132] In 1010, the processing device 400 (e.g., the determination module 420) may obtain a plurality of feature images of the group of images by extracting image features from each image of the group of images. Detailed description regarding the group of images please refer to elsewhere in the present disclosure, for example, the description of 510 in FIG. 5.

[0133] In some embodiments, each image of the group of images may be inputted into a corresponding convolution network in a fusion feature extraction submodel, which may include the feature extraction part of a VGG network, a ResNet, or an Inception network.

[0134] In some embodiments, each image of the group of images may be inputted into a plurality of convolution blocks connected in series in a fusion feature extraction submodel. The plurality of convolution blocks may extract features from the image to output a feature image. Merely by way of example, the plurality of convolution blocks may be five convolution blocks. In some embodiments, each of the plurality of convolutional blocks may include at least one convolution layer and a pooling layer. In some embodiments, each of a portion of the plurality of convolution blocks may include at least one convolution

layer and a pooling layer, and the last convolution layer may merely include at least one convolution layer. An image processed by the at least one convolution layer may be inputted into the pooling layer to reduce the size of the image. In some embodiments, the convolution block may include two or three convolution layers. The convolution layer may use a plurality of convolution kernels to convolute an image inputted into the convolution layer. In some embodiments, the size of the convolution kernel may be $3*3*d$, wherein d indicates a depth of the inputted image. In some embodiments, the size of the convolution kernel may be $5*5*d$, wherein d indicates a depth of the inputted image. In some embodiments, the size of the convolution kernel may be $2*2*d$, wherein d indicates a depth of the inputted image. In some embodiments, the size of the pooling layer may be $2*2$. An image of size $m*m*n$ may be inputted into the pooling layer, and the pooling layer may output an image of size $m/2*m/2*n$ after processing the image of size $m*m*n$. The size of the convolution kernel and pooling layer may be various and are not limited to the embodiments above.

[0135] In 1020, the processing device 400 (e.g., the determination module 420) may determine the fusion feature image by fusing the plurality of feature images.

[0136] In some embodiments, the plurality of feature images may be inputted into a fusion unit. The fusion unit may fuse the plurality of feature images to obtain a fusion feature image. In some embodiments, the fusion unit may include a concatenation layer and a convolution layer. The concatenation layer may be used to concatenate a plurality of matrixes of the plurality of feature images to obtain a stacked matrix. The convolution layer may be used to reduce the depth of the stacked matrix to output the fusion feature image.

[0137] In some embodiments, the concatenation layer may include a matrix concatenation function. The concatenation layer may concatenate the plurality of matrixes of the plurality of feature images using the matrix concatenation function to obtain the stacked matrix. In some embodiments, the convolution layer may further include a convolution kernel of size $1*1$. The convolution layer may be used to reduce the dimension of the stacked matrix while the length and width of the stacked matrix are fixed. Merely by way of example, an image of size $m*m*n$ may be inputted into the convolution layer, after processing the convolution layer may output an image of size $m*m*n/3$.

[0138] Merely by way of example, the process 1000 may be executed by the fusion feature extraction submodel 1100 as shown in FIG. 11. The fusion feature extraction submodel 1100 may include three independent convolution networks 1-3 and a fusion unit.

[0139] The group of images may include three images of different modalities. Each image may be inputted into a corresponding convolution network and a corresponding feature image may be outputted. Each convolution network (e.g., the convolution network 1, the convolution network 2, the convolution network 3) may include five convolution blocks. The first convolution block may include two convolution layers and a pooling layer. The second convolution block may include two convolution layers and a pooling

layer. The third convolution block may include three convolution layers and a pooling layer. The fourth convolution block may include three convolution layers and a pooling layer. The fifth convolution block may include three convolution layers. In some embodiments, the two convolution layers in the first convolution block may have convolution kernels of $64 \times 3 \times 3$ and $64 \times 3 \times 3$. In some embodiments, the two convolution layers in the second convolution block may have convolution kernels of $128 \times 3 \times 3$ and $128 \times 3 \times 3$. In some embodiments, the three convolution layers in the third convolution block may have convolution kernels of $256 \times 3 \times 3$, $256 \times 3 \times 3$, and $256 \times 3 \times 3$. In some embodiments, the three convolution layers in the fourth convolution block may have convolution kernels of $512 \times 3 \times 3$, $512 \times 3 \times 3$, and $512 \times 3 \times 3$. In some embodiments, each of the three convolution layers in the fifth convolution block may have 512 convolution kernels of size 3×3 . The pooling layer in five convolution blocks may be 2×2 . Each image of the group of images may have a size of $224 \times 224 \times 3$. The feature image outputted by the first convolution block may have a size of $112 \times 112 \times 64$ and is inputted into the second convolution block, the feature image outputted by the second convolution block may have a size of $56 \times 56 \times 128$ and is inputted into the third convolution block, the feature image outputted by the third convolution block may have a size of $28 \times 28 \times 256$ and is inputted into the fourth convolution block, the feature image outputted by the fourth convolution block may have a size of $14 \times 14 \times 512$ and is inputted into the fifth convolution block, the feature image outputted by the fifth convolution block may have a size of $14 \times 14 \times 512$. The group of images may include one or more objects with low resolution and less pixel information. In order to improve the detection of such objects, the last pooling layer of the convolution network may be removed, and thus the resolution of high-level image features may be improved, and more image details may be retained to prevent the loss of small object features caused by oversampling.

[0140] Three feature images may be inputted into the fusion unit. The fusion unit may include a fusion module and a convolution layer. The fusion module may include a matrix concatenation function. The convolution layer may include a convolution kernel of size 1×1 . The fusion module may concatenate three matrixes of three feature images in the depth dimension to obtain a stacked matrix. The convolution layer may reduce the depth of the stacked matrix to output a fusion feature image. Merely by way of example, each feature image may have a size of $14 \times 14 \times 512$, the stacked matrix may have a size of $14 \times 14 \times 1536$, and the fusion feature image may have a size of $14 \times 14 \times 512$.

[0141] In some embodiments, the convolution networks 1-3 may be trained independently. In some embodiments, the convolution networks 1-3 may be jointly trained with other parts of the object recognition model.

[0142] In some embodiments, the fusion feature image may be inputted into a Faster-CNN classification regression. The coordinate box regression and the classification may be outputted. The Faster-CNN

classification regression may include an RPN layer, an ROI pooling layer, and a classification layer. The classification layer may include two full connection layers. Detailed description regarding the RPN layer, the ROI pooling layer, and the classification layer please refer to elsewhere in the present disclosure, for example, the description of 720 in FIG. 7. The coordinate box regression may include the bounding box of the object presented in an image. The classification may be the category of the object. The category of the object may be one category of a plurality of categories that have been recognized in the training process.

[0143] FIG. 12 is a flowchart illustrating an exemplary process for determining the fusion feature image according to some embodiments of the present disclosure. In some embodiments, a process 1200 may be implemented as a set of instructions (e.g., an application) stored in the storage device 150 or storage device 250. The processing device 400, the processor 220, and/or the processing device 112 may execute the set of instructions, and when executing the instructions, the processing device 400, the processor 220, and/or the processing device 112 may be configured to perform the process 1200. The operations of the illustrated process presented below are intended to be illustrative. In some embodiments, the process 1200 may be accomplished with one or more additional operations not described, and/or without one or more of the operations discussed. Additionally, the order of the operations of the process 1200 illustrated in FIG. 12 and described below is not intended to be limiting.

[0144] In 1210, the processing device 400 (e.g., the determination module 420) may obtain a plurality of feature images of the group of images by extracting image features from each image of the group of images. Detailed description regarding the group of images please refer to elsewhere in the present disclosure, for example, the description of 510 in FIG. 5.

[0145] In some embodiments, the processing device 400 (e.g., the determination module 420) may extract image features from each image of the group of images using a fusion feature extraction submodel. The fusion feature extraction submodel may include a plurality of first convolution networks.

[0146] In some embodiments, each image of the group of images may be inputted into a first convolution network corresponding to the image. Each of the plurality of first convolution networks may extract features from each image of the group of images to output a feature image. The first convolution network may include at least one convolution block.

[0147] In some embodiments, the first convolution network may include a convolution block. The convolution block may include at least one convolution layer, for example, one convolution layer, two convolution layers, three convolution layers, or the like. The convolution layer may use a plurality of convolution kernels to convolute an image inputted into the convolution layer. The size of the convolution kernel may be $n*n*d$, wherein d indicates a depth of the inputted image, and n may be 2, 3, 4, 5, 6, 7, etc. In

some embodiments, the convolution block may further include a pooling layer. In some embodiments, the size of the pooling layer may be 2×2 .

[0148] In some embodiments, the first convolution network may include two or more convolution blocks connected in series, for example, the number of the convolution blocks may be 2, 3, 4, 5, 6, etc. Merely by way of example, the first convolution network may include four convolution blocks. Each of the four convolutional blocks may include at least one convolution layer and a pooling layer. In some embodiments, the convolution block may include two or three convolution layers. The convolution layer may use a plurality of convolution kernels to convolute an image inputted into the convolution layer. In some embodiments, the size of the convolution kernel may be $3 \times 3 \times d$, wherein d indicates a depth of the inputted image. In some embodiments, the size of the convolution kernel may be $5 \times 5 \times d$, wherein d indicates a depth of the inputted image. In some embodiments, the size of the convolution kernel may be $2 \times 2 \times d$, wherein d indicates a depth of the inputted image. In some embodiments, the size of the pooling layer may be 2×2 . An image of size $m \times m \times n$ may be inputted into the pooling layer, and the pooling layer may output an image of size $m/2 \times m/2 \times n$ after processing the image of size $m \times m \times n$. The size of the convolution kernel and pooling layer may be various and are not limited to the embodiments above.

[0149] In 1220, the processing device 400 (e.g., the determination module 420) may obtain a preliminary fusion feature image by fusing the plurality of feature images.

[0150] In some embodiments, the plurality of feature images may be inputted into a fusion unit. The fusion unit may fuse the plurality of feature images to obtain a preliminary fusion feature image. Detailed description regarding the fusion unit please refer to elsewhere in the present disclosure, for example, the description of 710 in FIG. 7, 810 in FIG. 8, and 1020 in FIG. 10.

[0151] In 1230, the processing device 400 (e.g., the determination module 420) may determine the fusion feature image by extracting features from the preliminary fusion feature image.

[0152] In some embodiments, the fusion feature extraction submodel may include a second convolution network.

[0153] In some embodiments, the preliminary fusion feature image may be inputted into the second convolution network. The second convolution network may extract features from the preliminary fusion feature image to output a fusion feature image. The second convolution network may include at least one convolution block. The larger the number (or count) of the convolution blocks in the second convolution network is, the smaller the number (or count) of the convolution block in the first convolution network may be. For example, the first convolution network may include four convolution blocks and the second convolution network may include one convolution block. As another example, the first convolution network may include one convolution block and the second convolution network may include four convolution blocks. As still

another example, the first convolution network may include two convolution blocks and the second convolution network may include three convolution blocks. Detailed description regarding the convolution block please refer to elsewhere in the present disclosure, for example, the description of 710 in FIG. 7, 810 in FIG. 8, and 1020 in FIG. 10.

[0154] Merely by way of example, the process 1200 may be executed by the object recognition model 1300 as shown in FIG. 13. The object recognition model 1300 may include three independent convolution networks, a fusion unit, a CNN feature extraction network, and a Faster-CNN classification regression.

[0155] The group of images may include three images of different modalities. In some embodiments, the three images of different modalities may be a color image, an infrared image, and a polarization image. Each image may be inputted into a convolution network of the three independent convolution networks and a feature image corresponding to the image may be outputted. In some embodiments, the convolution network may include a convolution layer. In some embodiments, each image of the group of images may have a size of $224 \times 224 \times 3$. The convolution layer may have 64 convolution kernels of size $3 \times 3 \times 3$. The feature image outputted by the convolution network may have a size of $224 \times 224 \times 64$.

[0156] Three feature images may be generated by the three independent convolution networks based on the three images of different modalities and inputted into the fusion unit and a preliminary fusion feature image may be outputted. The fusion unit may include a fusion module and a convolution layer. The fusion module may include a matrix concatenation function. The convolution layer may include a convolution kernel of size 1×1 . The fusion module may concatenate three matrixes of three feature images in the depth dimension to obtain a stacked matrix. The convolution layer may reduce the depth of the stacked matrix to output a fusion feature image. Merely by way of example, the stacked matrix may have a size of $224 \times 224 \times 192$ and the preliminary fusion feature image may have a size of $224 \times 224 \times 3$.

[0157] The preliminary fusion feature image may be inputted into the CNN feature extraction network and the fusion feature image may be outputted. Detailed description regarding the CNN feature extraction network please refer to elsewhere in the present disclosure, for example, the description of FIG. 9. The fusion feature image outputted by the CNN feature extraction network may have a size of $14 \times 14 \times 512$.

[0158] In some embodiments, the convolution network may include at least one convolution layer and a pooling layer. In some embodiments, the convolution network may include two convolution layers and a pooling layer. In some embodiments, the two convolution layers may have convolution kernels of $64 \times 3 \times 3 \times 3$ and $64 \times 3 \times 3 \times 64$. The pooling layer may be a size of 2×2 . Each image of the group of images may have a size of $224 \times 224 \times 3$. The feature image outputted by the convolution network may have a size of $112 \times 112 \times 64$.

[0159] The three feature images may be inputted into the fusion unit and the fusion unit may output a preliminary fusion feature image of size $112 \times 112 \times 64$.

[0160] The preliminary fusion feature image may be inputted into the CNN feature extraction network and the fusion feature image may be outputted. The CNN feature extraction network may include four convolution blocks. The first convolution block may include two convolution layers and a pooling layer. The second convolution block may include three convolution layers and a pooling layer. The third convolution block may include three convolution layers and a pooling layer. The fourth convolution block may include three convolution layers. In some embodiments, the two convolution layers in the first convolution block may have convolution kernels of $128 \times 3 \times 3 \times 64$ and $128 \times 3 \times 3 \times 128$. In some embodiments, the three convolution layers in the second convolution block may have convolution kernels of $256 \times 3 \times 3 \times 128$, $256 \times 3 \times 3 \times 256$, and $256 \times 3 \times 3 \times 256$. In some embodiments, the three convolution layers in the third convolution block may have convolution kernels of $512 \times 3 \times 3 \times 256$, $512 \times 3 \times 3 \times 512$, and $512 \times 3 \times 3 \times 512$. In some embodiments, each of the three convolution layers in the fourth convolution block may have 512 convolution kernels of size $3 \times 3 \times 512$. The pooling layer in four convolution blocks may be 2×2 . The fusion feature image outputted by the first convolution block may have a size of $56 \times 56 \times 128$ and is inputted into the second convolution block, the fusion feature image outputted by the second convolution block may have a size of $28 \times 28 \times 256$ and is inputted into the third convolution block, the fusion feature image outputted by the third convolution block may have a size of $14 \times 14 \times 512$ and is inputted into the fourth convolution block, the fusion feature image outputted by the fourth convolution block may have a size of $14 \times 14 \times 512$.

[0161] The fusion feature image may be inputted into the Faster-CNN classification regression. The coordinate box regression and the classification may be outputted. The Faster-CNN classification regression may include an RPN layer, an ROI pooling layer, and a classification layer. The classification layer may include two full connection layers. Detailed description regarding the RPN layer, the ROI pooling layer, and the classification layer please refer to elsewhere in the present disclosure, for example, the description of 720 in FIG. 7. The coordinate box regression may include the bounding box of the object presented in an image. The classification may be the category of the object. The category of the object may be one category of a plurality of categories that have been recognized in the training process.

[0162] Merely by way of example, the process 1200 may be executed by the object recognition model 1400 as shown in FIG. 14. The object recognition model 1400 may include three independent first convolution networks, a fusion unit, a second convolution network, and a Faster-CNN classification regression.

[0163] The group of images may include three images of different modalities. In some embodiments, the three images of different modalities may be a color image, an infrared image, and a polarization image. Each image may be inputted into a first convolution network of the three independent convolution networks and a feature image corresponding to the image may be outputted. The first convolution network may include four convolution blocks. The first convolution block may include two convolution layers and a

pooling layer. The second convolution block may include two convolution layers and a pooling layer. The third convolution block may include three convolution layers and a pooling layer. The fourth convolution block may include three convolution layers. In some embodiments, the two convolution layers in the first convolution block may have convolution kernels of $64 \times 3 \times 3 \times 3$ and $64 \times 3 \times 3 \times 64$. In some embodiments, the two convolution layers in the second convolution block may have convolution kernels of $128 \times 3 \times 3 \times 64$ and $128 \times 3 \times 3 \times 128$. In some embodiments, the three convolution layers in the third convolution block may have convolution kernels of $256 \times 3 \times 3 \times 128$, $256 \times 3 \times 3 \times 256$, and $256 \times 3 \times 3 \times 256$. In some embodiments, the four convolution layers in the fourth convolution block may have convolution kernels of $512 \times 3 \times 3 \times 256$, $512 \times 3 \times 3 \times 512$, and $512 \times 3 \times 3 \times 512$. The pooling layer in three convolution blocks may be 2×2 . In some embodiments, each image of the group of images may have a size of $224 \times 224 \times 3$. The feature image outputted by the first convolution block may have a size of $112 \times 112 \times 64$ and is inputted into the second convolution block, the feature image outputted by the second convolution block may have a size of $56 \times 56 \times 128$ and is inputted into the third convolution block, the feature image outputted by the third convolution block may have a size of $28 \times 28 \times 256$ and is inputted into the fourth convolution block, the feature image outputted by the fourth convolution block may have a size of $28 \times 28 \times 512$.

[0164] The three feature images may be generated by the three independent first convolution networks based on the three images of different modalities and inputted into the fusion unit and a preliminary fusion feature image may be outputted. The fusion unit may include a fusion module and a convolution layer. The fusion module may include a matrix concatenation function. The convolution layer may include a convolution kernel of size 1×1 . The fusion module may concatenate three matrixes of three feature images in the depth dimension to obtain a stacked matrix. The convolution layer may reduce the depth of the stacked matrix to output a fusion feature image. Merely by way of example, the stacked matrix may have a size of $28 \times 28 \times 1536$ and the preliminary fusion feature image may have a size of $28 \times 28 \times 512$.

[0165] The preliminary fusion feature image may be inputted into the second convolution network and the fusion feature image may be outputted. The second convolution network may include a pooling layer and three convolution layers. Each of the convolution layers may have 512 convolution kernels of size $3 \times 3 \times 512$ and the pooling layer may be 2×2 . The fusion feature image outputted by the second convolution network may have a size of $14 \times 14 \times 512$.

[0166] The feature fusion of the feature images of different modalities in the later stage of VGG-16 can effectively reduce the adverse impact of the alignment error caused by the pixel-level registration of the images of different modalities.

[0167] The fusion feature image may be inputted into the Faster-CNN classification regression. The coordinate box regression and the classification may be outputted. The Faster-CNN classification regression

may include an RPN layer, an ROI pooling layer, and a classification layer. The classification layer may include two full connection layers. Detailed description regarding the RPN layer, the ROI pooling layer, and the classification layer please refer to elsewhere in the present disclosure, for example, the description of 720 in FIG. 7. The coordinate box regression may include the bounding box of the object presented in an image. The classification may be the category of the object. The category of the object may be one category of a plurality of categories that have been recognized in the training process.

[0168] In some embodiments, the convolution networks 1-3 in FIG. 11 and the three independent first convolution networks in FIG. 14 may be trained independently.

[0169] Compare the fusion feature extraction submodel in FIGs. 9 and 13 with that of FIGs. 11 and 14, feature fusion is performed at the early stage of feature extraction, which greatly reduces the computational complexity of the model.

[0170] FIG. 15 is a schematic flowchart illustrating an exemplary process for determining **the recognition result based on the plurality of candidate recognition results and the confidence scores** according to some embodiments of the present disclosure. In some embodiments, a process 1500 may be implemented as a set of instructions (e.g., an application) stored in the storage device 150 or storage device 250. The processing device 400, the processor 220, and/or the processing device 112 may execute the set of instructions, and when executing the instructions, the processing device 400, the processor 220, and/or the processing device 112 may be configured to perform the process 1500. The operations of the illustrated process presented below are intended to be illustrative. In some embodiments, the process 1500 may be accomplished with one or more additional operations not described, and/or without one or more of the operations discussed. Additionally, the order of the operations of the process 1500 illustrated in FIG. 15 and described below is not intended to be limiting.

[0171] In 1510, the processing device 400 (e.g., the determination module 420) may obtain a plurality of candidate recognition results and confidence scores of the plurality of candidate recognition results according to the plurality of object recognition submodels based on the group of images. Detailed description regarding the group of images please refer to elsewhere in the present disclosure, for example, the description of operation 510 in FIG. 5. Each of the plurality of object recognition submodels may include a feature extraction module and an object recognition module. The plurality of object recognition submodels may include a Faster R-CNN, a YOLO network, or the like.

[0172] In some embodiments, each of the plurality of object recognition submodels may correspond to each image of the group of images. Merely by way of example, the group of images may include three images of different modalities, and the plurality of object recognition submodels may include three object recognition

submodels. In some embodiments, the three object recognition submodels may be the same type models. For example, the three object recognition submodels may be all the Faster R-CNN.

[0173] In some embodiments, each of the three object recognition submodels may include the Faster R-CNN. Each of the three object recognition submodels may output a candidate recognition result including one or more bounding boxes of one or more objects, one or more categories of the one or more objects, and one or more probabilities corresponding to the one or more categories of the one or more objects. A probability corresponding to a category of an object may be the probability of the object being a specific category. The one or more probabilities corresponding to the one or more categories of the one or more objects may be the confidence scores of the one or more objects in the candidate recognition result.

[0174] In some embodiments, each of the three object recognition submodels may be the YOLO network. Each of the three object recognition submodels may output a candidate recognition result including one or more bounding boxes of one or more objects, one or more categories of the one or more objects, and one or more confidences corresponding to the one or more objects. In some embodiments, a confidence may be the probability of an object being a specific category. In some embodiments, the confidence may include the probability of an object being a specific category and the distance between a predicted bounding box and a real bounding box. The one or more confidences corresponding to the one or more objects may be the confidence scores of the one or more objects in the candidate recognition result.

[0175] In 1520, the processing device 400 (e.g., the determination module 420) may determine the recognition result based on the plurality of candidate recognition results and the confidence scores.

[0176] In some embodiments, the processing device 400 may determine an average confidence score of the confidence scores of a recognized object. In some embodiments, the processing device 400 may determine the recognition result based on the average confidence score and a threshold. When the average confidence score is larger than the threshold, the processing device 400 may determine that the object is the specific category and output the bounding box of the object. In some embodiments, the processing device 400 may determine a sum of the confidence scores of a recognized object and determine the recognition result based on the sum of the confidence scores. In some embodiments, the processing device 400 may determine the recognition result based on the sum of the confidence scores and a threshold. When the sum of the confidence scores is larger than the threshold, the processing device 400 may determine that the object is the specific category and output the bounding box of the object. In some embodiments, the plurality of object recognition submodels may be trained independently. In some embodiments, the plurality of object recognition submodels may be jointly trained.

[0177] In some embodiments, the processing device 400 (e.g., the determination module 420) may determine the weight of each of the confidence scores of the candidate recognition result. In some

embodiments, the processing device 400 may determine the weight based on environment information. Merely by way of example, if the weather is good (e.g., a sunny day, a cloudy day), the weight for the color image may exceed the weight for the polarization image, and the weight for the polarization image may exceed the weight for the infrared image; if the weather is bad (e.g., rainy, snowy, foggy days) or in the night, the weight for the infrared image may exceed the weight for the color image, and the weight for the color image may exceed the weight for the polarization image.

[0178] In some embodiments, the weight may be determined in the training process and may be determined based on the model itself. In some embodiments, the weight factor may be set up by the user. In some embodiments, the environment information may be determined based on the group of images themselves. In some embodiments, the environment information may be determined based on inputted information, for example, the time information, the weather information, etc.

[0179] It should be noted that the above description regarding the process 1500 is merely provided for the purposes of illustration, and not intended to limit the scope of the present disclosure. For persons having ordinary skills in the art, multiple variations or modifications may be made under the teachings of the present disclosure. However, those variations and modifications do not depart from the scope of the present disclosure.

[0180] Having thus described the basic concepts, it may be rather apparent to those skilled in the art after reading this detailed disclosure that the foregoing detailed disclosure is intended to be presented by way of example only and is not limiting. Various alterations, improvements, and modifications may occur and are intended to those skilled in the art, though not expressly stated herein. These alterations, improvements, and modifications are intended to be suggested by this disclosure and are within the spirit and scope of the exemplary embodiments of this disclosure.

[0181] Moreover, certain terminology has been used to describe embodiments of the present disclosure. For example, the terms “one embodiment,” “an embodiment,” and “some embodiments” mean that a particular feature, structure, or characteristic described in connection with the embodiment is included in at least one embodiment of the present disclosure. Therefore, it is emphasized and should be appreciated that two or more references to “an embodiment” or “one embodiment” or “an alternative embodiment” in various portions of this specification are not necessarily all referring to the same embodiment. Furthermore, the particular features, structures, or characteristics may be combined as suitable in one or more embodiments of the present disclosure.

[0182] Further, it will be appreciated by one skilled in the art, aspects of the present disclosure may be illustrated and described herein in any of a number of patentable classes or context including any new and useful process, machine, manufacture, or composition of matter, or any new and useful improvement thereof.

Accordingly, aspects of the present disclosure may be implemented entirely hardware, entirely software (including firmware, resident software, micro-code, etc.) or combining software and hardware implementation that may all generally be referred to herein as a “module,” “unit,” “component,” “device,” or “system.”

Furthermore, aspects of the present disclosure may take the form of a computer program product embodied in one or more computer-readable media having computer-readable program code embodied thereon.

[0183] A computer-readable signal medium may include a propagated data signal with computer readable program code embodied therein, for example, in baseband or as part of a carrier wave. Such a propagated signal may take any of a variety of forms, including electro-magnetic, optical, or the like, or any suitable combination thereof. A computer-readable signal medium may be any computer-readable medium that is not a computer-readable storage medium and that may communicate, propagate, or transport a program for use by or in connection with an instruction execution system, apparatus, or device. Program code embodied on a computer-readable signal medium may be transmitted using any appropriate medium, including wireless, wireline, optical fiber cable, RF, or the like, or any suitable combination of the foregoing.

[0184] Computer program code for carrying out operations for aspects of the present disclosure may be written in any combination of one or more programming languages, including an object oriented programming language such as Java, Scala, Smalltalk, Eiffel, JADE, Emerald, C++, C#, VB, NET, Python or the like, conventional procedural programming languages, such as the “C” programming language, Visual Basic, Fortran 2003, Perl, COBOL 2002, PHP, ABAP, dynamic programming languages such as Python, Ruby, and Groovy, or other programming languages. The program code may execute entirely on the user's computer, partly on the user's computer, as a stand-alone software package, partly on the user's computer and partly on a remote computer or entirely on the remote computer or server. In the latter scenario, the remote computer may be connected to the user's computer through any type of network, including a local area network (LAN) or a wide area network (WAN), or the connection may be made to an external computer (for example, through the Internet using an Internet Service Provider) or in a cloud computing environment or offered as a service such as a Software as a Service (SaaS).

[0185] Furthermore, the recited order of processing elements or sequences, or the use of numbers, letters, or other designations therefore, is not intended to limit the claimed processes and methods to any order except as may be specified in the claims. Although the above disclosure discusses through various examples what is currently considered to be a variety of useful embodiments of the disclosure, it is to be understood that such detail is solely for that purpose and that the appended claims are not limited to the disclosed embodiments, but, on the contrary, are intended to cover modifications and equivalent arrangements that are within the spirit and scope of the disclosed embodiments. For example, although the implementation of various components

described above may be embodied in a hardware device, it may also be implemented as a software only solution, e.g., an installation on an existing server or mobile device.

[0186] Similarly, it should be appreciated that in the foregoing description of embodiments of the present disclosure, various features are sometimes grouped together in a single embodiment, FIGure, or description thereof for the purpose of streamlining the disclosure aiding in the understanding of one or more of the various embodiments. This method of disclosure, however, is not to be interpreted as reflecting an intention that the claimed subject matter requires more features than are expressly recited in each claim. Rather, claim subject matter lies in less than all features of a single foregoing disclosed embodiment.

WHAT IS CLAIMED IS:

1. A method for object detection implemented on a computing device including at least one processor and a storage device, comprising:

acquiring a group of images of an object, the group of images including at least three images of different modalities; and

determining a recognition result of the object based on the group of images according to an object recognition model, the recognition result including at least one of a position of the object or a category of the object.

2. The method of claim 1, wherein the object recognition model includes a feature extraction submodel and an object recognition submodel; and

wherein the determining a recognition result of the object based on the group of images according to an object recognition model includes:

obtaining a fusion feature image based on the group of images according to the fusion feature extraction submodel, the fusion feature image including at least one fusion feature of the group of images; and

determining the recognition result based on the fusion feature image according to the object recognition submodel.

3. The method of claim 2, wherein the obtaining a fusion feature image includes:

obtaining a fusion image by fusing the group of images; and

determining the fusion feature image by extracting features from the fusion image.

4. The method of claim 2, wherein the obtaining a fusion feature image includes:

obtaining a plurality of feature images of the group of images by extracting image features from each image of the group of images; and

determining the fusion feature image by fusing the plurality of feature images.

5. The method of claim 2, wherein the obtaining a fusion feature image includes:

obtaining a plurality of feature images of the group of images by extracting image features from each image of the group of images;

obtaining a preliminary fusion feature image by fusing the plurality of feature images; and

determining the fusion feature image by extracting features from the preliminary fusion feature image.

6. The method of claim 2, wherein the fusion feature extraction submodel includes a plurality of convolution blocks connected in series, each of a portion of the plurality of convolution blocks including at least one convolution layer and a pooling layer, and the last convolution layer including at least one convolution layer.

7. The method of claim 2, wherein the fusion feature extraction submodel includes:

a first convolution block including at least one convolution layer and a pooling layer;
a second convolution block including at least one convolution layer and a pooling layer;
a third convolution block including at least one convolution layer and a pooling layer;
a fourth convolution block including at least one convolution layer and a pooling layer; and
a fifth convolution block including at least one convolution layer.

8. The method of claim 2, wherein the obtaining a fusion feature image includes:

obtaining a stacked matrix by concatenating matrixes corresponding to the group of images using a matrix concatenation function, the matrixes corresponding to the group of images including matrixes of a plurality of feature images of the group of images or matrixes of images in the group of images;
determining a fused matrix by reducing a dimension of the stacked matrix; and
determining the fusion feature image based on the fused matrix.

9. The method of claim 1, wherein the object recognition model includes a plurality of object recognition submodels, each of the plurality of object recognition submodels corresponding to an image of a modality; and
wherein the obtaining a recognition result of the object based on the group of images according to an object recognition model includes:

obtaining a plurality of candidate recognition results and confidence scores of the plurality of candidate recognition results according to the plurality of object recognition submodels based on the group of images, each of the plurality of candidate recognition results corresponding to one of the plurality of recognition submodels; and

determining the recognition result based on the plurality of candidate recognition results and the confidence scores.

10. The method claim 9, wherein the determining the recognition result based on the plurality of candidate recognition results and the confidence scores includes:

determining a weight of each of the plurality of candidate recognition results based on environment information; and

determining the recognition result based on the weight of each of the plurality of candidate recognition results, the plurality of candidate recognition results, and the confidence scores.

11. The method of claim 1, wherein the group of images includes a color image, an infrared image, and a polarization image.

12. A system for object detection, comprising:

at least one storage device including a set of instructions; and

at least one processor configured to communicate with the at least one storage device, wherein, when the set of instructions are executed, the at least one processor is configured to instruct the system to perform operations, including:

acquiring a group of images of an object, the group of images including at least three images of different modalities; and

determining a recognition result of the object based on the group of images according to an object recognition model, the recognition result including at least one of a position of the object or a category of the object.

13. The system of claim 12, wherein the object recognition model includes a feature extraction submodel and an object recognition submodel; and

wherein the determining a recognition result of the object based on the group of images according to an object recognition model includes:

obtaining a fusion feature image based on the group of images according to the fusion feature extraction submodel, the fusion feature image including at least one fusion feature of the group of images; and

determining the recognition result based on the fusion feature image according to the object recognition submodel.

14. The system of claim 13, wherein the obtaining a fusion feature image includes:

obtaining a fusion image by fusing the group of images; and

determining the fusion feature image by extracting features from the fusion image.

15. The system of claim 13, wherein the obtaining a fusion feature image includes:

obtaining a plurality of feature images of the group of images by extracting image features from each image of the group of images; and

determining the fusion feature image by fusing the plurality of feature images.

16. The system of claim 13, wherein the obtaining a fusion feature image includes:

obtaining a plurality of feature images of the group of images by extracting image features from each image of the group of images;

obtaining a preliminary fusion feature image by fusing the plurality of feature images; and

determining the fusion feature image by extracting features from the preliminary fusion feature image.

17. The system of claim 13, wherein the fusion feature extraction submodel includes a plurality of convolution blocks connected in series, each of a portion of the plurality of convolution blocks including at least one convolution layer and a pooling layer, and the last convolution layer including at least one convolution layer.

18. The system of claim 13, wherein the fusion feature extraction submodel includes:

- a first convolution block including at least one convolution layer and a pooling layer;
- a second convolution block including at least one convolution layer and a pooling layer;
- a third convolution block including at least one convolution layer and a pooling layer;
- a fourth convolution block including at least one convolution layer and a pooling layer; and
- a fifth convolution block including at least one convolution layer.

19. The system of claim 13, wherein the obtaining a fusion feature image includes:

- obtaining a stacked matrix by concatenating matrixes corresponding to the group of images using a matrix concatenation function, the matrixes corresponding to the group of images including matrixes of a plurality of feature images of the group of images or matrixes of images in the group of images;
- determining a fused matrix by reducing a dimension of the stacked matrix; and
- determining the fusion feature image based on the fused matrix.

20. A non-transitory computer-readable storage medium storing computer instructions, and when a computer reads the computer instructions in the storage medium, the computer executes a method for object detection, comprising:

- acquiring a group of images of an object, the group of images including at least three images of different modalities; and
- determining a recognition result of the object based on the group of images according to an object recognition model, the recognition result including at least one of a position of the object or a category of the object.

21. A system for object detection, comprising:

- an obtaining module configured to acquire a group of images of an object, the group of images including at least three images of different modalities; and
- a determination module configured to determine a recognition result of the object based on the group of images according to an object recognition model, the recognition result including at least one of a position of the object or a category of the object.