[0005] At least the following problems exist in the related art: although deep learning is applied widely in the image classification, image segmentation, object detection, key point positioning, posture estimation, human face recognition and etc., due to the situations such as complicated and changeable scenes and/or the difficulty in identifying objects, the visual task processing model generated based on the training of deep learning does not have high prediction accuracy when processing visual tasks.

SUMMARY

[0006] Embodiments of the present disclosure provide a method and apparatus for processing images, a device and a storage medium to improve a prediction accuracy of a visual task processing model.

[0007] According to an embodiment of the present disclosure, there is provided a method for processing images. The method includes:

[0008] acquiring an original image and auxiliary information of the original image, wherein the original image is a single image or a video frame in a video, wherein in the case that the original image is the single image, the auxiliary information of the original image includes a background image corresponding to the original image; in the case that the original image is the video frame in the video, the auxiliary information of the original image includes a previous video frame and a response map of the previous video frame;

[0009] acquiring an object feature map by inputting the original image into a main path of a first visual task processing model, and acquiring an auxiliary feature map by inputting the auxiliary information into a branch path of the first visual task processing model; and

[0010] acquiring a response map of the original image by fusing the object feature map and the auxiliary feature map and inputting fused object feature map and auxiliary feature map into the main path of the first visual task processing model; wherein

[0011] in the case that the original image is the video frame in the video, the response map of the previous video frame is acquired in following manners:

[0012] when the previous video frame belongs to one of first N video frames of a video, the response map of the previous video frame is a response map acquired by inputting the previous video frame into a second visual task processing model; and

[0013] when the previous video frame does not belong to one of the first N video frames of the video, the response map of the previous video frame is a response map acquired by inputting the previous video frame into the first visual task processing model,

[0014] wherein the second visual task processing model has higher prediction accuracy than

the first visual task processing model, and N is a positive integer.

[0015] According to an embodiment of the present disclosure, there is provided an apparatus for processing images. The apparatus includes:

[0016] an original image and auxiliary information acquiring module configured to acquire an original image and auxiliary information of the original image, wherein the original image is a single image or a video frame in a video, wherein in the case that the original image is the single image, the auxiliary information of the original image includes a background image corresponding to the original image; in the case that the original image is the video frame in the video, the auxiliary information of the original image includes a previous video frame and a response map of the previous video frame;

[0017] a feature map acquiring module configured to acquire an object feature map by inputting the original image into a main path of a first visual task processing model, and acquire an auxiliary feature map by inputting the auxiliary information into a branch path of the first visual task processing model; and

[0018] an original image response map acquiring module configured to acquire a response map of the original image by fusing the object feature map and the auxiliary feature map and inputting fused object feature map and auxiliary feature map into the main path of the first visual task processing model; wherein

[0019] in the case that the original image is the video frame in the video, the response map of the previous video frame is acquired in following manners:

[0020] when the previous video frame belongs to one of first N video frames of a video, the response map of the previous video frame is a response map acquired by inputting the previous video frame into a second visual task processing model; and

[0021] when the previous video frame does not belong to one of the first N video frames of the video, the response map of the previous video frame is a response map acquired by inputting the previous video frame into the first visual task processing model,

[0022] wherein the second visual task processing model has higher prediction accuracy than the first visual task processing model, and N is a positive integer.

[0023] According to an embodiment of the present disclosure, there is provided a computer device for processing images. The computer device for processing images includes:

[0024] one or more processors; and

[0025] a memory configured to store one or more programs,

[0026] wherein the one or more programs is executed by the one or more processors to cause the one or more processors to realize the method according to the embodiments of the

the original image with the participation of the background image, the background image may provide relatively strong prior knowledge which helps to solve the problems which affect the prediction accuracy of the visual task processing model, such as complicated and changeable scenes and/or the difficulty in identifying objects, and thus the prediction accuracy of the visual task processing model is improved.

[0094] FIG. 4 is a flowchart of yet another method for processing images according to an embodiment of the present disclosure. The present embodiment may be applied to the situation of visual task processing. The method may be executed by an apparatus for processing images. The apparatus may be realized by software and/or hardware. The apparatus may be configured in a device, such as a computer or a mobile terminal. As shown in FIG. 4, the method includes step 310 to step 330.

[0095] In step 310, a current video frame, a previous video frame and a response map of the previous video frame are acquired.

[0096] In step 320, an object feature map is acquired by inputting the current video frame into a main path of a first visual task processing model, and an auxiliary feature map is acquired by inputting the previous video frame and the response map of the previous video frame into a branch path of the first visual task processing model.

[0097] In step 330, a response map of the original image is acquired by fusing the object feature map and the auxiliary feature map and inputting fused object feature map and auxiliary feature map into the main path of the first visual task processing model.

[0098] In some embodiments of the present disclosure, the response map of the previous video frame may be acquired by the following two manners.

[0099] A first Manner. When the previous video frame belongs to one of first N video frames of a video, the response map of the previous video frame is a response map acquired by inputting the previous video frame into a second visual task processing model. When the previous video frame does not belong to one of the first N video frames of the video, the response map of the previous video frame is a response map acquired by inputting the previous video frame into the first visual task processing model. The second visual task processing model has higher prediction accuracy than the first visual task processing model, and N is a positive integer.

[00100] A second manner. When the previous video frame belongs to one of first T video frames of a video frame sequence corresponding to the previous video frame, the response map of the previous video frame is a response map acquired by inputting the previous video frame into the second visual task processing model. When the previous video frame does not belong to one of the first T video frames of the video frame sequence corresponding to the previous video