CLAIMS

What is claimed is:

1. A method for processing images, comprising:

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 comprises 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 comprises a previous video frame and a response map of the previous video frame;

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

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

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:

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

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,

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

2. The method according to claim 1, wherein the response map of the previous video frame is acquired by following manners:

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 a second visual task processing model; and

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 frame, 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,

wherein the video frame sequence is one of a plurality of video frame sequences acquired by dividing a plurality of video frames of a video, the second visual task processing model has higher prediction accuracy than the first visual task processing model, and T is a positive integer.

3. The method according to claim 1, wherein the first visual task processing model is trained in following manners:

acquiring an original training image, annotation information of the original training image and auxiliary training information of the original training image;

acquiring an object training feature map by inputting the original training image into a main path of a convolutional neural network, and acquiring an auxiliary training feature map by inputting the auxiliary training information into a branch path of the convolutional neural network;

acquiring a response map of the original training image by fusing the object training feature map and the auxiliary training feature map and inputting fused object training feature map and auxiliary training feature map into the main path of the convolutional neural network;

acquiring a loss function of the convolutional neural network according to the annotation information of the original training image and the response map of the original training image; and

adjusting a network parameter of the convolutional neural network according to the loss function until an output value of the loss function is less than or equal to a preset threshold, and taking the convolutional neural network as the first visual task processing model.

4. The method according to claim 3, wherein the auxiliary training information is auxiliary training information acquired after being subjected to a data augmenting processing.

5. The method according to claim 4, wherein the data augmenting processing comprises at least one of translation, rotation, clipping, non-rigid transformation, noise interference and color transformation.

6. An apparatus for processing images, comprising:

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 comprises 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 comprises a previous video frame and a response map of the previous video frame;

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

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

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:

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

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,

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

7. A computer device for processing images, comprising:

one or more processors; and

a memory configured to store one or more programs,

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 any of claims 1 to 5.

8. A non-volatile computer-readable storage medium storing a computer program thereon, wherein when the computer program is executed by a processor, the method according to any of claims 1 to 5 is realized.