Claims

1. An object detection network training method comprising:
receiving a plurality of input images, each showing an object and comprising a ground truth bounding box, and ground truth mask, for the object;
for each image in the plurality of images:
a) extracting features from said image using a neural network;
b) applying one or more predicted bounding boxes to the image, based on the image features, using a bounding box generation network;
c) calculating a combined loss comprising both:
a bounding box regression loss between the one or more predicted bounding boxes and the ground truth bounding box for the object in the image;
an instance coverage loss of the one or more predicted bounding boxes relative to ground truth mask; and
d) updating the bounding box generation network using the combined loss; and
e) repeating steps a) to d) until convergence, thereby producing a trained bounding box generation network; and
the object detection network comprising the trained bounding box regression network and being configured to determine a final bounding box based on the predicted boxes.

2. The method of claim 1, wherein the bounding box regression loss is based on, for each image:
a distance between each predicted bounding box and the ground truth bounding box for the respective image; and
an area of each predicted bounding box relative to an area of the ground truth bounding box.

3. The method of claim 1 or 2, wherein the instance coverage loss is based on, for each object, a distribution of all said predicted bounding boxes around the object.

4. The method of claim 3, wherein the instance coverage loss is determined, at least in part, by calculating a box mask intersection over union (BMIoU) between each predicted bounding box and the ground truth mask.

5. The method of claim 4, wherein the instance coverage loss is based on, for each object, a percentage of each respective predicted bounding box that comprises an instance of the object.

6. The method of claim 5, wherein the instance coverage loss is based on, for each object, uniformity of the distribution of all said predicted bounding boxes around the object.

7. The method of claim 6, wherein uniformity of the distribution of all said proposed bounding boxes (UBD) is determined by calculating a number of pixels that are included in the boxes for each instance and then computing the distribution loss as:

$$L_{UBD} = 1 - \frac{\sum_{i=1}^{N_{box}} \sum_{j \in (A_B \cap A_M) \; and \; j! \in B_{Ins}} B_{ins} + 1}{\sum_{i \in A_M} M_{ins}}$$

where N_{box} is the number of proposed bounding boxes on an object, A_B is the area of box B, A_M is area of mask M, M_{ins} is the number of pixels of object, and B_{ins} is number of object pixels in all predicted boxes.

8. The method of any one of claims 1 to 7, wherein the object detection model is configured to determine the final bounding box by applying non-maximum suppression on all predicted boxes.

9. A method for assigning a bounding box to a further object in a further image, comprising:
receiving the further image; and
applying an object detection network trained in accordance with the method of any one of claims 1 to 8, to the further image, thereby producing the bounding box for the further object.

10. A system for training an object detection network, the system comprising:

memory;

at least one processor;

an object detection module comprising:

an image feature extractor network; and

the object detection network, the object detection network comprising:

a bounding box generation network; and

final bounding box predictor model,

the memory storing instructions that, when executed by the at least one processor, cause the system to:

obtain a plurality of input images, each showing an object and comprising a ground truth bounding box, and ground truth mask, for the object;

for each image in the plurality of images:

a) extract features from said image using the image feature extractor network;

b) apply one or more predicted bounding boxes to the image, based on the image features, using the bounding box generation network;

c) calculate a combined loss comprising both:

a bounding box regression loss between the one or more predicted bounding boxes and the ground truth bounding box for the object in the image;

an instance coverage loss of the one or more predicted bounding boxes relative to ground truth mask; and

d) update the bounding box generation network using the combined loss; and

e) repeat steps a) to d) until convergence, thereby training the bounding box generation network; and

employ the final bounding box predictor model to determine a final bounding box based on the predicted boxes.

11. The system of claim 10, wherein the bounding box regression loss is based on, for each image:

a distance bet ween each predicted bounding box and the ground truth bounding box for the respective image; and

an area of each predicted bounding box relative to an area of the ground truth bounding box.

12. The system of claim 10 or 11, wherein the instance coverage loss is based on, for each object, a distribution of all said predicted bounding boxes around the object.

13. The system of claim 12, wherein the instance coverage loss is determined, at least in part, by calculating a box mask intersection over union (BMIoU) between each predicted bounding box and the ground truth mask.

14. The system of claim 13, wherein the instance coverage loss is based on, for each object, a percentage of each respective predicted bounding box that comprises an instance of the object.

15. The system of claim 14, wherein the instance coverage loss is based on, for each object, uniformity of the distribution of all said proposed bounding boxes around the object.

16. The system of claim 15, wherein uniformity of the distribution of all said proposed bounding boxes (UBD) is determined by calculating a number of instances that are included in the boxes for each instance and then computing the distribution loss as:

$$L_{UBD} = 1 - \frac{\sum_{i=1}^{N_{box}} \sum_{j \in (A_B \cap A_M) \; and \; j! \in B_{Ins}} B_{ins} + 1}{\sum_{i \in A_M} M_{ins}}$$

where N_{box} is the number of proposed bounding boxes on an object, A_B is the area of box B, A_M is area of mask M, M_{ins} is the number of pixels of object, and B_{ins} is number of object pixels in all predicted boxes.

17. The system of any one of claims 10 to 16, wherein the object detection model is configured to determine the final bounding box by applying non-maximum suppression on all predicted boxes.

18. The system of any one of claims 10 to 17, being used for assigning a bounding box to a further object in a further image, the system comprising memory and at least one processor, the memory storing instructions that, when executed by the at least one processor, cause the system to determine that the bounding box generation module is trained, receive a new image and apply the object detection model to the new image to generate a final bounding box for the new image.