November 2022

# APPLICATION OF ONE-SHOT LEARNING DETECTOR TO AVOID UNWANTED OBJECTS TO BE BLURRED/HIDDEN IN VIDEO CONFERENCING

HP INC

# *Application of One-shot learning detector to avoid unwanted objects to be blurred/hidden in video conferencing*


## Abstract

In the era of hybrid working mode, more and more people prefer online video meetings to face-to-face meetings. Considering that people may have online meetings from home, there are many solutions such as background blur or virtual background to protect user privacy.


However, if a user tries to demonstrate certain object when the background blur/virtual background enabled, the object may be viewed as a part of background and then loss the visibility as the picture below.

This invention disclosure proposes the idea of utilizing one-shot learning to allow users to specify the objects they want to be visible in a background blur/virtual background enabled video conferencing, along with implementing a post-processing to make any of one-shot learning models perform better for object tracking in video conferencing user scenario.

In this disclosure, the user scenario we use to explain the idea is set to show an object in a background blur/virtual background enabled video conferencing. But the application of this idea is not limited to this scenario, for example, it can be also used for auto framing on any of objects users specify instead of only on human faces. Besides, even though this scenario combines image segmentation, gesture detection, and one-shot learning deep learning models, we will only discuss more details on one-shot learning model in this disclosure because it's the most critical part for the invention, and the rest of the two models are very common in existing applications.

## Prior Solutions

Prior solution for avoiding unexpected blurring on objects users want to show:

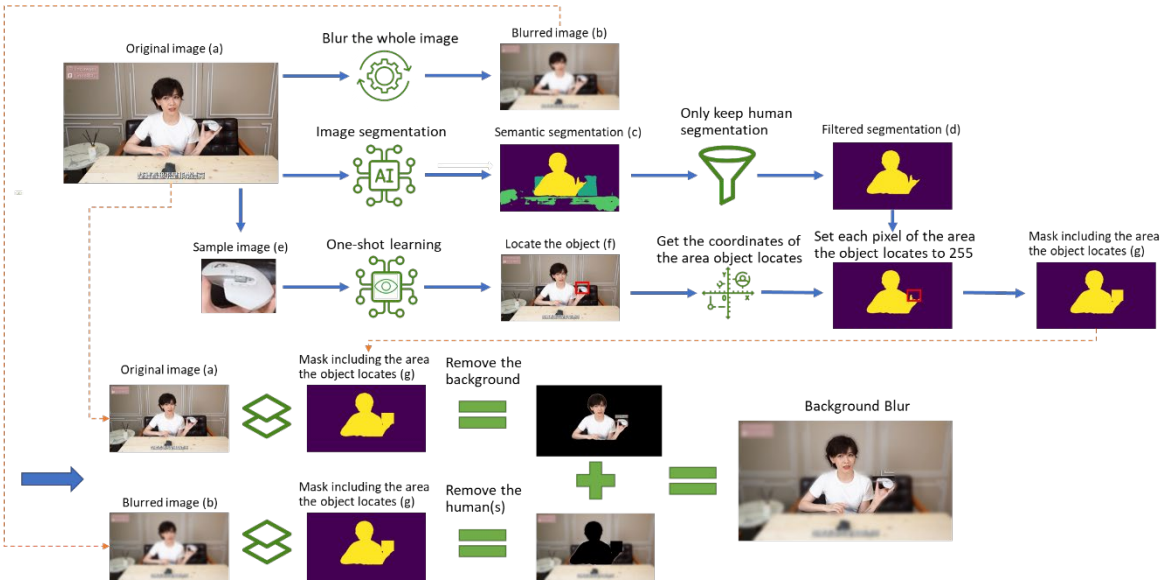
- Simply put the objects within the torso area so the background blur/virtual background will not mis-classify the object as part of background (see the video on the right). But this way will only give users very restricted space to demonstrate objects.
- Monocular depth estimation: this method is to use a model that can do depth estimation. Thus, a user can configure how deep the blurring is applied. If an object is placed within the depth, it is always visible. However, everything within the depth will also be seen.

Prior solution for detecting specified objects with one or a few samples:

- Comparing the sample image and frames pixel by pixel: If the similarity is over certain threshold, it is viewed as matched. The disadvantage of this approach is obvious. If the background of the object is different from the sample one's, it's very possible to be inferred as unmatched.
- Comparing the sample image and frames feature by feature: Since it compares the extracted features from both images, it has better tolerance to different background and noise signal.

However, if the object is slightly rotated or if the size of the object is different due to moving it towards/away from camera, it's also likely to be viewed as unmatched.

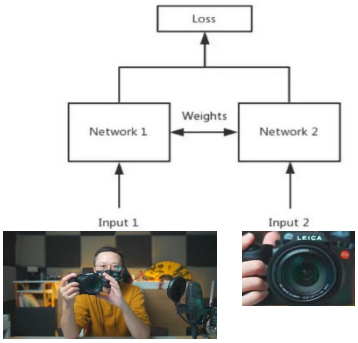
## Invention Description




Here explains how the proposed solution extends the typical background blur (see the flow chart in the previous slide) to be able to allow users to choose which object to not get blurred. It uses one sample image (e) of the target object to have the one-shot learning model learn detecting the target object. Then it tries to locate the object in the following frames from the video stream. Once it locates the object, it sends the coordinates of the area object locates to a mask maker that determines which portion should be unblurred. The mask maker sets each pixel of the area the object locates to 255 so that in the rest of the steps will not blur the area.

A one-shot learning model usually uses Siamese Network as its feature extractor architecture (see the diagram below), which is commonly used for comparing the similarity between two images. In our case, the input 1 and input 2 can be a frame from video stream and a sample image. The network 1 and network 2 are CNN layers, and the weights are shared between the two networks. So, the parameters in both networks will be adjusted together during model training. Therefore, they will have consistent way to extract features.

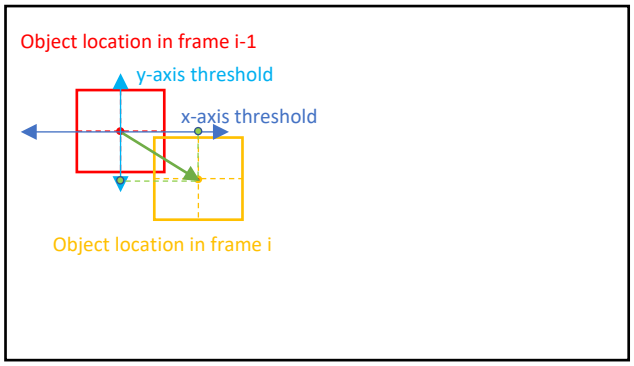
Siamese Network



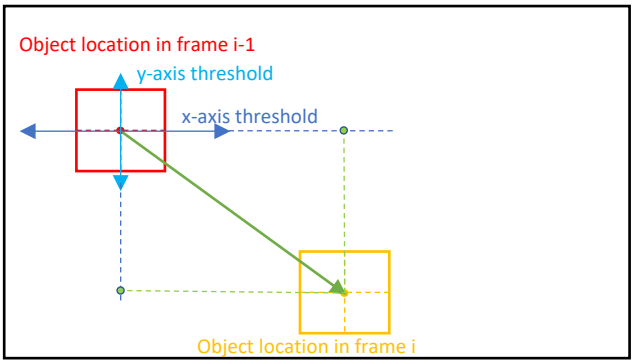

Considering the object should not have sudden long movement from one position to another in the scenario of demonstrating an object in video conferencing, we added a filter mechanism to filter out the location predictions that have unreasonable distance of movement. The algorithm of the filter mechanism is as described below:

1. The model predicts a list of unblurred area candidates, which have reduced by NMS (non-maximum suppression) and sorted by confidence scores from high to low.

2. The filter calculates translation distances on x and y axis separately between the centroids of the first unblurred area candidate in the list of the current frame (frame i) and the unblurred area in the previous frame (frame i-1) (see the diagram below).
3. If the distances on x and y axis are both less than the threshold of x and y axis respectively (see the left diagram below), the filter will output the candidate and end the filtering process.
4. If any of the distances on x and y axis is over the threshold of the corresponding axis (see the right diagram below). The filter will check the next candidate to calculate the translation distances as step 2 until finding the one that has shorter distances than the thresholds.
5. If none of the candidates has both shorter distances than the thresholds, the filter will conclude that the target object is not in the camera view and will stop tracking the object. This way will prevent the model to unblur the less likely area to impact the privacy protection.




The translation distances on x-axis and y-axis are less than threshold. The filter outputs the prediction and ends the




The translation distances on x-axis and y-axis are both over threshold. The filter discards the current prediction candidate and moves to check the next predicted unblurred area

## Advantages

- The disclosure invention effectively solves the pain points of showing objects in background blur/virtual background enabled video conferencing.
- It has better privacy protection than depth-based background blur
- It provides more convenient way than the typical background blur/virtual background, in which users can only show objects within the range of user torso.
- Only single or a few images are needed to immediately make the model able to detect arbitrary objects without re-training it.
- It has better tolerance to background difference between the sample image and frames from video stream because it finds the matches with extracted features rather than the way of comparing pixel by pixel.
- Utilizing Spatial Transformation Network to make the model still able to detect the target object even if it's rotated, moving towards/away from camera, or changing the locations.
- Considering of the object moving distance between consecutive frames makes the object tracking more accurate in video conferencing user scenario.

*Disclosed by Hong-Wei Chou / Albert Ma / Jerry Lu / Cheri Wang, HP Inc.*