

1. *The article does not detail the improvements to the algorithm, which is expected in a scientific article. Otherwise, it resembles press material of a commercial offer. In conclusion, the area of the article should be expanded with a detailed description of the improvements to the saliency prediction network TA-MSNet method.*

We thank you for your suggestion. The following descriptions have been added in 3.2 (Software Architecture):

Since the backbone of TA-MSNet, namely EML-Net, is very slow, we replace it with EfficientNet-b0 [17] to speed up the computation of the network. The reason for choosing EfficientNet is that its convolution operation is depth-wise separable convolution, which is designed to accelerate the CPU inference time of the model on the edge device. Furthermore, TA-MSNet uses a transformer for the saliency fusion of multiple images. The number of parameters in the transformer was considerably large, resulting in TA-MSNet having up to 108.07 million parameters. Therefore, we replaced the transformer with a graph convolutional network. The number of parameters then reduces to 57.48 million.

2. *The comparison of the developed method with the TA-MSNet method is, in my opinion, also insufficiently supported by results and information on test conditions, test assumptions, comparison factors, measurement technique, etc.*

We thank you for your suggestion. We have rewritten the Section 5 (Experiment) to explain test conditions, test assumptions, comparison factors, and measurement technique more clearly. Some of the contents are as follows:

The TA-MSNet dataset is derived from the 17 "camera moving" videos in the VR-EyeTracking dataset [26]. The reason for choosing "camera moving" videos is that the camera is moving while shooting the video as if a drone is exploring the scene. After these videos are selected, the unnecessary noise is further removed. Finally, five sub-images are randomly cropped from each frame in each video to build the TA-MSNet dataset. Thirteen videos of the TA-MSNet dataset are used for training and the remaining videos are used for testing. TA-MSNet and the proposed model were trained, tested, and evaluated using the same data set and metrics. Regarding the CPU inference time test, TA-MSNet and the proposed model were performed on the same machine (Intel(R) Core(TM) i9-10900X CPU @ 3.70GH). More details can be found in Section 5 (Experiment).

The metrics used to evaluate the performance were similarity (SIM), correlation co-efficient (CC), and Kullback–Leibler divergence (KL). Similarity (cosine

similarity) [27] can be used to measure the degree of similarity between two vectors. The method can be applied in many fields, for example, to calculate the degree of similarity between two images. The correlation coefficient [28] is a statistical indicator that can be used to assess the degree of correlation between two variables. The closer the value of the correlation co-efficient is to 1, the higher the correlation between them; the closer the value is to 0, the lower the correlation between them. Kullback-Leibler divergence [29] can be used to measure the degree of difference between two independent probability distributions in the same sample space.