Supplemental Material for: Measuring the Importance of Temporal Features in Video Saliency

Matthias Tangemann[1][0000−0001−9734−8692], Matthias Kümmerer[1][0000−0001−9644−4703], Thomas S.A. Wallis[1,2][0000−0001−7431−4852], and Matthias Bethge[1,2]

[1] University of Tübingen, Tübingen, Germany
[2] Amazon Research, Tübingen, Germany
{matthias.tangemann,matthias.kuemmerer,tom.wallis,matthias}@bethgelab.org

1 DHF1K Dataset

The gaze maps of the DHF1K dataset [6] contain artifacts in the gaze maps that make it impossible to properly evaluate the gold standard model and most likely affect model scores. Therefore, as stated in the main paper, we did not evaluate models on DHF1K. Here we provide more details on this issue.

For the DHF1K dataset, gaze positions have been collected from 17 subjects and provided as binary gaze maps for every frame. In Figure 1 we plot a histogram of the number of gaze positions per frame. The histogram clearly shows that substantially more positions than subjects are given per gaze map for all frames. On average, the number of gaze positions is ten times higher than the number of subjects. The example gaze maps in Figure 2 show that the subject's gaze positions are represented as irregular clusters of multiple pixels and large, grid-like structures in the map.

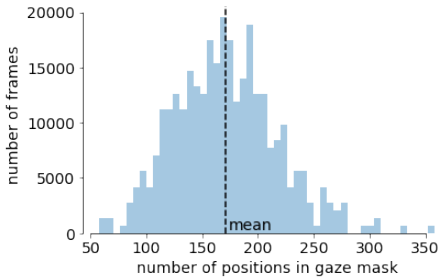

Fig. 1: Histogram of the number of positions in the binary gaze maps provided by the DHF1K dataset (17 subjects).

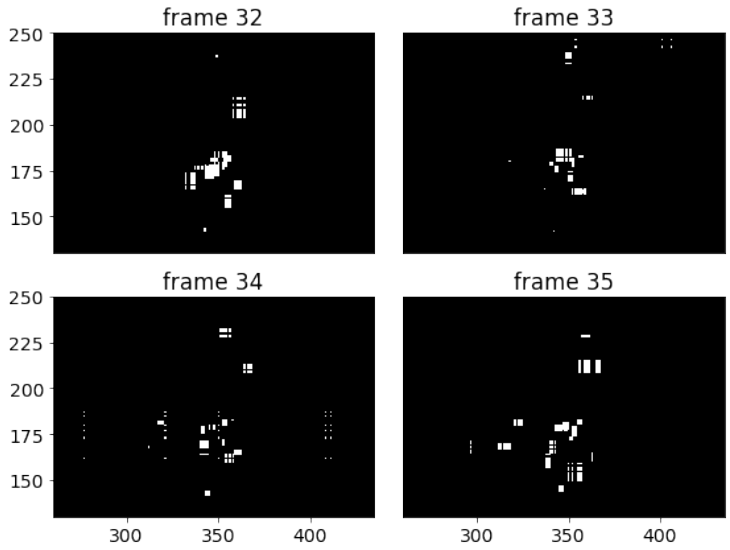

Fig. 2: Example gaze maps from sample "0601" of the DHF1K dataset. The maps represent gaze positions as irregular clusters of several pixels and contain large, grid-like structures.

The gold standard model is based on leave-one-out cross validation across subjects. The gaze maps provided with the DHF1K dataset however don't allow to determine the gaze locations for the individual subjects which makes it impossible to properly evaluate the gold standard model. Furthermore we expect those artifacts to affect the metrics used to evaluate gaze prediction models: depending on how many pixels are contained in any of the pixel clusters (which have very diverse sizes), that cluster will contribute more or less to the loss on this frame. For those reasons, we do not use the DHF1K dataset in our work.

Nevertheless, we evaluated the performance of our proposed model on the DHF1K validation set. As the results in table 1 show, DeepGaze MR performs better than many video saliency models. We did not adapt any hyperparameters for this dataset, so it is likely that the performance of our method could be improved further. The best performing model SalEMA [4] is based on an exponentially moving average, so similarly to our baseline model it cannot model temporal effects by design either. Summing up, those results suggest that temporal effects are of minor importance also for DHF1K.

2 Architecture Search

The architecture of DeepGaze MR described in the main paper has some important hyperparameters: We determined the number of input frames using a grid

	DHF1K		
Model	IG	AUC	NSS
Center bias	0	0.853	1.674
DeepVS [1]	-	0.854	1.067
DeepGaze II [2]	0.238	0.881	1.833
ACLNet [6]	-	0.893	2.412
DeepGaze MR	**0.702**	0.897	2.587
TASED-Net [5]	-	0.901	2.822
SalEMA [4]	-	**0.905**	**2.849**
Gold Standard	-	-	-

Table 1: Performance of state-of-the-art models on DHF1K. Due to the artifacts in the provided gaze maps the gold standard performance cannot be evaluated.

search and the depth and number of channels in the readout network using a random search. In the following, we present the respective results in more detail.

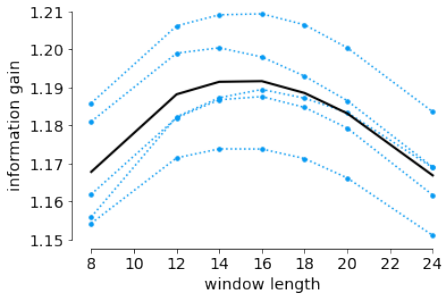

Fig. 3: Performance of a linear readout of VGG features averaged over time for different window lengths. Each blue line represents a single iteration, for which the same seed has been used for all window lengths. The black line represents the average performance per window length.

To find the optimal **window length** we used a linear instantiation of our model (i.e., only one convolutional layer in the readout network). We trained this model as the model described in the main paper, except that we used a learning rate of 0.001 which was decreased by a factor of 10 after 4 epochs. For the grid search we trained the model using 9 different window lengths from 8 to 24 frames and repeated the grid search for 5 different seeds. For all window lengths, the first 32 frames have been ignored for each video.

According to the results show in Figure 3, the optimal window length is 16 frames. However, the parameters seems to be not too sensitive as window lengths of 12–18 frames yielded a very similar performance.

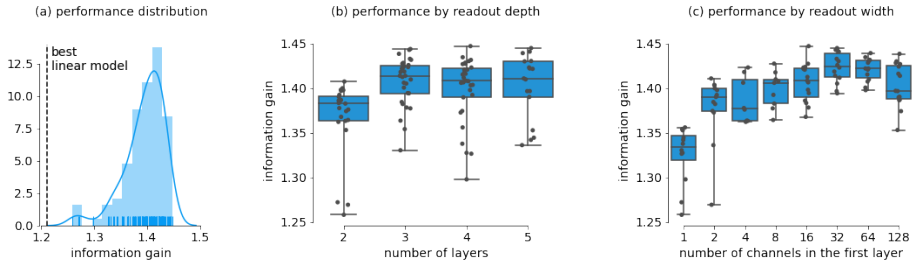

Fig. 4: Results of the random search for the optimal configuration of the readout network. (a) The distribution of performances achieved. The dashed line indicates the performance of the best linear model. (b) The performance by the number of layers in the readout network. (c) The performance by the number of channels in the first layer of the readout network.

Using the optimal window length of 16 frames, we then performed a random search to find the optimal **architecture of the readout network**. We trained 100 models having two to five layers and [1, 2, 4, ..., 128] channels in each layer.

The results are summarized in Figure 4. As the distribution of model performances in Figure 4a shows, most models clearly outperform the best linear model from the previous experiment. So a non-linear readout network is clearly needed for a good model. However, several different configurations of the readout network achieved a similar and good performance. So the detailed configuration of the readout network seems to be of minor importance.

Nevertheless, the random search revealed some trends. In Figure 4b, we plot the model performance of the networks by the number of convolutional layers used. The results indicate that the readout network should have a depth of at least 3 layers. A readout network consisting of only two layers is clearly too shallow and typically performs worse than deeper models.

Finally, we analyze the model performance depending on the number of channels in the first convolutional layer. Typically, this layer contains the majority of the readout network's parameters and therefore has a large impact on the model's capacity. As the results Figure 4c show, the number of channels in the first layer has to be high enough. Having only one layer doesn't allow the model to learn feature interactions and is clearly outperformed by the readout networks with higher capacity. Having 32 channels in the first layer seems to be optimal. Interestingly, we didn't see any signs of overfitting even for the biggest readout networks. So moderately sized readout networks appear sufficient to capture all available information for our architecture.

One of the simplest models reaching a top performance has 3 layers with 32, 32 and 1 channel, respectively. This is the architecture presented in the main paper.

Meta-Benchmark: LEDOV & DIEM							
Model	IG	%	AUC	CC	KLDiv	NSS	SIM
Center bias	0	0	0.869	0.258	2.607	1.918	0.190
DeepVS [1]	-	-	0.855	0.327	2.580	2.115	0.218
SalEMA [4]	-	-	0.890	0.470	2.454	2.613	0.389
ACLNet [6]	-	-	0.893	0.473	1.916	2.547	0.369
STRA-Net [3]	-	-	0.896	0.498	2.345	2.699	0.397
TASED-Net [5]	-	-	0.903	**0.567**	2.625	**3.078**	**0.448**
DeepGaze II [2]	0.326	14.8	0.903	0.445	1.545	2.041	0.354
DeepGaze MR	**0.799**	**36.2**	**0.913**	0.543	**1.325**	3.069	0.419
Gold Standard	2.207	100	0.952	-	-	5.490	-

Table 2: Performance of state-of-the-art models on a variant of our proposed meta-benchmark, using DeepGaze II as a baseline instead of DeepGaze MR. Comparing to the results of the original meta-benchmark, DeepGaze MR achieves much better results.

3 DeepGaze II as Baseline for the Meta-Benchmark

In our main work, we proposed a meta-benchmark consisting of those frames for which the information gain of the new DeepGaze MR model is more than 1bit worse than the gold standard model. By using the DeepGaze MR model as a baseline for defining the new benchmark, its results are disproportionally worse than that of other models.

As a comparison, we report the results of the benchmark using DeepGaze II as baseline model in Table 2. This way, our model performs similar to DeepGaze II on the original meta-benchmark whereas DeepGaze II now performs substantially worse. The results of the remaining models consistently improved compared to the original meta-benchmark. This indicates, that the benchmark variant defined using DeepGaze II is easier than the benchmark proposed in the paper.

References

1. Jiang, L., Xu, M., Liu, T., Qiao, M., Wang, Z.: DeepVS: A Deep Learning Based Video Saliency Prediction Approach. In: The European Conference on Computer Vision (ECCV). pp. 602–617 (Sep 2018)
2. Kümmerer, M., Wallis, T.S., Gatys, L.A., Bethge, M.: Understanding Low- and High-Level Contributions to Fixation Prediction. In: The IEEE International Conference on Computer Vision (ICCV). pp. 4799–4808 (Oct 2017)
3. Lai, Q., Wang, W., Sun, H., Shen, J.: Video Saliency Prediction using Spatiotemporal Residual Attentive Networks. IEEE Transactions on Image Processing **29**, 1113–1126 (2020). https://doi.org/10.1109/TIP.2019.2936112
4. Linardos, P., Mohedano, E., Nieto, J.J., O'Connor, N.E., Giro-i-Nieto, X., McGuinness, K.: Simple vs complex temporal recurrences for video saliency prediction. In: British Machine Vision Conference (BMVC) (Sep 2019)
5. Min, K., Corso, J.J.: TASED-Net: Temporally-Aggregating Spatial Encoder-Decoder Network for Video Saliency Detection. In: The IEEE International Conference on Computer Vision (ICCV). pp. 2394–2403 (Oct 2019)
6. Wang, W., Shen, J., Guo, F., Cheng, M.M., Borji, A.: Revisiting Video Saliency: A Large-scale Benchmark and a New Model. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4894–4903 (Jun 2018)