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

Abstract. Where people look when watching videos is believed to be heavily influenced by temporal patterns. In this work, we test this assumption by quantifying to which extent gaze on recent video saliency benchmarks can be predicted by a static baseline model. On the recent LEDOV dataset, we find that at least 75% of the explainable information as defined by a gold standard model can be explained using static features. Our baseline model "DeepGaze MR" even outperforms state-of-the-art video saliency models, despite deliberately ignoring all temporal patterns. Visual inspection of our static baseline's failure cases shows that clear temporal effects on human gaze placement exist, but are both rare in the dataset and not captured by any of the recent video saliency models. To focus the development of video saliency models on better capturing temporal effects we construct a meta-dataset consisting of those examples requiring temporal information.

Keywords: gaze prediction · saliency · video · temporal modelling · model evaluation

1 Introduction

The human visual system processes information from the environment selectively. Several attention mechanisms limit the amount of information to be processed and thus enable efficient perception of the world (e.g., [9]). The most obvious form of attention is the shifting of gaze, which orients the high-resolution fovea towards areas of interest.

Modelling those gaze shifts is an important topic in computer vision. Predictive models of human gaze have the potential to advance our understanding of human visual attention, for example by aiding the development of hypotheses that can be tested with human subjects [7]. Besides their scientific usefulness, such models have various technical applications. They can be used for graphic design [6], automated cropping, video compression [11] or other computer vision tasks (e.g., [48]).

Great progress has been made recently in predicting where people look in still images. With the use of pre-trained models the performance improved from

1/3 to more than 80% of explainable information explained (e.g., [27, 25]). Since the human visual system developed in a dynamic environment, there is growing interest to also model human gaze on videos. Previous studies revealed that motion patterns are an important factor attracting visual attention [39, 16, 8]. All recent video gaze models therefore are based on temporal modeling components such as recurrent units or spatiotemporal convolutions to capture those dynamic patterns.

To which degree those temporal patterns influence human gaze on natural videos and to which degree the recent performance improvements in video gaze prediction can be attributed to capturing these effects, however, has not been evaluated thoroughly so far. With our work we aim at filling this gap, by providing a method to measure the influence of temporal patterns on human gaze. We construct a static baseline model that by design cannot capture temporal effects and compare its performance to a gold standard model estimating the total information in the ground truth gaze data. The performance difference to the gold standard then represents an upper bound to the influence of temporal effects on the respective dataset. Furthermore, by looking at the largest failure cases of our static baseline, we can identify situations in the dataset where human gaze is driven by temporal patterns. Evaluating gaze prediction models on those situations then lets us draw conclusions about the capabilities of models to predict temporal effects.

Applying this method to the recent LEDOV dataset [20] and state-of-the-art video gaze models we arrive at the following conclusions: (1) Human gaze placement on the videos contained in the LEDOV dataset is largely driven by spatial appearance. (2) Clearly identifiable temporal effects on human visual attention exist, but occur rarely in the videos considered. (3) We need to construct suitable video data sets to enable learning based models to capture temporal effects. Indeed, we show that all other recent video gaze models with the capacity for temporal modelling fail in the same situations as our restricted model.

We explicitly note that the main contribution of our work are above findings and the proposed evaluation method that we need to come to those findings, but not the static baseline model that is required for our analysis. Interestingly though, our baseline model outperforms state-of-the-art video gaze prediction models on the LEDOV and DIEM [34] datasets—despite deliberately ignoring all temporal information.

To enable other researches to apply our proposed evaluation method more easily, we collect a meta-benchmark from existing datasets that contains the situations requiring temporal information revealed by our analysis. The performance of new models on this meta-benchmark indicates how much an improved predictive performance can be attributed to better handling of temporal effects. We will make this meta-benchmark as well as our pre-trained baseline model publicly available.

2 Related Work

Substantial progress has been made on the task of gaze prediction for free viewing of images. While the influential model by Itti and Koch [18], inspired by Treisman and Gelade's feature integration theory [45], was devised to explain effects observed in visual search originally, it also achieved first successes in predicting where people look. Since then, more than 50 models have been proposed predicting probable gaze locations based on image content (for a recent comparison see, e.g., [12]). As in other areas of computer vision, the advent of deep learning gave rise to models greatly improving state-of-the-art performance [27, 35, 15, 24, 46]. DeepGaze II [27], the current state of the art model on the MIT Saliency Benchmark [5], captures 81% of the explainable information gain on that dataset (explainable information gain is an information-theoretic analogue of explainable variance, see [25] for details).

In contrast, gaze prediction for videos only recently attracted more attention. Several datasets and models have been developed, but neither a widely accepted benchmark nor an estimate of the amount of explainable information in those datasets exist. This makes an evaluation of the state of the field difficult.

Recently, two video gaze datasets have been introduced that are large enough to train deep neural network based models: LEDOV [20] and DHF1K [47]. More recently, Wang et al. also provided gaze recordings for video segmentation datasets [48]. The gaze recordings provided by Mathe and Sminichescu [32] for the Hollywood and UCF-Sports dataset are large enough for deep learning based approaches, but most of the subjects have not been recorded in the free-viewing setting. Several small datasets exists that provide high-quality recordings (e.g., DIEM [34], for an overview see [20]).

Starting with an extension of the Itti and Koch model to videos [17, 16], several models predicting gaze specifically for videos have been proposed [51, 14, 41, 38, 40, 52, 10, 53, 13, 30]. The performance of video gaze models has been greatly improved with the advent of deep learning. Bazzani et al. [3] trained a recurrent neural network based on features extracted from a spatiotemporal DNN predicting gaze using a mixture of Gaussians. The models by Wang et al. [47] and Wu et al. [50] pair convolutional LSTM units with an attention mechanism. Bak et al. [2] proposed a two-stream network using optical flow in parallel to RGB features. This two-stream approach has also been combined with convolutional LSTM units by [19, 20] and with convolutional GRU and an attention mechanism by [28]. Linardos et al. [31] proposed a model based on an exponential moving average of frame-wise features. Very recently, [33] and [43] proposed spatio-temporal encoder-decoder networks for video gaze prediction.

3 Methods

The main objective of our work is to evaluate the influence of temporal patterns on human gaze. To that end, we propose a baseline model that cannot learn temporal patterns by design but predicts human gaze on videos solely relying

on static appearance. This baseline model is then compared to a gold standard model as an estimate of the total information in the ground truth gaze data. The performance difference between those models represents an upper bound of the influence of temporal patterns on human gaze placement.

3.1 Center Bias

The center bias is an important lower baseline. It is obtained by blurring and normalizing a histogram of all gaze positions in the training set. As humans tend to look at the center of images [44] and videos are usually recorded such that interesting objects are in the middle of videos, there is a clear bias in the gaze data towards the center of the videos. The center bias therefore represents a prior distribution of gaze position independent of visual content. Predicting this spatial prior for every frame yields a lower baseline, comparable to the chance level performance in classification tasks.

The center bias is much stronger in the beginning of each video due to the subjects fixating the center of the screen before each trial. As described later, we ignore this effect by not evaluating on the first 15 frames and confirmed experimentally that a stationary center bias models the remaining data well. Furthermore, we optimized the blur size using a grid search.

3.2 Gold Standard Model

The maximal performance that gaze prediction models can achieve is limited by the consistency of the subjects and varies from frame to frame. We use a gold standard model [49] to measure the inter-subject variability of the gaze positions. The model predicts where each subject looked given the ground truth information from all other subjects on the same frame. This is done by blurring the gaze positions of all but one subject and performing leave-one-out cross validation. Moreover, the prediction of the gold standard model is mixed with a uniform distribution to handle outliers. The gold standard therefore predicts subjects to look where other subjects look with a high probability, and to randomly look anywhere on the image with a small probability defined by the mixing coefficient. The optimal blur size of the gaussian filter and the mixing weight of the uniform distribution are determined using a grid search.

A high gold standard performance indicates very consistent gaze locations across all subjects and vice versa. Therefore, the gold standard model yields an estimate of the maximal performance that can be achieved for every frame. All reported gold standard performances refer to the leave-one-subject-out performance.

3.3 Static Baseline Model

Our proposed evaluation method requires a static baseline model that cannot handle temporal effects by design. Initial experiments revealed that DeepGaze II

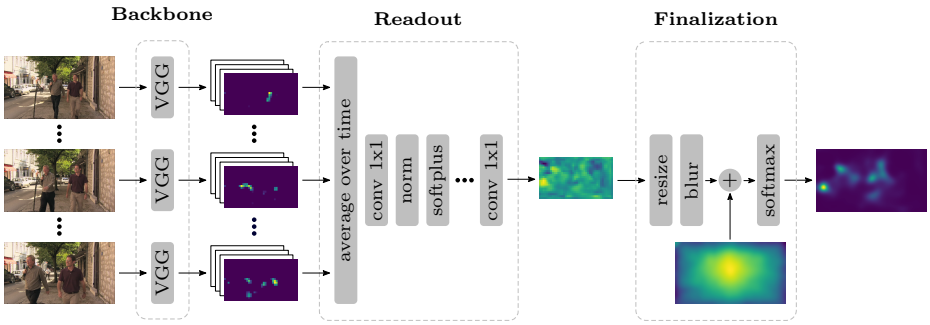

Fig. 1: Architecture of our static baseline model "DeepGaze MR": A feature representation is extracted from individual frames in a fixed size window using the VGG-19 network. A non-linear readout network transforms this representation into a priority map by first averaging the feature channels over time, and then applying a series of 1x1 convolutions. The resulting map is then resized, blurred, weighted by the center bias, and normalized to obtain the final prediction.

[27], the current state-of-the-art model for images, achieves a very competitive performance when simply applied to videos frame-by-frame (see section 4). However, this instantaneous model ignores delays due to the required processing in the human brain. This suggests a way to improve the DeepGaze II architecture for videos by averaging deep features over multiple recent video frames. Based on this approach, we propose a space-time separable variant of DeepGaze II using a temporal box filter as static baseline model (see Figure 1), which we call *DeepGaze Mean Readout (DeepGaze MR)*.

Input to our model is a fixed length window of consecutive frames, which is used to predict the gaze distribution on the last frame ("target frame") in this window. We use a window length of 16 frames, which was the optimal value found using a grid search (see supplement for details).

Backbone. Our model applies the VGG-19 network pretrained on Imagenet [42] to every frame individually and extracts the representation from the last convolutional layer (`conv5_4`) after the nonlinearity. We keep the parameters of the backbone fixed to prevent overfitting.

Readout. A non-linear readout network is used to transform the feature representation into a priority map of probable gaze locations. The readout network first averages the feature representation over time. A series of 1x1 pixel convolutions is then used to non-linearly combine the feature channels to the priority map. Layer Normalization [1] is used after all but the last convolutional layer to stabilize training. As non-linearity we use the softplus function, which is a smooth approximation of the commonly used ReLU and avoided units zeroing out early in training. We use three convolutional layers with 32, 32, and one channel, respectively. This optimal instantiation of the readout network has been found using a random search (see supplement for details).

Finalization. Finally, the output of the readout network is turned into the predicted probability distribution: First, the priority map is resized to the resolution of the input. It is then smoothed using a Gaussian with learnable standard deviation per x and y dimension. The logarithm of the center bias density from the training set is added to the map using a learnable weight, acting as a spatial prior. Finally, a softmax is applied to obtain the predicted spatial probability distribution of gaze locations.

Training. Our model is trained using maximum-likelihood learning (Kümmerer et al. [26] suggest that this allows for best metric scores in all classic saliency metrics). Thus, the loss function is the average log-density at gaze locations for each frame. We use the Adam optimizer [23] with a learning rate of 0.01, which is decreased by a factor of ten after one and five epochs. In each epoch, only one random target frame per video is used for training. Experiments confirmed, that this training scheme is sufficient for our model to converge.

Since our model averages features over time, it is by design not able to represent temporal patterns such as movements, or appearing and disappearing objects.

4 Experiments

In this section, we evaluate our baseline models described above on recent video gaze datasets. We then analyse the baseline predictions in comparison to state-of-the-art video gaze models to better understand the importance of temporal effects in video saliency.

The evaluation of gaze prediction models comes with challenges: different evaluation protocols and metrics typically lead to inconsistent model rankings. Building on recent work to better understand this evaluation process, we first describe and motivate the model evaluation approach used in this work.

4.1 Metrics

A large number of metrics exist that are used to evaluate gaze predictions (for a review see [4]). As typically used, these metrics give rise to inconsistent model rankings. Kümmerer et al. [26] proposed to adapt a probabilistic setting, i.e., to formulate models so that they predict spatial probability distributions, train them for log-likelihood and differentiate between predictions and derived saliency maps. In this way, consistent model ratings can be achieved.

We adopt this setting in our work, and use information gain (average log-likelihood per fixation compared to the center bias, [25]) as our primary metric. To enable a comparison to models that did not use a probabilistic approach, we additionally evaluate the AUC [22], NSS [36], CC [21], KLDiv [37, 29] and SIM [22] metrics to judge the performance of our model relative to state-of-the-art. To obtain an overall score for a model, the metrics are applied to the prediction for every frame individually, and then averaged first over frames and then over videos.

4.2 Datasets

The main dataset for this work is the LEDOV dataset [19]. It contains 538 short videos (11s on average) with eye tracking data of 32 subjects. The authors removed smooth pursuits and saccades and artificially stabilized fixations during their preprocessing, so this dataset does not allow to investigate the precise dynamics of individual gaze trajectories. However, the dataset covers the common factors driving human gaze placement sufficiently well to develop and compare models. All videos have been rescaled to 640x360px and resampled to 30Hz. Models from other groups are evaluated using the resolutions and frame rates that the respective models have been trained on.

For additional analyses we are using the DIEM dataset [34] (84 videos, 66 subjects on average, mean duration 95.2s). The videos have been padded to match the viewing conditions reported in the paper and rescaled to 640x480px.

The DHF1K dataset [47] is comparable in scope to LEDOV, but contains artifacts in the provided gaze maps. As those artifacts affect the model scores and make it impossible to properly evaluate the gold standard model, we excluded this dataset from our analysis. In the supplemental information we provide more details on this issue together with overall performance results which suggest that our conclusions are also valid for DHF1K.

For all datasets, the subject had to fixate the center of the screen before each trial. We do not evaluate models on the first 15 frames to ignore the centered gaze due to the experimental paradigm.

4.3 Performance Results

In Table 1, we show the performances of our baselines and other recent gaze models on LEDOV. Despite deliberately ignoring all temporal effects, DeepGaze MR performs very well and explains as much as 75% of the explainable information (as a comparison, the state-of-the-art for images on MIT1003 is 81%). Moreover, DeepGaze MR performs substantially better than DeepGaze II which confirms the effectiveness of our proposed adaptations. Interestingly, in AUC our model matches the gold standard performance, which might be due to the fact that AUC saturates very quickly. Furthermore, the AUC metric might suffer from the leave-one-subject-out cross validation applied in the gold standard.

We further compare the performances of our baselines to recent video gaze prediction models: The DeepVS model [20, 19] allows the most direct comparison, as it was trained on the LEDOV dataset as well. ACLNet [47], SalEMA [31], TASED-Net [33] and STRA-Net [28] are recent video gaze models developed on the DHF1K dataset [47]. For all models, we used the published weights and adapted size and frame rate of the input videos to match the samples encountered in the respective model training. As the results in Table 1 show, DeepGaze MR clearly outperforms all evaluated previous state-of-the-art models on the LEDOV dataset across all metrics, despite being designed as a static baseline model.

Additionally, we evaluated the models on the DIEM dataset. The size of the dataset is rather small (84 videos), therefore we did not train but only

LEDOV val

Model	IG	%	AUC	CC	KLDiv	NSS	SIM
Center bias	0	0	0.833	0.157	3.521	1.546	0.062
TASED-Net [33]	-	-	0.887	0.647	3.214	3.498	0.496
STRA-Net [28]	-	-	0.890	0.610	2.315	3.324	0.460
SalEMA [31]	-	-	0.890	0.596	2.573	3.331	0.466
ACLNet [47]	-	-	0.892	0.587	1.905	3.156	0.430
DeepVS [19]	-	-	0.894	0.397	2.445	3.098	0.210
DeepGaze II [27]	1.216	62.8	0.908	0.588	1.259	3.368	0.434
DeepGaze MR	**1.445**	**74.6**	**0.917**	**0.665**	**1.105**	**3.857**	**0.498**
Gold standard	1.961	100	0.917	-	-	4.992	-

LEDOV test

Model	IG	%	AUC	CC	KLDiv	NSS	SIM
Center bias	0	0	0.844	0.142	3.689	1.585	0.057
SalEMA [31]	-	-	0.897	0.590	2.377	3.152	0.465
TASED-Net [33]	-	-	0.897	0.650	2.965	3.361	0.505
ACLNet [47]	-	-	0.898	0.573	1.667	2.922	0.435
STRA-Net [28]	-	-	0.899	0.597	2.024	3.130	0.466
DeepVS [19]	-	-	0.903	0.394	2.398	3.081	0.218
DeepGaze II [27]	1.117	61.0	0.909	0.606	1.195	3.403	0.447
DeepGaze MR	**1.367**	**75.5**	**0.920**	**0.667**	**1.035**	**3.657**	**0.506**
Gold standard	1.810	100	0.920	-	-	4.676	-

Table 1: Performance comparison of recent gaze prediction models on the LEDOV dataset. The information gain can only be evaluated for models that predict a spatial probability distribution. All models have been applied using the published weights. TASED-NET, SalEMA, ACLNet and STRA-Net have been trained on the DHF1K dataset, DeepGaze II on SALICON and MIT1003.

evaluate the models on this dataset. As the results in Table 2 show, our model performs clearly better than all other video saliency methods on this dataset except TASED-Net [33]. Interestingly, the original DeepGaze II model for images performs even better than the variant adapted to videos.

The performances on DIEM are worse than those on LEDOV for two reasons. First, this dataset is much harder as the videos in this dataset contain much more temporal activity. Second, the domain gap to LEDOV is rather large, as DIEM contains cuts and many objects not present in LEDOV. The good performance of DeepGaze II on this dataset could therefore be explained by the much broader range of objects it has seen during training. Moreover, DeepGaze II is applied purely frame-by-frame, so it probably copes better with the many cuts in DIEM.

4.4 Analyzing Temporal Effects

In the following, we try to better understand the influence of temporal information on gaze placement. As motivated earlier, we use the information gain

	DIEM						
Model	IG	%	AUC	CC	KLDiv	NSS	SIM
Center bias	0	0	0.892	0.436	1.511	2.053	0.341
DeepVS [19]	-	-	0.853	0.424	2.070	2.096	0.309
SalEMA [31]	-	-	0.911	0.576	1.743	2.987	0.465
STRA-Net [28]	-	-	0.914	0.595	1.975	3.069	0.477
TASED-Net [33]	-	-	0.914	**0.621**	2.098	**3.194**	**0.493**
ACLNet [47]	-	-	0.914	0.558	1.468	2.826	0.428
DeepGaze MR	0.660	43.1	0.920	0.602	1.091	3.116	0.471
DeepGaze II [27]	**0.674**	**44.0**	**0.926**	0.619	**1.058**	2.898	0.477
Gold standard	1.531	100	0.940	-	-	4.659	-

Table 2: Performance comparison of recent gaze prediction models on the full DIEM dataset. Due to the small number of videos, none of the models has been trained on this dataset.

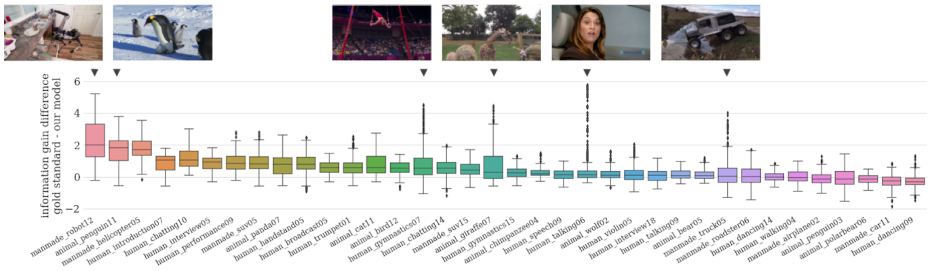

Fig. 2: Distribution of the unexplained information across frames in the LEDOV validation set (x-axis shows distinct videos). The remaining explainable information is estimated by the difference of our model to the gold standard in bit using the information gain metric. The videos marked are the largest failure cases of our model.

difference of the gold standard and DeepGaze MR as an estimator of the information that is not captured by our model. Since the model cannot learn temporal patterns by design, temporal effects on human gaze placement should result in large differences to the gold standard.

In Figure 2 we plot the distribution of those differences grouped by video. The median remaining information is close to 0bit for roughly half of the videos in the validation set. This indicates that our static baseline model successfully predicts gaze positions on a large number of frames. However, the results also clearly show two kinds of failure cases: (1) There are some videos for which the average performance gap to the gold standard is large. For the first three videos in the plot, the median difference is almost 2bit. (2) For other videos there is a large number of outlier frames whose performance gap is much greater than for most of the other remaining frames in the video. As our model is not able to exploit temporal structure by design, they should include cases in which temporal patterns affect human gaze placement.

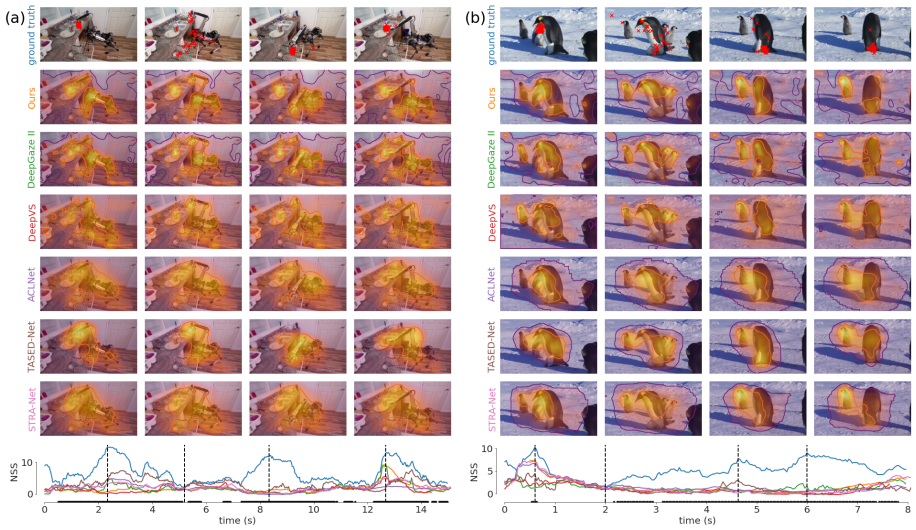

Fig. 3: Failure cases with a high average difference to the gold standard: **(a)** Most of the subjects look at the robot's hand while it puts a glass into a dishwasher. The models however distribute their prediction over the whole robot. **(b)** After roughly two seconds a small penguin becomes visible under the big penguin in the foreground, shifting the gaze of most subjects to the small penguin for the rest of the video. Markers on the x-axis of the NSS plots indicate frames that are part of our proposed meta-benchmark (see Section 4.5).

We analyze the found failure cases in more detail by visualizing them in Figures 3 and 4. We plot the NSS scores of the models over time (bottom) and visualize the model log predictions at interesting frames (top, frame position indicated by dashed lines in the NSS plot). As SalEMA averages features and thus cannot handle temporal effects by design, we don't consider it in this case study. The figures reveal three common factors that strongly influence where people look and are difficult for all models:

Interactions between objects occur in several of the videos. Here, most subjects look at the interaction point, not at the objects themselves. This is clearly observable in Figure 4b, when the child is feeding the giraffe or in Figure 3a when the robot is grabbing objects.

Suddenly appearing objects have a very strong ability to attract human attention as well. As can be seen in Figure 4a the shifting of the gaze to the appearing text is very consistent across all subjects. We assume that this effect can be observed with suddenly appearing objects in general, but cannot verify this hypothesis properly due to the small number of samples. A related effect is the appearing of the two persons due to the camera motion in Figure 4d. They also clearly attract attention, however much less than the sudden appearing of the text in Figure 4a.

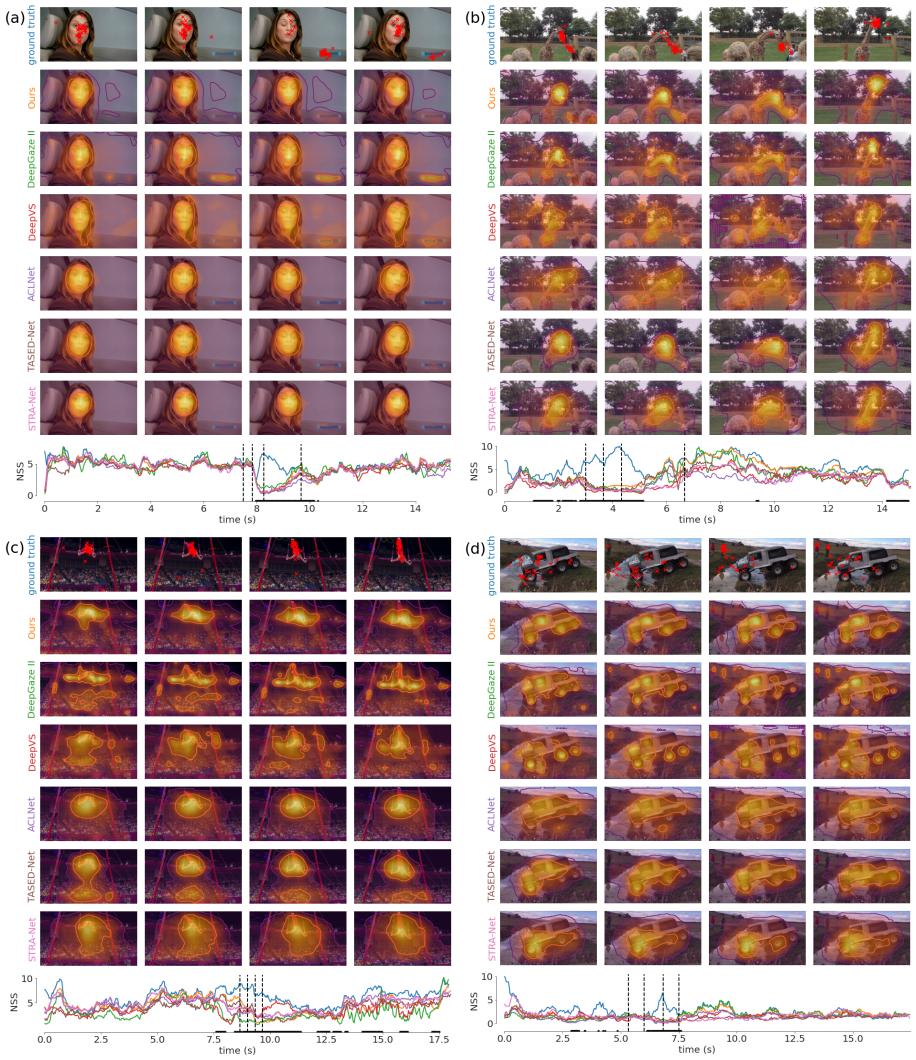

Fig. 4: Failure cases due to localized events: **(a)** A text suddenly appearing draws almost all attention for a short time, whereas the models predict people to mainly look at the person talking. **(b)** When a child is feeding a giraffe, the subjects' attention focuses at their interaction point and not at the giraffe's head that is looked at during the remainder of the video. **(c)** Gaze concentrates on the gymnast's torso during a swinging exercise, whereas other body parts are much less looked at. **(d)** Two persons enter the scene due to the movement of the camera, which temporarily attracts the attention of most of the subjects.

Movements of objects also clearly have the potential to change which parts of a scene are observed. In Figure 4c, none of the subjects looks at the gymnast's arms or hands, but all are looking at his torso that is moving in the respective scene. This stands in contrast to most cases in which humans appear, where subjects tend to look at people's hands or faces. Also global camera movements seem to be able to shift people's gaze towards the side of the direction of the movement, as indicated in Figure 4d. However the effect in this example is small and entangled with the appearing persons. A closer investigation would be necessary to address this effect.

The temporal effects described are compiled from the qualitative analysis of our model's largest failure cases. As their number is small, the list given is most likely not exhaustive. Moreover, it is not possible to reliably draw any general conclusions about the relative strengths of those effects. However, the cases presented clearly reveal the existence of such temporal effects and show that they are not captured by all recent video gaze models that should have the capacity to model them.

4.5 Evaluating Temporal Modelling

The detailed analysis of the failure cases in the previous section showed that none of the considered models was able to correctly predict cases in which temporal information influences where people look. As our proposed method requires training and evaluating two baseline models, the hurdle to apply it is quite high. To facilitate applying our method to new models, we propose a principled new meta-benchmark consisting of those hard cases.

Our meta-dataset contains all frames of videos where our static baseline's information gain is at least 1bit worse than the gold standard (indicated by markers on the x-axis of the NSS plots in Figures 3 and 4). We propose to run the models on the full videos, but only average the performances over the frames included in our meta-dataset. This evaluation scheme discards roughly 80% of the frames in LEDOV and 65% of the frames in DIEM. As our model cannot learn temporal effects by design, gaze on the discarded frames can be explained by spatial features. The performance on the remaining frames reflects the ability of models to handle cases in which temporal information is necessary much better than existing benchmarks.

In Table 3 we report the model performances on this meta-benchmark derived from LEDOV and DIEM. As indicated by our previous analysis, all models considered in this work perform poorly. As this benchmark was derived from failure cases of our model, the performance reduction of our model is disproportionally large. When using DeepGaze II as a baseline model, our model performs much better in this meta-benchmark (see supplement for details).

5 Discussion

Human gaze on dynamic stimuli such as videos is hypothesized to be strongly driven by temporal patterns in the stimuli, e.g., temporal popup and motion

	Meta-Benchmark: LEDOV & DIEM						
Model	IG	%	AUC	CC	KLdiv	NSS	SIM
Center bias	0	0	0.853	0.274	2.580	1.679	0.195
DeepVS [19]	-	-	0.854	0.337	2.599	2.152	0.225
SalEMA [31]	-	-	0.887	0.477	2.584	2.596	0.394
STRA-Net [28]	-	-	0.889	0.497	2.681	2.658	0.39
ACLNet [47]	-	-	0.891	0.483	2.044	2.579	0.377
TASED-Net [33]	-	-	0.893	**0.583**	2.995	**2.855**	**0.430**
DeepGaze MR	0.528	24.2	0.898	0.454	1.568	2.458	0.365
DeepGaze II [27]	**0.787**	**36.1**	**0.908**	0.507	**1.420**	2.693	0.389
Gold Standard	2.182	100.0	0.948	-	-	5.093	-

Table 3: Performance of state-of-the-art models on our proposed meta-benchmark, which discards frames in which the information gain of our model is more than 1bit less as the gold standard. As our model cannot exploit temporal patterns, the reported performances reflect the ability to handle cases in which temporal information is needed to predict where people look much better.

[16, 8]. In this work, we measured the importance of temporal features in video saliency. To that end, we developed and analysed DeepGaze MR, a static baseline model predicting gaze positions on the LEDOV dataset, and compared it's performance to a gold standard model. DeepGaze MR is adapted from the successful DeepGaze II model for still images and is not able to learn temporal patterns by design. Nevertheless, our model outperforms previous state of the art with a large margin on the LEDOV dataset and captures 75% of the explainable information gain.

When we analyzed failure cases of our model, we found clear temporal effects that drove the subject's gaze such as sudden appearances and movements and, to a certain degree, also interactions. We found that the gold standard performance and therefore the consistency among subjects is very high in those cases. This confirms the hypothesis that temporal patterns are an important factor influencing human gaze placement.

Given this importance of temporal effects, we would expect a good video saliency model to predict human gaze in those cases well. While our model wasn't able to capture those effects by design, we found that all other models we tested consistently failed to capture those effects either. In particular, this is the case also for models like DeepVS, ACLNet, STRA-Net and TASED-Net that have explicitly been designed to capture temporal patterns.

We argue that the main reason for this shortcoming is a deficiency of the datasets used to train video saliency models. Temporal patterns in the videos can influence gaze placement in ways that are highly consistent over subjects (Figure 4, see also [16, 8]). However, these effects turn out to be rare compared to the influence of spatial patterns such as faces on gaze placement. We suppose that they are so rare that current state-of-the-art models do not benefit from investing modelling capacity into modelling them. This difficulty for learning-

based approaches to handle rare, but important, events correctly is a general problem relevant for many fields. In autonomous driving, for example, it is crucial to handle rare events correctly, e.g. when children running onto the street.

Several aspects can contribute to tackling this issue: the model architecture, the loss function and the datasets.

Adding general temporal modelling components, as done by previous works on video saliency, has shown to be ineffective to learn temporal effects. However, our study reveals distinct temporal effects on human gaze. Models might benefit from adding modules that are explicitly designed to detect effects that we know to be relevant, such as appearing objects.

To evaluate models predicting gaze on videos, image-based metrics are typically applied per frame and averaged. As a result, some of the failure cases seen above do not substantially affect the overall model performance as those effects tend to be short compared to the whole video. This is opposed to our subjective impression of the clear failure of the model on those samples. A loss function that penalizes such failures more visibly would align the benchmark results better with human judgement.

We see the most fundamental need for improvement in the datasets. Obviously, one could explicitly collect and add cases of relevant temporal patterns to the training datasets. In particular, it would be possible to have multiple validation datasets tailored towards effects that might be considered relevant, such as appearing objects, motion and interactions. In this way, one can quantitatively judge how well new models incorporate effects that researchers consider relevant for understanding behaviour, but that are rare in the usual training datasets.

Finally, we introduced a meta-benchmark derived from existing datasets that allows to quantify the ability of models to handle those temporal effects much better: Instead of averaging performances over all frames, we only consider frames in which the information gain of our model is more than 1bit smaller than the gold standard. As our model cannot learn temporal patterns, only frames are discarded in which spatial information is sufficient. The low performance of existing models on this meta-dataset confirms our previous analysis. We will make a list of the frames we considered in this study available. In the future, our proposed benchmark could be improved by considering more datasets and by improving our spatial baseline model.

Acknowledgements. This work was supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): Germany's Excellence Strategy – EXC 2064/1 – 390727645 and SFB 1233, Robust Vision: Inference Principles and Neural Mechanisms, TP 3, project number: 276693517. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Matthias Tangemann.

References

1. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer Normalization. arXiv:1607.06450 [cs, stat] (Jul 2016)
2. Bak, C., Kocak, A., Erdem, E., Erdem, A.: Spatio-Temporal Saliency Networks for Dynamic Saliency Prediction. IEEE Transactions on Multimedia **20**(7), 1688–1698 (Jul 2018). https://doi.org/10.1109/TMM.2017.2777665
3. Bazzani, L., Larochelle, H., Torresani, L.: Recurrent Mixture Density Network for Spatiotemporal Visual Attention. In: ICLR 2017 (2017)
4. Borji, A., Itti, L.: State-of-the-Art in Visual Attention Modeling. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(1), 185–207 (Jan 2013). https://doi.org/10.1109/TPAMI.2012.89
5. Bylinskii, Z., Judd, T., Borji, A., Itti, L., Durand, F., Oliva, A., Torralba, A.: MIT Saliency Benchmark. http://saliency.mit.edu/
6. Bylinskii, Z., Kim, N.W., O'Donovan, P., Alsheikh, S., Madan, S., Pfister, H., Durand, F., Russell, B., Hertzmann, A.: Learning Visual Importance for Graphic Designs and Data Visualizations. In: Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology. pp. 57–69. UIST '17, ACM, New York, NY, USA (2017). https://doi.org/10.1145/3126594.3126653
7. Cichy, R.M., Kaiser, D.: Deep Neural Networks as Scientific Models. Trends in Cognitive Sciences **23**(4), 305–317 (Apr 2019). https://doi.org/10.1016/j.tics.2019.01.009
8. Dorr, M., Martinetz, T., Gegenfurtner, K.R., Barth, E.: Variability of eye movements when viewing dynamic natural scenes. Journal of Vision **10**(10), 28–28 (Aug 2010). https://doi.org/10.1167/10.10.28
9. Eysenck, M.W., Keane, M.T.: Cognitive Psychology: A Student's Handbook. Psychology Press, London, sixth edn. (Jan 2010)
10. Fang, Y., Lin, W., Chen, Z., Tsai, C.M., Lin, C.W.: A Video Saliency Detection Model in Compressed Domain. IEEE Transactions on Circuits and Systems for Video Technology **24**(1), 27–38 (Jan 2014). https://doi.org/10.1109/TCSVT.2013.2273613
11. Guo, C., Zhang, L.: A Novel Multiresolution Spatiotemporal Saliency Detection Model and Its Applications in Image and Video Compression. IEEE Transactions on Image Processing **19**(1), 185–198 (Jan 2010). https://doi.org/10.1109/TIP.2009.2030969
12. He, S., Tavakoli, H.R., Borji, A., Mi, Y., Pugeault, N.: Understanding and Visualizing Deep Visual Saliency Models. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10206–10215 (Jun 2019)
13. Hossein Khatoonabadi, S., Vasconcelos, N., Bajic, I.V., Shan, Y.: How Many Bits Does it Take for a Stimulus to Be Salient? In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5501–5510 (Jun 2015)
14. Hou, X., Zhang, L.: Dynamic visual attention: Searching for coding length increments. In: Koller, D., Schuurmans, D., Bengio, Y., Bottou, L. (eds.) Advances in Neural Information Processing Systems 21. pp. 681–688. Curran Associates, Inc. (2009)
15. Huang, X., Shen, C., Boix, X., Zhao, Q.: SALICON: Reducing the Semantic Gap in Saliency Prediction by Adapting Deep Neural Networks. In: The IEEE International Conference on Computer Vision (ICCV). pp. 262–270 (Dec 2015)
16. Itti, L.: Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. Visual Cognition **12**(6), 1093–1123 (Aug 2005). https://doi.org/10.1080/13506280444000661

17. Itti, L., Koch, C.: Computational modelling of visual attention. Nature Reviews Neuroscience **2**(3), 194–203 (Mar 2001). https://doi.org/10.1038/35058500
18. Itti, L., Koch, C., Niebur, E.: A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence **20**(11), 1254–1259 (Nov 1998). https://doi.org/10.1109/34.730558
19. 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)
20. Jiang, L., Xu, M., Wang, Z.: Predicting Video Saliency with Object-to-Motion CNN and Two-layer Convolutional LSTM. arXiv:1709.06316 [cs] (Sep 2017)
21. Jost, T., Ouerhani, N., von Wartburg, R., Müri, R., Hügli, H.: Assessing the contribution of color in visual attention. Computer Vision and Image Understanding **100**(1-2), 107–123 (2005). https://doi.org/10.1016/j.cviu.2004.10.009
22. Judd, T., Durand, F., Torralba, A.: A Benchmark of Computational Models of Saliency to Predict Human Fixations. MIT Tech Report (Jan 2012)
23. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization. In: ICLR 2015 (May 2015)
24. Kümmerer, M., Theis, L., Bethge, M.: Deep Gaze I: Boosting Saliency Prediction with Feature Maps Trained on ImageNet. In: ICLR Workshops 2015 (May 2015)
25. Kümmerer, M., Wallis, T.S.A., Bethge, M.: Information-theoretic model comparison unifies saliency metrics. Proceedings of the National Academy of Sciences **112**(52), 16054–16059 (Dec 2015). https://doi.org/10.1073/pnas.1510393112
26. Kümmerer, M., Wallis, T.S.A., Bethge, M.: Saliency Benchmarking Made Easy: Separating Models, Maps and Metrics. In: The European Conference on Computer Vision (ECCV). pp. 770–787 (Sep 2018)
27. 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)
28. 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
29. Le Meur, O., Baccino, T.: Methods for comparing scanpaths and saliency maps: Strengths and weaknesses. Behavior Research Methods **45**(1), 251–266 (Mar 2013). https://doi.org/10.3758/s13428-012-0226-9
30. Leborán, V., García-Díaz, A., Fdez-Vidal, X.R., Pardo, X.M.: Dynamic Whitening Saliency. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(5), 893–907 (May 2017). https://doi.org/10.1109/TPAMI.2016.2567391
31. 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)
32. Mathe, S., Sminchisescu, C.: Actions in the Eye: Dynamic Gaze Datasets and Learnt Saliency Models for Visual Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **37**(7), 1408–1424 (Jul 2015). https://doi.org/10.1109/TPAMI.2014.2366154
33. 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)
34. Mital, P.K., Smith, T.J., Hill, R.L., Henderson, J.M.: Clustering of Gaze During Dynamic Scene Viewing is Predicted by Motion. Cognitive Computation **3**(1), 5–24 (Mar 2011). https://doi.org/10.1007/s12559-010-9074-z

35. Pan, J., Ferrer, C.C., McGuinness, K., O'Connor, N.E., Torres, J., Sayrol, E., Giro-i-Nieto, X.: SalGAN: Visual Saliency Prediction with Generative Adversarial Networks. arXiv:1701.01081 [cs] (Jan 2017)
36. Peters, R.J., Iyer, A., Itti, L., Koch, C.: Components of bottom-up gaze allocation in natural images. Vision Research **45**(18), 2397–2416 (Aug 2005). https://doi.org/10.1016/j.visres.2005.03.019
37. Rajashekar, U., Cormack, L.K., Bovik, A.C.: Point-of-gaze analysis reveals visual search strategies. In: Human Vision and Electronic Imaging IX. vol. 5292, pp. 296–306. International Society for Optics and Photonics (Jun 2004). https://doi.org/10.1117/12.537118
38. Ren, Z., Gao, S., Chia, L.T., Rajan, D.: Regularized Feature Reconstruction for Spatio-Temporal Saliency Detection. IEEE Transactions on Image Processing **22**(8), 3120–3132 (Aug 2013). https://doi.org/10.1109/TIP.2013.2259837
39. Rosenholtz, R.: A simple saliency model predicts a number of motion popout phenomena. Vision Research **39**(19), 3157–3163 (Oct 1999). https://doi.org/10.1016/S0042-6989(99)00077-2
40. Rudoy, D., Goldman, D.B., Shechtman, E., Zelnik-Manor, L.: Learning Video Saliency from Human Gaze Using Candidate Selection. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1147–1154 (Jun 2013)
41. Seo, H.J., Milanfar, P.: Static and space-time visual saliency detection by self-resemblance. Journal of Vision **9**(12), 15–15 (Nov 2009). https://doi.org/10.1167/9.12.15
42. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. In: ICLR 2015 (May 2015)
43. Sun, Z., Wang, X., Zhang, Q., Jiang, J.: Real-Time Video Saliency Prediction Via 3D Residual Convolutional Neural Network. IEEE Access **7**, 147743–147754 (Oct 2019). https://doi.org/10.1109/ACCESS.2019.2946479
44. Tatler, B.W.: The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. Journal of Vision **7**(14), 4–4 (Nov 2007). https://doi.org/10.1167/7.14.4
45. Treisman, A.M., Gelade, G.: A feature-integration theory of attention. Cognitive Psychology **12**(1), 97–136 (Jan 1980). https://doi.org/10.1016/0010-0285(80)90005-5
46. Vig, E., Dorr, M., Cox, D.: Large-Scale Optimization of Hierarchical Features for Saliency Prediction in Natural Images. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2798–2805 (2014)
47. 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)
48. Wang, W., Song, H., Zhao, S., Shen, J., Zhao, S., Hoi, S.C.H., Ling, H.: Learning Unsupervised Video Object Segmentation Through Visual Attention. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3064–3074 (Jun 2019)
49. Wilming, N., Betz, T., Kietzmann, T.C., König, P.: Measures and Limits of Models of Fixation Selection. PLOS ONE **6**(9), e24038 (Dec 2011). https://doi.org/10.1371/journal.pone.0024038
50. Wu, X., Wu, Z., Zhang, J., Ju, L., Wang, S.: SalSAC: A Video Saliency Prediction Model with Shuffled Attentions and Correlation-based ConvLSTM. In: Thirty-Fourth AAAI Conference on Artificial Intelligence. AAAI Press (Feb 2020)

51. Zhang, L., Tong, M.H., Cottrell, G.W.: SUNDAy: Saliency Using Natural Statistics for Dynamic Analysis of Scenes. In: Proceedings of the 31st Annual Meeting of the Cognitive Science Society. pp. 2944–2949. AAAI Press, Cambridge, MA (2009)
52. Zhong, S.h., Liu, Y., Ren, F., Zhang, J., Ren, T.: Video Saliency Detection via Dynamic Consistent Spatio-Temporal Attention Modelling. In: Twenty-Seventh AAAI Conference on Artificial Intelligence. AAAI Press (Jul 2013)
53. Zhou, F., Bing Kang, S., Cohen, M.F.: Time-Mapping Using Space-Time Saliency. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3358–3365 (Jun 2014)