Temporal Regularization of Saliency Maps in Egocentric Videos

Panagiotis Linardos[1], Eva Mohedano[2], Cathal Gurrin[2], Monica Cherto[1], and Xavier Giro-i-Nieto[1]

[1] UPC Universitat Politcnica de Catalunya, Campus Nord, Calle Jordi Girona, 1-3, 08034 Barcelona
[2] DCU Dublin City University, Glasnevin, Whitehall, Dublin 9, Ireland
linardos.akis@gmail.com

Abstract. This work explores how temporal regularization in egocentric videos may have a positive or negative impact in saliency prediction depending on the viewer behavior. Our study is based on the new EgoMon dataset, which consists of seven videos recorded by three subjects in both free-viewing and task-driven set ups. We predict a frame-based saliency prediction over the frames of each video clip, as well as a temporally regularized version based on deep neural networks. Our results indicate that the NSS saliency metric improves during task-driven activities, but that it clearly drops during free-viewing. Encouraged by the good results in task-driven activities, we also computed and publish the saliency maps for the EPIC Kitchens dataset.

Keywords: Computer Vision · Egocentric Vision· Saliency · Visual Attention

1 Introduction

Saliency prediction refers to the task of estimating which regions of an image have a higher probability of being observed by a viewer. The result of such predictions is expressed under the form of a saliency map (heat map), in which higher values are aligned with the pixel locations with higher chances to attract the viewer's attention. This information can be used for multiple applications, such as a higher quality coding of the salient regions [24], spatial-aware feature weighting [17], or image retargeting [22].This task has been extensively explored in set ups where the viewer is asked to observe an image [10, 8, 13, 2] or video [23] depicting a scene.

Like in most aspects of computer vision, deep learning has been dominating conventional methods in the study of saliency and static scenes have received most of the scientific interest. Datasets in the context of static scene viewing are measured accurately with eye-trackers or quite less accurately but at a much larger scale using mouse clicks or webcams.

Our work focuses in the case of egocentric vision, which presents the particularity of having the viewer immersed in the scene. Saliency information is collected using a wearable sensor, which detects the gaze of the subject as he interacts with the environment. In this case, the user is not only free to fixate the gaze over any region, but also to change the framing of the scene with his head motion. When collecting datasets, this set up also differs from others in which the same image or video is shown to many viewers, as in this case each recording and scene is unique for each user. Egocentric saliency prediction has received the attention of other researchers in the past [6, 21], which we extend by assessing a state of the art model in video saliency prediction in an egocentric set up. In particular, we observe that including a temporal regularization over frame-based prediction results into a gain or loss of performance depending on whether the viewer is engaged in an activity or is just free-viewing the scene.

We developed our study on this new egocentric video dataset, named *EgoMon*, and added a temporal regularization on the SalGAN model [15] for image saliency prediction. Both the dataset and trained models are publicly available [3].

2 Related Work

2.1 Egocentric Datasets

The recording of an egocentric video dataset requires a wearable camera, but also a wearable eye tracker. This specificity in the hardware, together with the privacy constraints, limits the amount of public datasets for this task

The GTEA Gaze dataset was collected using Tobii eye-tracker glasses [6]. The more updated version of the dataset (EGTEA+) contains 28 hours of cooking activities from 86 unique sessions of 32 subjects. These videos come with audios and gaze tracking (30Hz) and provided with human annotations of actions (human-object interactions) and hand masks. In this work [6], saliency prediction models based on SVMs are trained separately for each activity, while in our case we train a single model and apply it to any activity.

The UT Ego Dataset [21] was collected using the Looxcie wearable (head-mounted) camera and contains four videos. Each video is 3-5 hours long, captured in a natural, uncontrolled setting. The videos capture a variety of activities such as eating, shopping, attending a lecture, driving, and cooking.

Why do you describe the approach followed in the 'Fathi2012' paper (SVMs for each activity) but you do not mention anything about the approach used in the 'UTego' paper? Maybe add one sentence describing what they do?

Maybe add a few sentences here 'selling' the EgoMon dataset. Why is it better/ what does it has that these other two dataset haven't? (i.e diversity on the activities vs just cooking activities (for the GTEA), anything to point regarding the UTego dataset? Maybe more videos, users?...)

2.2 Computational Models for Fixation Prediction in Videos

The recent success of deep neural networks for solving computer vision tasks has been recently explored in the context of video saliency prediction. These works

[3] https://imatge-upc.github.io/saliency-2018-videosalgan/

have basically applied to this domain the architectures developed in the field of action recognition.

Two-stream networks [20] combining video frames and optical flow were applied in [1] for saliency prediction, while temporal sequences modeled with RNN [5] were explored for saliency in [11]. A shortcoming in these methods is that they don't utilize the existing large-scale static fixation datasets. A more recent approach, however, has been shown to produce better results with an arguably simpler architecture based on the integration of an existing state of the art saliency model in conjunction with a ConvLSTM module [23] while utilizing a large-scale dataset which the authors presented (DHF1K). An interesting novelty in their model was the incorporation of a supervised attention mechanism into the network structure. Meanwhile, a VGG16 architecture extracts a deeper representation of the input frame. The intra-frame salient features extracted by the supervised attention mechanism are then enhanced on this representation frame by element-wise multiplication. The final output is fed to a sequential model for detection of the inter-frame temporal features. We train our model on the same large-scale DHF1K dataset, which contains 700 annotated videos for video saliency prediction.I think it is very important to justify why do we use the DHF1K for training and not directly an Egocentric dataset. Something like: ”Even though the DHF1K dataset do not consists in egocentric videos, we train our model on it since it is the largest annotated video saliency model publicity available to date, where each video is annotated by several users, in contrast to egocentric datasets where each video represents a unique scene annotated by one single user.

Similarly, the authors of [7] propose a complex convolutional architecture with four branches fused with a temporal-aware ConvLSTM layer. The structure is focused to work on videos that include other actors. Specifically, there is a pathway that follows the gaze of other actors in the scene and two pathways for when there is a rapid change, which can either be an actor leaving the scene or an entire scene change. These pathways hold no particular interest in the study of egocentric vision as there are no real rapid scene changes in real life and there is not necessarily always a focus on other people (the kitchen setting is a good example where your attention is entirely focused on objects). Their work however also uses one other pathway called the ”saliency pathway” which is simply a state of the art saliency model in conjunction with a ConvLSTM. In their ”saliency pathway” the output of the static saliency predictor is directly fed to the ConvLSTM rather than using it as an enhancer to a deeper frame representation.

In line with the reasoning of the aforementioned recent works [7, 23] we aim to focus on another state of the art model, that has not been tested in these works and has been shown to outperform most of the used saliency predictors, namely SalGAN [15].

Finally, regarding egocentric saliency prediction with deep models, Huang *et al.* [9] propose to model the bottom-up and top-down attention mechanisms on the GTEA Gaze dataset. Their approach combines a saliency prediction with a task-dependent attention that explicitly models the temporal shift of gaze fixations during different manipulation tasks.

3 EgoMon Gaze & Video Dataset

The dataset we are presenting here, namely EgoMon Gaze & Video Dataset, was produced using video cameras mounted on glasses and consists of 7 videos of 30 minutes on average. The *Tobii glasses* (Fig. 1), which are a head-mounted eye tracking system, were used for the construction of this dataset. The acquisition of the data was carried out by three different wearers of the *Tobii glasses*. The videos were recorded in Dublin (Ireland) during the months of March to May 2016.

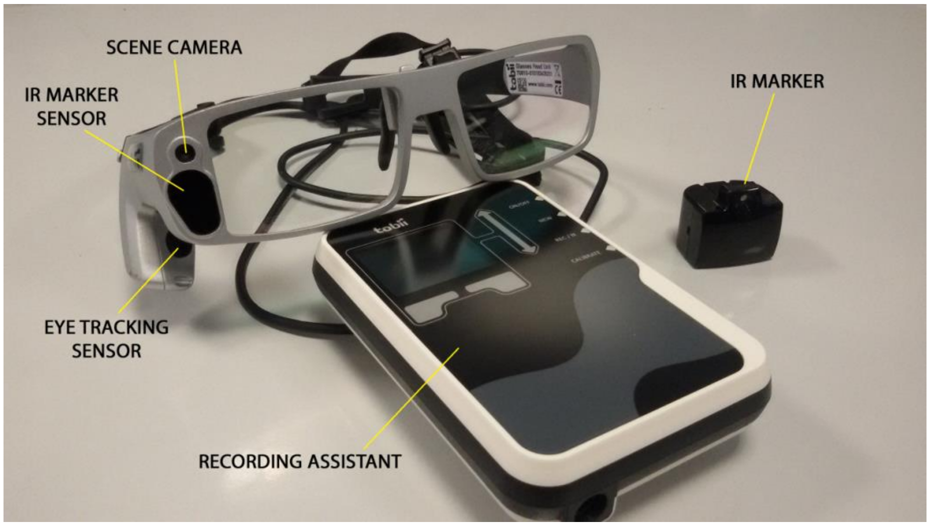

Fig. 1. A picture of the *Tobii glasses*. The eye-tracker is monocular, which means that the glasses only consider the gaze from only one eye. This device shows the exact point of gaze in real time and is complemented with a recording assistant and IR Markers. The recording assistant is used to process the recordings along with the gaze data and store them in memory, while the IR Makers are used to calibrate the glasses.

In the past, most of the scientific focus regarding saliency has been centered around the free-viewing type, where saliency arises according to the intrinsic visual characteristics of a scene (bottom-up). There have, however, also been some efforts to exploit task-driven saliency (top-down) [14]. In the recordings of EgoMon we include both free-viewing activities: a walk in a park, walking to the office, a walk in the botanic gardens, a bus ride (Fig. 3), as well as task-oriented activities: cooking an omelette (Fig. 2), listening to an oral presentation and playing cards. These amount to a total of 7 videos of average 30 minutes length.

In the case of the botanic gardens another piece of equipment was also used to extract additional information: the Narrative Clip. This is a small wearable camera that can be put through a clip on clothes and automatically takes an image every 30 seconds.

Table 1. Main features of the EgoMon Gaze & Video Dataset.

7 videos	Recorded with the Tobii eye-tracker Glasses. With gaze information plotted on them
7 videos	Clean (without the gaze information)
13428 images	Each image corresponds to 1 fps from each video.
7 text files	Gaze data extracted from each video.
73 Images corresponding of one video	Taken with the Narrative Clip.

Notable differences with other egocentric datasets can be summarized as follows: a higher diversity of recording environments (indoor and outdoor) and a high variance on the videos that constitute the dataset (movements of the wearer of the camera and changes and movements in the environment)

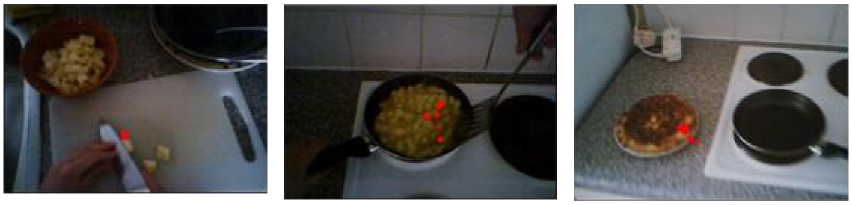

Fig. 2. Frames from an example task-driven recording in EgoMon. The wearer was engaged in the task of cooking the traditional Spanish omelette dish. The gaze points demonstrate how the participant follows with her gaze the objects that are most significant in cooking.

4 Model Architecture

SalGAN. The whole structure of our video saliency pipeline is presented in Fig.4. We use the SalGAN generator, pre-trained on the SALICON dataset [12]. The generator consists of an encoder and a decoder; the layer structure following the typical VGG-16 [19] sequence of convolutions (with the same sequence reversed in the decoder and with max pooling replaced by up-samplings). The generator was originally trained using the binary cross entropy loss function but

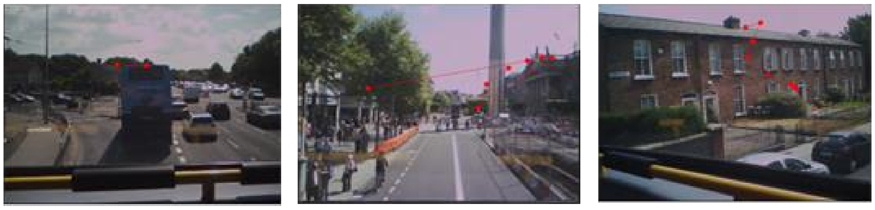

Fig. 3. Frames from an example free-viewing recording in EgoMon. The wearer was on a bus and gazed with no particular task in mind at the changing scene outside the window. The red dots represent where his gaze fell during the recording.

also through adversarial training pitted against a discriminator, composed of six 3x3 kernel convolutions interspersed with three pooling layers, and followed by three fully connected layers [15].

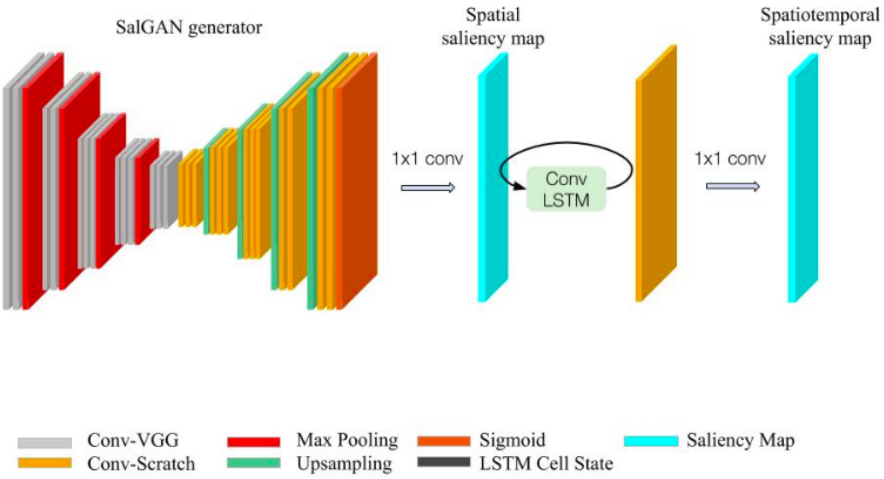

Fig. 4. Overall architecture of the proposed video saliency map estimation pipeline.

ConvLSTM. The pre-trained SalGAN generator is given an input video $\{I_t\}_t$, and outputs a sequence of static saliency maps $\{S_t\}_t$. Then the maps are fed into a ConvLSTM [18] as input. An LSTM is an autoregressive architecture that controls the flow of information in the network using 3 gates : update, forget, output. That way, at time step t, the network may control which parts of the input to keep (update gate), what is to forget as it is deemed no longer useful (forget gate) and what to output as hidden state $\{H_t\}_t$ (output gate). In the case

of ConvLSTMs, the operations at each gate are convolutions. In the following equations ’$\circ$’ represents the element-wise product, ’$*$’ a convolution operation, ’σ’ the sigmoid logistic function and ’tanh’ the hyperbolic tangent.
Update Gate:

$$u_t = \sigma({W_u}^S * S_t + {W_u}^H * H_{t\text{-}1} + {W_u}^C \circ C_{t\text{-}1} + b_u) \tag{1}$$

Forget Gate:

$$f_t = \sigma({W_f}^S * S_t + {W_f}^H * H_{t\text{-}1} + {W_f}^C \circ C_{t\text{-}1} + b_f) \tag{2}$$

Output Gate:

$$o_t = \sigma({W_o}^S * S_t + {W_o}^H * H_{t\text{-}1} + {W_o}^C \circ C_{t\text{-}1} + b_o) \tag{3}$$

Cell State Update:

$$C_t = f_t \circ C_{t\text{-}1} + u_t \circ tanh({W_C}^S * S_t + {W_C}^H * H_{t\text{-}1} + b_C) \tag{4}$$

Hidden State Update:

$$H_t = o_t \circ tanh(C_t) \tag{5}$$

To obtain a saliency map a 1x1 convolution is used at the output hidden state, so as to filter out all channels but one. We sequentially pass static saliency maps to the ConvLSTM input and, hence, obtain a sequence of time-correlated saliency maps. Our ConvLSTM model will then learn to leverage these temporal features during training.

5 Training

The content loss is computed per-pixel; concretely, this means that each pixel of the predicted saliency map is compared with its corresponding pixel in the ground truth map. The binary cross entropy, which we will be using, is the average of the individual binary cross entropies (BCE) across all pixels in this case:

$$L_{BCE} = -\frac{1}{N}\sum_{n=1}^{N} S_n log(Q_n) + (S_n - 1) log(Q_n - 1) \tag{6}$$

Where S represents the predicted saliency map and Q represents the ground truth saliency map.

To train the model, we chose the DHF1K dataset [23] because of its large-scale and diversity of contents. The dataset contains 700 annotated videos at 640 × 360 resolution. We extract the frames at their original 30 fps rate and then feed them to the pre-trained SalGAN. For that purpose, we used the model with its tuned weights as it was developed by their original authors and made publicly available on github [15]. After inferring the saliency maps from SalGAN, we resize them to 64x36 and load them into our ConvLSTM model using a batch size of 1 (corresponding to 1 video sequence at a time). At training time we backpropagate the loss through time as far as 10 frames to avoid exceeding memory capacity and potential vanishing or exploding gradients.

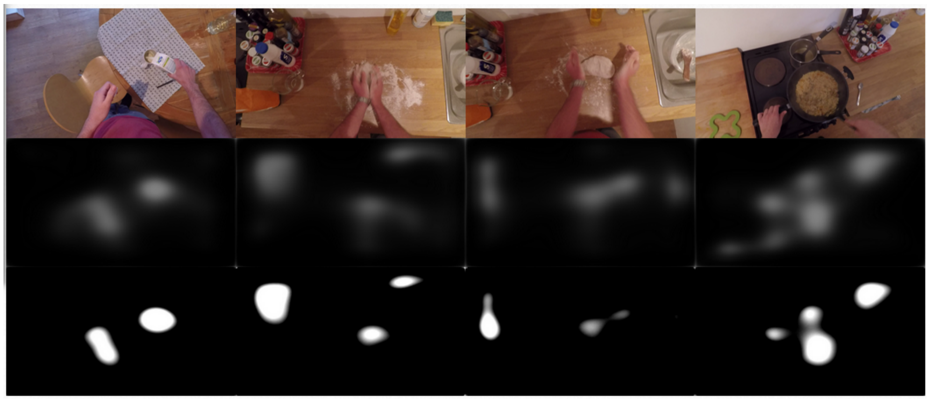

Fig. 5. We run our model on the EPIC-KITCHENS dataset [4]
. Results from 4 random frames are demonstrated here: the middle row corresponds to predictions of the vanilla SalGAN, the last row corresponds to predictions of our temporally aware model.

Table 2. Performance on DHF1K.

	AUC-J ↑	sAUC ↑	NSS ↑	CC ↑	SIM ↑
static	**0.930**	**0.834**	**2.468**	0.372	0.264
dynamic	0.744	0.722	2.246	0.302	0.260
SOA [23]	0.885	0.553	2.259	**0.415**	**0.311**

6 Evaluation

The proposed model was assessed firstly on the DHF1K dataset to evaluate its quality with respect to the state of the art in non-egocentric datasets, and later on the proposed EgoMon dataset to draw our conclusions in the egocentric domain. There is, however, an important aspect of egocentric vision datasets, such as EgoMon, that differentiates them from the large scale video saliency datasets. It's the fact that every sampled video corresponds to only one person, who is wearing the *Tobii glasses*. That means that instead of a ground truth saliency map we are limited to using fixation locations, as the gaze will fall to one particular point at any given frame and it is not possible to average over multiple subjects' gazes. Lastly, we also run our model on the newly released EPIC-KITCHENS dataset [4], but as annotations have not been made publicaly available at the time of writing of this paper, we could not assess its performance yet (some examples are shown in Fig. 5).

To evaluate our performance on SalGAN we will be using the following popular saliency metrics: Normalized Scanpath Saliency (NSS), Similarity Metric (SIM), Linear Correlation Coefficient (CC), AUC-Judd (AUC-J), and shuffled AUC (s-AUC) [3]. In the case of the egocentric datasets though, we will restrict our evaluation to just the NSS metric [16].

Table 3. Performance on all datasets (NSS metric).

	DHF1K	EgoMon
static	**2.468**	**2.079**
dynamic	2.246	1.247

To our surprise, the vanilla Salgan outperformed all the models that have been evaluated on DHF1K by [23] (table2). The surprising factor is that the vanilla version has not been trained to extract any of the temporal cues that attribute to video saliency.

Furthermore we performed evaluation on EgoMon. Vanilla SalGAN performs decently, especially considering this is a more challenging task where the gaze corresponds to the subject directly interacting with the environment. Evaluating our augmented version shows a drop in the average performance; however, there is an interesting observation if we look at the performance of each video separately (table 4). Looking at the performance it seems like our augmented version does better on all the tasks that correspond to top-down saliency while doing much worse on the bottom up saliency tasks. We theorize that this is a result of the fact that task-driven recordings have much more prevalent temporal patterns, while free-viewing is inherently random. Since our model is trained on extracting temporal patterns from saliency maps it follows in line with this theory that the performance should be higher on task-driven samples. Probably a similar observation can be drawn in the HDF1K and worth to include here if you have time. Check dynamic and static score per video, and visualize those videos with higher/lower temporal score than the static. Is there any pattern there?

Table 4. Performance on different EgoMon tasks (NSS metric). Static refers to vanilla SalGAN, while Temporal refers to the augmented version.

	free-viewing recordings (bottom-up saliency)				
	bus ride	botanical gardens	dcu park	walking office	average
Static	1.618	1.182	4.374	3.434	2.652
Temporal	0.827	0.576	1.171	1.040	0.903
	task-driven recordings (top-down saliency)				
	playing cards	presentation	tortilla		average
Static	0.971	1.360	1.617		1.315
Temporal	1.141	1.897	2.076		1.705

7 Conclusions

Our first contribution in this work is the publication of a new and innovative dataset called EgoMon Gaze & Video Dataset. To the best of our knowledge, this dataset contains the highest diversity of tasks when it comes to datasets in egocentric vision and includes both free-viewing tasks, such as a simple walk in the park, as well as mentally engaging tasks, such as cooking or playing cards.

Additionally, we extended the work on a fairly recent state of the art saliency network called SalGAN to the task of video saliency and tried an augmented version that would be trained to consider temporal information. The vanilla version of this network significantly outperformed other models that have been evaluated to the same task; however, the augmented version showed no improvements and proved to be actually worse when evaluated at free-viewing samples. When focused entirely on task-driven egocentric recordings on the other hand, there is an interesting boost in performance.

Considering our augmented model was trained on the DHF1K which is characterized by an abundance of free-viewing videos, we hypothesize that training the model on task-driven clips instead will make it better at recognizing key temporal patterns that arise at these scenarios. Another concern we would like to asses in future research, is to include relevant temporal information, that was missed at the first step by SalGAN, as input to the ConvLSTM.

References

1. Bak, C., Kocak, A., Erdem, E., Erdem, A.: Spatio-temporal saliency networks for dynamic saliency prediction. IEEE Transactions on Multimedia (2017)
2. Borji, A.: Boosting bottom-up and top-down visual features for saliency estimation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
3. Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., Durand, F.: What do different evaluation metrics tell us about saliency models? (2016). https://doi.org/10.1109/TPAMI.2018.2815601
4. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al.: Scaling egocentric vision: The epic-kitchens dataset. arXiv preprint arXiv:1804.02748 (2018)
5. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2625–2634 (2015)
6. Fathi, A., Li, Y., Rehg, J.M.: Learning to recognize daily actions using gaze. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **7572 LNCS**(PART 1), 314–327 (2012)
7. Gorji, S., Clark, J.J.: Going from image to video saliency: Augmenting image salience with dynamic attentional push. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7501–7511 (2018)
8. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: Neural Information Processing Systems (NIPS) (2006)
9. Huang, Y., Cai, M., Li, Z., Sato, Y.: Predicting gaze in egocentric video by learning task-dependent attention transition. arXiv preprint arXiv:1803.09125 (2018)
10. 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 (PAMI) (20), 1254–1259 (1998)

11. Jiang, L., Xu, M., Wang, Z.: Predicting video saliency with object-to-motion cnn and two-layer convolutional lstm. arXiv preprint arXiv:1709.06316 (2017)
12. Jiang, M., Huang, S., Duan, J., Zhao, Q.: Salicon: Saliency in context. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1072–1080 (June 2015). https://doi.org/10.1109/CVPR.2015.7298710
13. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: IEEE International Conference on Computer Vision (ICCV) (2009)
14. Murabito, F., Spampinato, C., Palazzo, S., Pogorelov, K., Riegler, M.: Top-Down Saliency Detection Driven by Visual Classification (2017)
15. 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 (2017)
16. Peters, R., Iyer, A., Itti, L., Koch, C.: Components of bottom-up gaze allocation in natural images **45**, 2397–416 (08 2005)
17. Reyes, C., Mohedano, E., McGuinness, K., O'Connor, N.E., Giro-i Nieto, X.: Where is my phone?: Personal object retrieval from egocentric images. In: Proceedings of the first Workshop on Lifelogging Tools and Applications. pp. 55–62. ACM (2016)
18. Shi, X., Chen, Z., Wang, H., Yeung, D., Wong, W., Woo, W.: Convolutional LSTM network: A machine learning approach for precipitation nowcasting. CoRR **abs/1506.04214** (2015), http://arxiv.org/abs/1506.04214
19. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. ArXiv e-prints (Sep 2014)
20. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Advances in neural information processing systems. pp. 568–576 (2014)
21. Su, Y.C., Grauman, K.: Detecting engagement in egocentric video. In: European Conference on Computer Vision. pp. 454–471. Springer (2016)
22. Theis, L., Korshunova, I., Tejani, A., Huszár, F.: Faster gaze prediction with dense networks and fisher pruning. arXiv preprint arXiv:1801.05787 (2018)
23. Wang, W., Shen, J., Guo, F., Cheng, M.M., Borji, A.: Revisiting video saliency: A large-scale benchmark and a new model. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4894–4903 (2018)
24. Zhu, S., Xu, Z.: Spatiotemporal visual saliency guided perceptual high efficiency video coding with neural network. Neurocomputing **275**, 511–522 (2018)