Two-Stream Convolutional Networks for Dynamic Saliency Prediction

Cagdas Bak
Hacettepe University
Ankara, TURKEY
cagdasbak@hacettepe.edu.tr

Aykut Erdem
Hacettepe University
Ankara, TURKEY
aykut@cs.hacettepe.edu.tr

Erkut Erdem
Hacettepe University
Ankara, TURKEY
erkut@cs.hacettepe.edu.tr

Abstract

In recent years, visual saliency estimation in images has attracted much attention in the computer vision community. However, predicting saliency in videos has received relatively little attention. Inspired by the recent success of deep convolutional neural networks based static saliency models, in this work, we study two different two-stream convolutional networks for dynamic saliency prediction. To improve the generalization capability of our models, we also introduce a novel, empirically grounded data augmentation technique for this task. We test our models on DIEM dataset and report superior results against the existing models. Moreover, we perform transfer learning experiments on SALICON, a recently proposed static saliency dataset, by finetuning our models on the optical flows estimated from static images. Our experiments show that taking motion into account in this way can be helpful for static saliency estimation.

1. Introduction

Visual saliency models have gained significant popularity in recent years. The reason behind this growing interest lies in the effective use of these models in various computer vision problems such as segmentation, object detection, video summarization and compression where extracted saliency maps are employed either as a visual feature or as a feature selection mechanism. Broadly speaking, saliency models can be split into two categories in terms of whether they try to predict human eye fixations [2] or detect salient objects [1]. The models can be further divided into static and dynamic saliency models according to which type of input they process. While static models take still images as input, dynamic models work on video sequences.

Predicting saliency in videos poses great challenges for researchers as compared to performing the same task in still images. First and foremost, the dynamic models need to consider both the spatial and the temporal characteristics of the scene when computing saliency maps. While static saliency models employ visual features such as intensity, color and orientation, for the dynamic saliency one need to focus more on the motion features since humans have a tendency to fixate their eyes on the objects in motion. In that regard, the early examples of the dynamic saliency models extend the static saliency models so that they take into account additional motion features [8, 7, 5, 30]. In addition, there are a limited number of dynamic saliency models which approach the saliency prediction in videos from a novel point of view [10, 23, 29].

There is a recent interest in applying deep learning to saliency prediction in still images [19, 25, 18, 32, 37]. These models all employ deep neural networks, and give the state-of-the-art results in most of the benchmark datasets. In this paper, our contributions are three-fold. First, inspired by the success of these models, we investigate convolutional neural networks for dynamic saliency prediction. We study the use of two-stream convolutional neural network architectures which integrate the spatial stream with the temporal stream. These network models predict a saliency map for a given video frame by simultaneously exploiting the appearance and motion information via filters learned in an end-to-end fashion. In particular, we propose two different models which consider late or early fusion strategies. To our knowledge, we are the first to apply a two-stream deep model for dynamic saliency. Second, we propose a data augmentation technique for this task to improve the generalization of the convolutional networks. Experiments on DIEM dataset [24] validate the effectiveness of our models and our data augmentation strategy. Third, we demonstrate transfer learning could be used to predict static saliency by exploitin optical flow information extracted from still images [33]. Experiments on SALICON dataset [13] show that taking motion into account can improve the prediction accuracy.

2. Related Work

Saliency estimation models in the literature are generally classified into two main groups as bottom-up approaches [22, 26, 35] and top-down approaches [16, 36]. Bottom-up approaches usually try to identify the salient re-

gions where humans fixate their eyes in an image by using low-level visual cues such as color, intensity, orientation and they don't employ any prior information about image content and context. In contrast, top-down approaches directly use semantic cues related to context and content information about scenes, and generally consider a specific task such as finding a person in a scene.

In this study, we are interested in task-free bottom-up modeling of visual attention for dynamic scenes. The early models of dynamic saliency extend the static saliency models to include motion features. For example, Cui *et al*. [5] identify salient parts of a video frame via a frequency modulated model. In particular, they carry out a spectral residual analysis on the Fourier spectrum of a video frame along the spatial and temporal planes to extract the foreground objects in motion from the background. Guo *et al*. [7] perform a similar spectral analysis on the phase spectrum of the video frames. Harel *et al*. [8] propose a saliency model where the extracted feature maps are represented by means of fully connected graphs and the final saliency map is estimated via a graph-theoretic approach. Seo and Milanfar [30] employ self similarities of spatio-temporal volumes to predict saliency.

Some researchers devise novel models that are specifically designed for dynamic saliency. For instance, Hou [10] propose a model which takes into account rarity of visual features and extracts saliency maps by performing entropy maximization over the extracted visual features. Mathe *et al*. [23] formulate the dynamic saliency prediction as a classification task and propose a learning-based saliency model to integrate several visual cues. In another study, Rudoy *et al*. [29] also propose a learning based framework for saliency prediction. Their formulation differs from Mathe *et al*.'s model [23] in the sense that they consider a sparse set of gaze locations thorough which they try to predict conditional gaze transitions over consecutive video frames.

In recent years, deep neural networks have been applied to many computer vision problems such as image classification [9], object detection [6], activity recognition [38], semantic segmentation [21]and video classification [15], giving superior results. These approaches perform hierarchical feature learning for a particular task, which generally provide better results than the handcrafted features. Motivated by these successes, a number of deep learning based saliency models have also been introduced recently [19, 25, 18, 32, 37]. Vig *et al*. [32] employ an ensemble of deep networks which are based on biologically inspired hierarchical features when predicting saliency maps. Kruthiventi *et al*. [18] adopt fully convolutional neural networks architecture for saliency estimation task. Kümmerer *et al*. [19] use deep features learned via different layers of the AlexNet [17] model and learn how to combine them for saliency estimation. Zhao *et al*. [37] perform salient object detection through a deep learning framework considering local and global image context. Pan *et al*. [25] very recently propose two convolution network based models with different layer sizes by formulating saliency prediction as a regression task. It is important to note that all these models are proposed for predicting saliency in still images not videos.

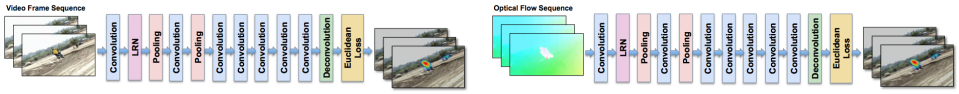
Motivated by these deep learning based saliency models, in our paper, we investigate the use of convolutional neural networks for dynamic saliency estimation. More specifically, we extract temporal information via optical flow between consecutive video frames and investigate different ways to use this additional information in saliency prediction. We model two different two-stream convolutional networks that combine spatial and temporal information by employing early and late fusion strategies. Best to our knowledge, our models are the first two-stream convolutional neural network based dynamic saliency models in the literature.

3. Models

The purpose of this study is to investigate different deep architectures for dynamic saliency prediction. Recently, deep convolutional networks provided drastically superior performance in many classification and regression tasks in computer vision. While the lower layers of these networks respond to primitive image features such as edges, corners and shared common patterns, the higher layers extract semantic information like object parts or faces. As mentioned before, such low and high-level features are shown to be both important and complementary in estimating visual saliency. Towards this end, we examine two baseline single frame networks in Figure 1(a) and 1(b) (spatial and temporal) and two two-stream networks [31] that combine spatial and temporal cues at different levels of granularity by implementing two different fusion strategies: late fusion and early fusion, as shown in Figure 2(a) and 2(b). We describe these models in detail below.

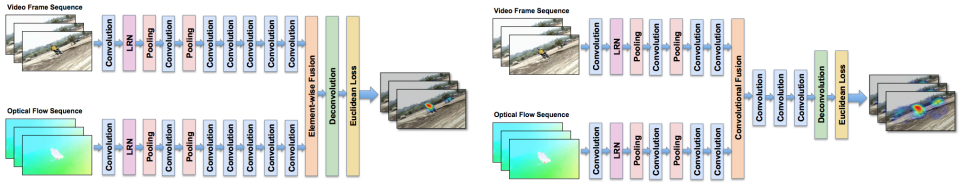
Spatial Stream Net. For the basic single-frame baseline model, we use the recently proposed static saliency model in [25]. As shown in Figure 1(a), this convolutional network resembles the VGG-M model [4] – the main difference being that the final layer is a deconvolution (fractionally strided convolution) layer to upsample to the original image size. Note that it does not use any temporal information and exploits only appearance information to predict saliency in still video frames.

Temporal Stream Net. What makes saliency prediction in videos inherently different than in images is that our attention is hugely influenced from the local motion contrast of foreground objects. To understand the contribution of temporal information to the saliency prediction, by itself, we develop a second single frame baseline. As shown in



(a) Spatial Stream Net (b) Temporal Stream Net

Figure 1. Single frame networks for dynamic saliency prediction. While one utilizes only spatial (appearance) information and accepts still video frames, the other one exploits only temporal information whose input is given in the form of optical flow images.



(a) Late Fusion Net (b) Early Fusion Net

Figure 2. The proposed two-stream convolutional network architectures for dynamic saliency prediction. The first one performs late fusion by using element-wise fusion before the deconvolution layer, whereas the second one employs Conv fusion at an early stage, after the fourth convolution layers.

Figure 1(b), this model is just a replica of the spatial stream net but the input is provided in the form of optical flow images, as in [31], computed from two subsequent frames. Some sample optical flow images are shown in Figure 3.

Late Fusion Net. The Late Fusion model is illustrated in Figure 2(a). It takes both a video frame and the corresponding optical flow image as inputs and merges together the spatial and temporal single-frame networks via element-wise fusion at the latest convolution layer. In this sense, it directly combines spatial and temporal information at the highest level. After this fusion step, it also employs a deconvolution layer to produce an upsampled saliency map as the final result.

Early Fusion Net. Early Fusion model integrates spatial and temporal streams in an early stage, to be specific by applying Conv fusion after the 4th convolution layer of the single-frame models Conv fusion. That is the corresponding feature maps from single-frame model are stacked together and them combined with a bank of 1×1 convolutions. As illustrated in Figure 2(b), this is followed by a number of convolutions and a final deconvolution layer to produce the saliency map.

3.1. Implementation Details

Network Architectures. For the single-frame models, we employ the deep convolution network proposed in [25], which takes inputs of size $640 \times 480 \times 3$ pixels and can be summarized as $C(96, 7, 3) \rightarrow LRN \rightarrow P \rightarrow C(256, 5, 2) \rightarrow P$, $C(512, 3, 1) \rightarrow C(512, 5, 2) \rightarrow C(512, 5, 2) \rightarrow C(256, 7, 3) \rightarrow C(128, 11, 5) \rightarrow C(32, 11, 5) \rightarrow C(1, 13, 6) \rightarrow D$, where $C(d, f, p)$ denotes a convolutional layer with d filters of size $f \times f$ applied to the input with padding p and stride 1. LRN symbolizes a local response normalization layer that performs a kind of lateral inhibition, and P indicates a max pooling layer over 3×3 regions with stride 2. Finally, D is a deconvolution layer with filters of size $8 \times 8 \times 1$ with stride 4 and padding 2 which upscales the final convolution results to the original size. All convolutional layers except the last one are followed by a rectified linear unit non- linearity (ReLU) layer. The spatial and the temporal stream models differ in their inputs, that is, while one accepts still images, the other accepts optical flow images.

The proposed two-stream models employ different fusion strategies to fuse together the spatial and temporal convolutional networks at different stages, as illustrated in Figure 2. In the Late Fusion model, as its name suggests, the single stream stream networks are combined after the last convolutional layer $C(1, 13, 6)$ by applying element-wise max operation, which is followed by the same deconvolution layer D in the single-frame models. On the other hand, the Early Fusion model performs Conv fusion after the fourth convolutional layer $C(512, 5, 2)$. That is, the resulting feature maps are stacked together and integrated by a convolution layer $C(512, 1, 0)$ whose weights are initialized with identity matrices. The remaining layers are the same with those of the single-frame models.

Preprocessing. In our experiments, we use DIEM (Dynamic Images and Eye Movements) dataset [24] which be described in detail in the Experiments section. Since our networks accept inputs of size $640 \times 480 \times 3$ pixels and outputs saliency maps of the same size, all videos and ground truth fixation density maps are rescaled to this size prior to training. We use the publicly available implementation of

DeepFlow [34] and we additionally extract optical flow information from the rescaled versions of subsequent video frames. Optical flow images are then generated by stacking horizontal and vertical flow components and the magnitude of the flow together. Some example optical flow images are shown in Figure 3.

Data Augmentation. Data augmentation is a widely used approach to reduce the effect of overfitting and improve generalization of neural networks. For saliency prediction, however, classical techniques such as cropping, horizontal flipping, or RGB jittering are not very suitable since they alter the setup used in the eye tracking experiments in collecting the data. The experiments in [14] reveal that humans are quite consistent about where they look on high and low-resolution versions of the same images. Hence, we process all video sequences and produce their low-resolution versions by downsampling them by a factor of 2 and 4. We note that in reducing the resolution of optical flow images the magnitude should also be rescaled to match with the downsampling rate.

Training. We use the weights of the pretrained model in [25] to set the initial weights of the spatial and temporal stream networks. In optimizing the models, we use Caffe framework [12] and employed Stochastic Gradient Descent with Euclidean distance between the predicted saliency map and the ground truth. The networks were trained over 200K iterations where we used a batch size of 2 images, momentum of 0.9 and weight decay of 0.0005, which is reduced by a factor of 0.1 at every 10K iterations.

4. Experimental Results

In the following, we first present experimental evaluation of the proposed network architectures against the state-of-the-art dynamic saliency models on DIEM dataset [24]. We then describe our transfer learning experiments on SALICON [13] dataset where we demonstrate that static saliency estimation can also benefit from using motion information.

4.1. Experiments on DIEM

We experimentally validate the effectiveness of the proposed deep dynamic saliency networks on the DIEM dataset [24]. This dataset consists of 84 high-definition natural videos including movie trailers, advertisements, etc. Each video sequence has eye fixation data collected from approximately 50 different human subjects. In our evaluation, we evaluate all of our proposed deep dynamic saliency networks (Spatial Stream Net, Temporal Stream Net, Late Fusion Net, Early Fusion Net) by considering the same experimental setup reported by Borji *et al.* in [3]. Specifically, we train each one of these networks with 64 video sequences,and test them on the remaining 20 representative videos.

In Figure 4, we provide sample qualitative results of the proposed networks on one of sample video frames along with the ground truth human fixation map. The results clearly demonstrate the importance of the motion in dynamic saliency estimation. The Spatial Stream Net, which does employ appearance information but not motion, provides an inaccurate saliency map and misses the foreground object in motion. The Temporal Stream Net gives better results, but does identify all the moving regions as salient. Late Fusion Net results in a more accurate result as it integrates the appearance features with the motion features in its final prediction layer. Early Fusion Net gives the best results as it combines spatial and temporal information in early layers, which allows to learn filters which work on combined appearance and motion information in higher layers. Sample results are also provided in the supplementary material.

We quantitatively evaluate the proposed network models by using the shuffled AUC metric [27] and the χ^2 distance. The Area-under-curve (AUC) metric treats the saliency maps as a classification map and employs the receiver operator characteristics curve to estimate the effectiveness of the predicted saliency maps in capturing the ground truth eye fixations. In particular, we employ the shuffled version of AUC (sAUC) which accounts for the center bias observed in the saliency datasets. The χ^2 distance, on the other hand, considers the saliency maps as a probability distribution map and compares the predicted map with the ground truth human fixation map accordingly. A perfect prediction model needs to give a score of 1 for the sAUC metric and needs to provide a distance close to 0 for the χ^2 distance. For each test sequence we compute the sAUC scores and the χ^2 distances at every frame and average them out. Table 1 presents the quantitative results of the proposed dynamic saliency networks. As can be seen from the table, Spatial Stream network provides the worst results in terms of both metrics. Early Fusion network in general gives better results than all the other network architectures. It can be argued that the reason behind this success lies in an early integration of the appearance and the motion features in earl layers, which allows the filters in the higher levels of the hierarchy to learn more effective features for the saliency prediction task. When we employ the data augmentation strategy that we discussed in the previous section, it further boosts the scores.

We additionally compare our Early Fusion Net model with four different methods from the literature. These are GVBS [8], PQFT [7], Hou and Zhang's [10] and Rudoy *et al.*'s [29] dynamic saliency models, which are the best performing models on the DIEM dataset. While Figure 5 presents the qualitative results on some sample video frames, Table 2 provides the quantitative evaluation results. As can be seen from these results, the proposed two-stream

Figure 3. Sample optical flow images generated for some frames of a video sequence from DIEM dataset.

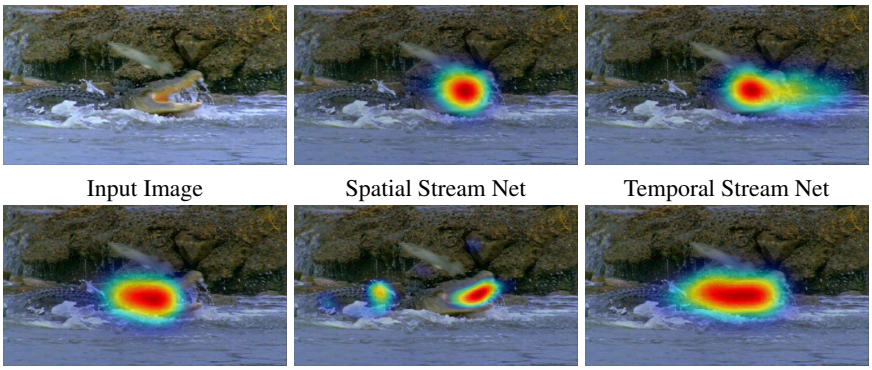


Figure 4. Comparison of the proposed network architectures. For this frame the Early Fusion Net provides the most accurate prediction as compared to the other network models.

convolutional network model with early fusion strategy outperforms all the existing models in terms of sAUC score and χ^2 metric.

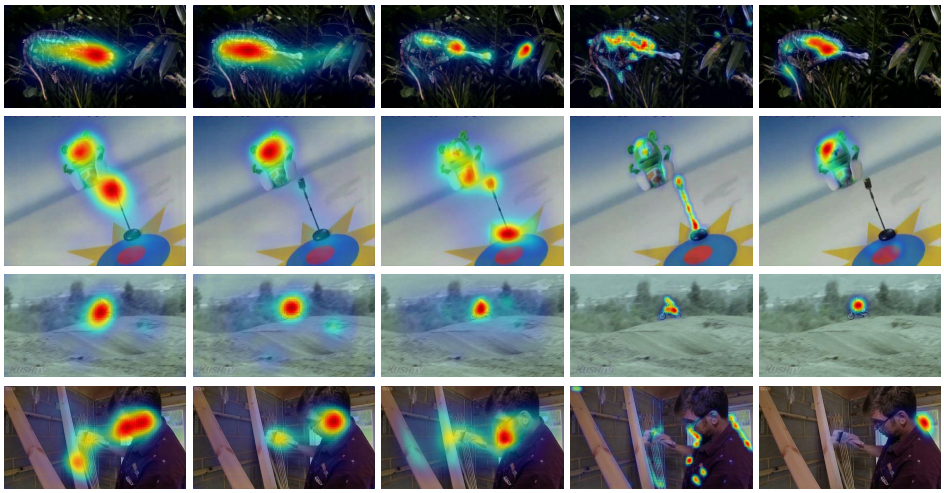
4.2. Transfer Learning Experiments on SALICON

A still image is captured in an instant, but that single image frame has enough information which allows to predict the inherent motion of the scene imaged. Motivated with this observation Walker *et al*. [33] have proposed a deep optical flow prediction model which works for static images. For this last set of experiments, we employ this prediction model to estimate the related motion map of an input image, and use it together with the original image as inputs to our two-stream Early Fusion network model. In Figure 6, we present a sample image containing a man in front of a house. It is a still image, yet we can understand the scene clearly that the man is throwing a frisbee. As illustrated, the optical flow model correctly extracts the frisbee as the moving object in the scene. This allows our dynamic saliency network to better understand the image and provide a more accurate saliency map as opposed the Deep Conv model which does only use the appearance information.

We perform out experiments on the recently proposed SALICON dataset [13]. This large-scale dataset contains 20000 natural images, all of which are taken from the MS-COCO dataset [20] and enriched with the fixation data collected via mouse cursor tracking. For evaluation, the dataset is split into 10000 training, 5000 validation and 5000 testing images. The fixation data is available only for the training and validation splits, and the tests are carried out externally by an evaluation server. For the experiments, we finetune our Early Fusion model on the training images and the optical flow images extracted from these static images using [33]. In Figure 7, we present some sample saliency maps generated by our Early Fusion network. We also present some quantitative results in Table 3 where we compare our results with some state-of-the-art static saliency models. For this evaluation, we use AUC, sAUC and Cross Correlation (CC) metrics. As shown in the table, our proposed Early Fusion network model provides the best sAUC score among the existing models, and gives highly competitive results for the remaining metrics as compared to Deep Conv network [25]. Although motion might not be the primary factor for all the images in the SALICON dataset, our results demonstrate that motion information predicted from still images could be useful.

5. Conclusion

We have presented two novel architectures for saliency prediction in videos. Our models are based on two-stream convolutional networks, which are trained in an end-to-end fashion on a large and diverse dataset, and they provide effective ways of combining spatial and temporal information. We also propose a well-founded and effective data



Ground Truth Early Fusion Net GBVS [8] PQFT [7] Hou & Zhang [10]

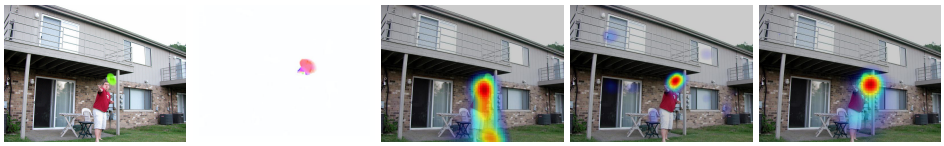
Figure 5. Qualitative comparison of the Early Fusion Net against other dynamic saliency models. Our model clearly produces better results.

	Spatial Stream	Temporal Stream	Late Fusion	Early Fusion	Early Fusion w/ Augmentation
sAUC	0.69	0.77	0.81	0.83	**0.84**
χ^2	0.48	0.40	0.31	0.29	**0.28**

Table 1. Comparison of the proposed deep dynamic saliency networks on the DIEM dataset.

	Early Fusion w/ Aug.	Rudoy	GBVS	PQFT	Hou and Zhang
sAUC	**0.84**	0.74	0.70	0.67	0.64
χ^2	**0.28**	0.31	0.47	0.52	0.57

Table 2. Comparison against the state-of-the art models on the DIEM dataset. The proposed Early Fusion network outperforms all the existing models in terms of all metrics.



Input Image Optical Flow Map Deep Conv [25] Early Fusion Net Ground Truth

Figure 6. Transfer learning for static saliency estimation. Utilizing optical flow information estimated from static images can improve the prediction quality. Our finetuned Early Fusion Net produces a better saliency map than Deep Conv [25], a static deep saliency model.

Figure 7. Sample static saliency results on some images from SALICON test set. Top row: Input images, Bottom row: Predicted saliency maps.

	AUC	sAUC	CC
Early Fusion	0.84	**0.73**	0.59
Deep Conv [25]	**0.85**	0.72	**0.62**
Shallow Conv [25]	0.83	0.66	0.59
Rare 2012 Improved [28]	0.81	0.66	0.51
Xindian	0.80	0.68	0.48
GBVS [8]	0.78	0.63	0.42
Itti [11]	0.66	0.61	0.20

Table 3. Quantitative comparison of our early fusion network against the state-of-the art models on the SALICON dataset. Although the proposed network is designed for dynamic saliency prediction, it outperforms all the existing models in terms of sAUC and gives very competitive results for the remaining metrics.

augmentation method which employs low-resolution versions of the video frames and the ground truth saliency maps. We demonstrate that our models quantitatively outperform the state-of-the-art on DIEM dataset. In addition, we show that static saliency prediction can also benefit from motion information where we finetune our model on SALICON dataset by exploiting automatically predicted optical flows from static images.

References

[1] A. Borji, M.-M. Cheng, H. Jiang, and J. Li. Salient object detection: A benchmark. *IEEE Trans. Image Processing*, 24(12):5706–5722, 2015. 1

[2] A. Borji and L. Itti. State-of-the-art in visual attention modeling. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(1):185–207, 2013. 1

[3] A. Borji, D. N. Sihite, and L. Itti. Quantitative analysis of human-model agreement in visual saliency modeling: a comparative study. *IEEE Trans. Image Processing*, 22(1):55–69, 2013. 4

[4] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *BMVC*, 2014. 2

[5] X. Cui, Q. Liu, and D. Metaxas. Temporal spectral residual: fast motion saliency detection. In *ACM MM*, pages 617–620, 2009. 1, 2

[6] R. Girshick. Fast R-CNN. In *ICCV*, pages 1440–1448, 2015. 2

[7] C. Guo, Q. Ma, and L. Zhang. Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. In *CVPR*, pages 1–8, 2008. 1, 2, 4, 6

[8] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *NIPS*, pages 545–552, 2006. 1, 2, 4, 6, 7

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2

[10] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *CVPR*, pages 1–8, 2007. 1, 2, 4, 6

[11] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, (11):1254–1259, 1998. 7

[12] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093, 2014. 4

[13] M. Jiang, S. Huang, J. Duan, and Q. Zhao. Salicon: Saliency in context. In *CVPR*, pages 1072–1080, 2015. 1, 4, 5

[14] T. Judd, F. Durand, and A. Torralba. Fixations on low-resolution images. *Journal of Vision*, 11(4):14–14, 2011. 4

[15] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, pages 1725–1732, 2014. 2

[16] A. Kocak, K. Cizmeciler, A. Erdem, and E. Erdem. Top down saliency estimation via superpixel-based discriminative dictionaries. In *BMVC*, 2014. 1

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012. 2

[18] S. S. S. Kruthiventi, K. Ayush, and R. V. Babu. Deepfix: A fully convolutional neural network for predicting human eye fixations. arXiv preprint arXiv:1510.02927, 2015. 1, 2

[19] M. Kummerer, L. Theis, and M. Bethge. Deep gaze I: Boosting saliency prediction with feature maps trained on imagenet. In *ICLR Workshop*, 2015. 1, 2

[20] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. 2014. 5

[21] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015. 2

[22] R. Margolin, A. Tal, and L. Zelnik-Manor. What makes a patch distinct? In *CVPR*, pages 1139–1146, 2013. 1

[23] S. Mathe and C. Sminchisescu. Dynamic eye movement datasets and learnt saliency models for visual action recognition. In *ECCV*, pages 842–856, 2012. 1, 2

[24] P. K. Mital, T. J. Smith, R. L. Hill, and J. M. Henderson. Clustering of gaze during dynamic scene viewing is predicted by motion. *Cognitive Computation*, 3(1):5–24, 2011. 1, 3, 4

[25] J. Pan, E. Sayrol, X. Giro-i Nieto, K. McGuinness, and N. O'Connor. Shallow and deep convolutional networks for saliency prediction. In *CVPR*, 2016. 1, 2, 3, 4, 5, 6, 7

[26] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung. Saliency filters: Contrast based filtering for salient region detection. In *CVPR*, pages 733–740, 2012. 1

[27] N. Riche, M. Duvinage, M. Mancas, B. Gosselin, and T. Dutoit. Saliency and human fixations: state-of-the-art and study of comparison metrics. In *ICCV*, pages 1153–1160, 2013. 4

[28] N. Riche, M. Mancas, M. Duvinage, M. Mibulumukini, B. Gosselin, and T. Dutoit. Rare2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis. *Signal Processing: Image Communication*, 28(6):642–658, 2013. 7

[29] D. Rudoy, D. Goldman, E. Shechtman, and L. Zelnik-Manor. Learning video saliency from human gaze using candidate selection. In *CVPR*, pages 1147–1154, 2013. 1, 2, 4

[30] H. J. Seo and P. Milanfar. Static and space-time visual saliency detection by self-resemblance. *Journal of Vision*, 9(12):15–15, 2009. 1, 2

[31] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, pages 568–576, 2014. 2, 3

[32] E. Vig, M. Dorr, and D. Cox. Large-scale optimization of hierarchical features for saliency prediction in natural images. In *CVPR*, pages 2798–2805, 2014. 1, 2

[33] J. Walker, A. Gupta, and M. Hebert. Dense optical flow prediction from a static image. In *ICCV*, pages 2443–2451, 2015. 1, 5

[34] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid. Deepflow: Large displacement optical flow with deep matching. In *ICCV*, pages 1385–1392, 2013. 4

[35] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang. Saliency detection via graph-based manifold ranking. In *CVPR*, pages 3166–3173, 2013. 1

[36] J. Yang and M.-H. Yang. Top-down visual saliency via joint CRF and dictionary learning. In *CVPR*, pages 2296–2303, 2012. 1

[37] R. Zhao, W. Ouyang, H. Li, and X. Wang. Saliency detection by multi-context deep learning. In *CVPR*, pages 1265–1274, 2015. 1, 2

[38] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *NIPS*, pages 487–495, 2014. 2