How much time do you have? Modeling multi-duration saliency

Camilo Fosco*[1], Anelise Newman*[1], Pat Sukhum[2], Yun Bin Zhang[2], Nanxuan Zhao[3], Aude Oliva[1], and Zoya Bylinskii[4]

[1]Massachusetts Institute of Technology, [2]Harvard University, [3]City University of Hong Kong,
[4]Adobe Research
{camilolu, apnewman, oliva}@mit.edu, {psukhum, ybzhang}
@g.harvard.edu, nanxuanzhao@gmail.com, bylinski@adobe.com

Abstract

What jumps out in a single glance of an image is different than what you might notice after closer inspection. Yet conventional models of visual saliency produce predictions at an arbitrary, fixed viewing duration, offering a limited view of the rich interactions between image content and gaze location. In this paper we propose to capture gaze as a series of snapshots, by generating population-level saliency heatmaps for multiple viewing durations. We collect the CodeCharts1K dataset, which contains multiple distinct heatmaps per image corresponding to 0.5, 3, and 5 seconds of free-viewing. We develop an LSTM-based model of saliency that simultaneously trains on data from multiple viewing durations. Our Multi-Duration Saliency Excited Model (MD-SEM) achieves competitive performance on the LSUN 2017 Challenge with 57% fewer parameters than comparable architectures. It is the first model that produces heatmaps at multiple viewing durations, enabling applications where multi-duration saliency can be used to prioritize visual content to keep, transmit, and render.

1. Introduction

How long an observer has to examine an image determines what they notice and what tasks they can complete. Despite this dependency of viewing behavior on viewing time, most models of visual attention predict saliency at an arbitrary duration because they are trained on data accumulated over 3 or 5 seconds of viewing [9, 22, 30, 33, 43]. On the other hand, scanpath models, which predict individual gaze trajectories over time, struggle to summarize the attention patterns of a population [2, 3, 27, 42, 53].

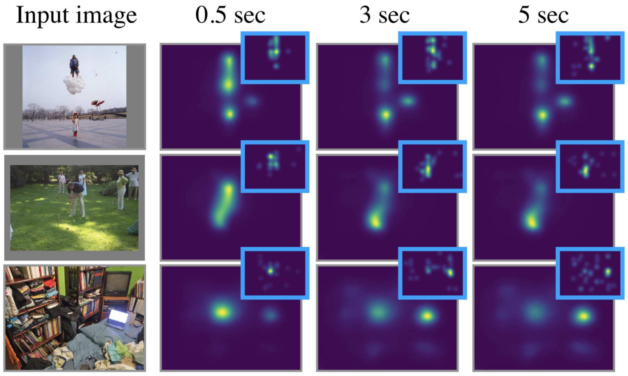


Figure 1: Predictions of our Multi-Duration Saliency Excited Model at three viewing durations. Images are from the Abnormal Objects [46], SALICON [33], and Eye-Crowd [34] datasets (top to bottom). Insets with blue borders contain human ground truth collected using the CodeCharts UI.

In this paper, we introduce the concept of **multi-duration saliency**, which captures multiple attention snapshots corresponding to different viewing durations (Fig. 1)[1]. This offers richer insight into how gaze evolves over time than conventional saliency, while providing a more robust representation than scanpaths (Fig. 2). We leverage an efficient crowdsourcing methodology for collecting large scale human attention data at several viewing durations (Fig. 3). We use it to assemble CodeCharts1K, a dataset of 1000 images with viewing patterns at three durations: 0.5, 3, and 5 seconds. Our data shows that human gaze patterns are highly consistent at each viewing duration but can differ across durations, proving that saliency depends predictably on viewing duration.

*Equal contribution.

[1] Data, code, and models available at: http://multiduration-saliency.csail.mit.edu/

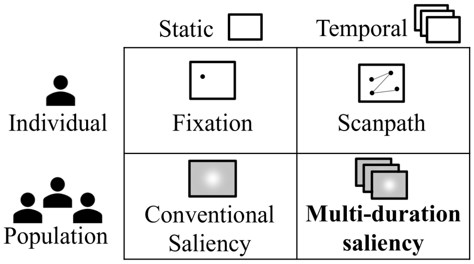


Figure 2: Multi-duration saliency compared to other gaze prediction tasks: combining the stability and generalizability of a population-level metric with rich temporal data.

To capture this dependency, our multi-duration saliency model takes an image as input and predicts three distinct saliency maps for three different durations. This lightweight model achieves competitive performance when evaluated at a single duration, and outperforms baseline models trained to predict multiple durations. We show that the predicted saliency maps can be used as input to applications such as image cropping, compression/rendering, and captioning to tailor them to different contexts based on viewing time.

2. Related work

Crowdsourcing attention: Recent efforts at collecting large-scale attention data have used interfaces that can be run remotely, without an eye tracker [11, 33, 37, 38, 41, 45]. Moving-window approaches like SALICON [33] and BubbleView [37] involve using the mouse cursor to inspect small, bubble-shaped regions of blurred images. While providing cheap, high-fidelity data, these interfaces have two downsides: (1) blurring images distorts the visual content and interferes with feature sizes at different resolutions, and (2) moving the cursor requires a different process than moving the eyes, which can impact which image regions are explored [37, 52]. **The CodeCharts UI** [41, 49] captures attention *without* distorting the underlying image or relying on mouse movements. It can account for over 80% of human consistency and outperforms other attention-capturing interfaces, including BubbleView, at approximating eye movements [41]. Because it allows for fine-grained control of image presentation time, we use it to capture multi-duration attention data.

Saliency modeling: The large-scale attention data captured using SALICON [33] and BubbleView [37] enabled training neural network models of saliency (e.g., [11, 22, 30, 43]). The top performers on the MIT Saliency Benchmark [9] were trained on SALICON data and have opened a wide performance gap to the previous, traditional models of saliency [12]. Driven by such improvements in efficiency and accuracy, saliency models have found wide use in applications like image cropping, retargeting, and view-finding for improved composition [6, 15, 24, 54].

Scanpath modeling: Ground truth saliency maps are computed by accumulating gaze locations of multiple observers over a fixed viewing duration, which averages out paterns in gaze location over time. A complementary approach to representing and modeling human attention is via scanpaths: the sequence of gaze locations that an observer makes on an image over time. Scanpath analysis and modeling is complicated by the fact that individual differences are huge at the level of single gaze locations [3, 39]. This hides the fact that different permutations of traversing image content may nevertheless correspond to a similar allocation of attention to the respective image regions.

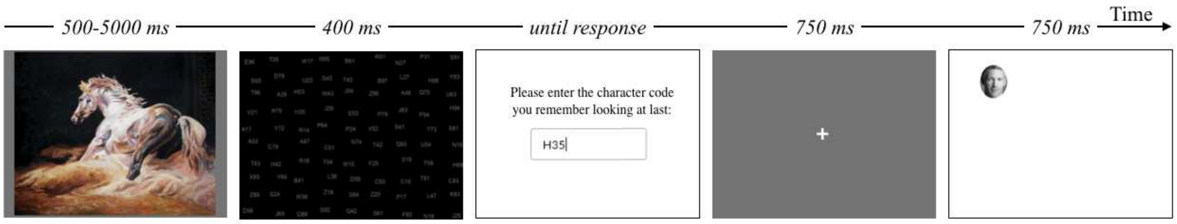
How does multi-duration saliency relate to conventional saliency and scanpaths? We propose an intermediate representation of attention that maintains the robustness of population-level saliency modeling and the temporal resolution of scanpaths (Fig. 2). We introduce **multi-duration saliency** as a way to snapshot attention at a few distinct time points. Unlike conventional saliency, multi-duration saliency accounts for the effect of viewing duration on gaze patterns and provides insight into how attention evolves over time. However, in contrast to scanpath prediction, multi-duration saliency is a population-level metric that produces stable, interpretable, and generalizable attention heatmaps. This framing addresses questions like: what content do people prioritize, and what is initially attention grabbing versus noticeable only after seconds of viewing?

3. Collecting multi-duration saliency

In this section, we introduce a scalable approach to measuring multi-duration saliency that uses a web-based interface as a proxy for eye tracking.

CodeCharts UI: In the CodeCharts methodology [41, 49], participants view an image for 500-5000 milliseconds followed by a jittered grid of three-character codes ("codechart"). They then self-report the first three-character code they see when the image vanishes (Fig. 3a-c). By construction, participants report the region of the image they were looking at last. The steps in Fig. 3a-d are repeated for dozens of images. The task also contains validation trials consisting of randomly placed cropped faces [5], where we expect the participant to enter a code that overlaps with the face. To ensure data quality, we filter out participants who enter nonexistent codes, fail over 25% of validation images, or look at the same spot repeatedly (more details in the supplement). We collect 50 gaze points per image per viewing duration, which produces on average 44 valid gaze points after filtering. We blur all gaze points (with a Gaussian sigma of 50 pixels) to produce an attention heatmap.

Pilot experiments: We ran an initial experiment with 50 images from the OSIE dataset [55] to analyze differences in gaze patterns across viewing durations. We collected gaze


500-5000 ms
400 ms
until response
750 ms
750 ms
Time
Please enter the character code you remember looking at last:
H35
+

a) image for viewing b) codechart c) code entry d) fixation cross e) validation trial

Figure 3: CodeCharts UI task flow. (a) Participants view an image for a duration chosen by the experimenter. (b) A codechart is briefly flashed on the screen. (c) Participants report the character code they remember looking at last, which indicates where they were gazing on an image. This repeats for a sequence of images, separated by a fixation cross (d) to re-center gaze. Validation trials (e) are interspersed among the experimental sequence to check if participants move their gaze as expected.

locations for 6 durations: 0.5, 1, 2, 3, 4, and 5 seconds. Attention heatmaps at 0.5, 3, and 5 seconds of viewing were the most distinct from each other (Fig. 4), so we collect further data at these 3 durations. Attention heatmaps at 3 seconds most closely matched the ground truth OSIE data, originally also collected at 3 seconds (Pearson's CC = .62, averaged over 50 OSIE images). This further validates the ability of CodeCharts data to model natural human gaze.

CodeCharts1K: Our collected dataset contains a variety of image types to provide a broad picture of differences in attention over time. We used 500 images from SALICON [33], 130 from LaMem [36], 120 from CAT2000 [7][2], 100 from EyeCrowd [34], 100 from a mix of Abnormal Objects [46] and Out-of-Context Objects [19], and 50 from the Stanford 40K Actions dataset [56][3]. Images were padded to the same aspect ratio and resized in-browser to fit in a 700 × 1000 pixel window. The task sequence included 6 practice images to pre-screen for attentiveness, 50 dataset images, and 5 validation trials of faces spaced throughout the sequence. We used Amazon's Mechanical Turk and paid participants at an hourly rate of $10. Data collection cost $4.90 per image for 150 unique gaze points (50 participants each at 0.5, 3, and 5 second viewing durations).

What does CodeCharts measure? We conducted an analysis to understand what aspect of gaze CodeCharts data captures. We had two hypotheses: CodeCharts approximates either (1) people's last fixation before the end of the viewing duration, or (2) all fixations within the viewing duration. For 100 images from the CAT2000 Action category, we used the sequential ground truth eye fixations and assumed that they were equally distributed within the 5-second viewing interval. We then generated two sets of ground truth heatmaps by sampling an equal number of fixations according to our two hypotheses. The similarity between ground truth and CodeCharts heatmaps best supported hypothesis 2 (Pearson's CC of .57 versus .54), which gives us reason to believe that CodeCharts measures what was most salient within the entire viewing interval.

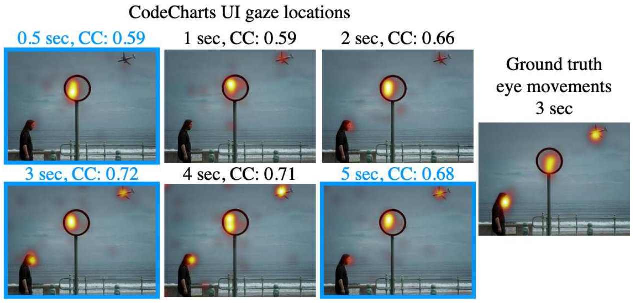


Figure 4: CodeCharts gaze locations collected on an OSIE dataset image at 6 viewing durations (left). The highest correlation with ground truth eye movements (Pearson's CC) occurs when the viewing duration was 3 seconds, the same as the duration used for the eye tracking data collection (right). For further CodeCharts data collection, we used viewing durations of 0.5, 3, and 5 seconds, as they were most distinct from each other.

4. Data analysis

4.1. Is multi-duration saliency predictable?

To measure whether gaze patterns across participants are consistent for a given viewing duration, we perform a split-half consistency analysis on the CodeCharts1K data. We divide participants into two groups, generate a heatmap from each group's gaze points, and compute Pearson's Correlation Coefficient (CC) between the heatmaps. We repeat this computation over 10 splits of participant data and average the scores. To measure whether the gaze patterns vary systematically across durations, we select participants from different viewing duration conditions.

Saliency is predictable across viewing durations: The split-half consistency between participants is high across all

[2] Using 100 "Action" [56] and 20 "Low Resolution" [35] images.

[3] We used action classes that explicitly contained an interaction of a person and an object, by selecting 10 images each of: shooting an arrow, throwing a frisby, walking the dog, writing on a board, writing on a book.

durations (CC=.76 at 0.5 sec, CC=.68 at 3 sec, CC=.67 at 5 sec). While the highest consistency occurs at the briefest duration [8, 51], consistency remains high across the longer viewing durations. These findings hold across the image sets tested (more results in the supplement).

Different things are salient at different durations: When there are differences in what is salient at different durations, CC scores between participants viewing an image at the same duration are higher than CC scores between participants viewing an image at different durations. Gaze patterns are different between .5 and 3 sec for 51% of images from CodeCharts1K; 55% of images show differences between .5 and 5 sec, and 27% of images show differences between 3 and 5 sec.

These analyses indicate that gaze data collected using the CodeCharts UI contains a consistent signal at each of the viewing durations and the signal differs between viewing durations. This suggests that saliency is predictable at different viewing durations, setting the stage for the computational model in Sec. 5.

4.2. What is salient when?

Things and stuff: We used COCO segmentation maps [13] of the SALICON images to compute gaze counts per object class over time. From 0.5 to 3 seconds, gaze frequently moves away from people and towards objects and furniture (e.g., paper, bottle, table). From 3 to 5 seconds, there is an increase in attention on "stuff" (like grass, carpet, and road) that may contain other objects. At these longer durations people gaze more at small and distant objects.

Faces: We know that gaze is attracted by faces [12, 14]. For a finer-grained analysis, we ran a face detection network [26] over images in CodeCharts1K. Across the 266 images where faces were detected, we computed a measure of face saliency at different durations. At each duration, we counted all the gaze points that land on a face region and normalized by the number of gaze points per image across all 3 durations, so face saliency ranges between 0 and 1. Fig. 5a plots face saliency as a function of viewing duration for each image. Across CodeCharts1K, we find a dominant "boomerang" pattern (found in 33% of images with faces): people notice faces at 0.5 sec, their gaze shifts elsewhere at 3 sec, and returns to faces at 5 sec. The second most prevalent pattern is a decrease in gaze on faces over time (24%). Other patterns, like an increase in face saliency over time, were in the minority. These observations are consistent with the phenomenon known as inhibition of return (IOR) [31, 47], or the relative suppression of visual cues that were recently attended to. Samuel and Kat [50] found that IOR lasts for approximately 3 seconds, which might explain why attention tends to shift away from faces between 0.5 and 3 sec but often returns to faces at 5 sec.

Qualitatively, human gaze frequently moves from the actor (at 0.5 sec) to the action (at 3 and 5 sec). Sometimes this shift in attention is gradual: saliency at 3 sec is a combination of saliency at 0.5 and 5 sec (Fig. 5c). In other cases, saliency at 5 sec is more similar to that at 0.5 sec; in these cases it seems that people explore an image before returning to the most interesting regions (Fig. 5b).

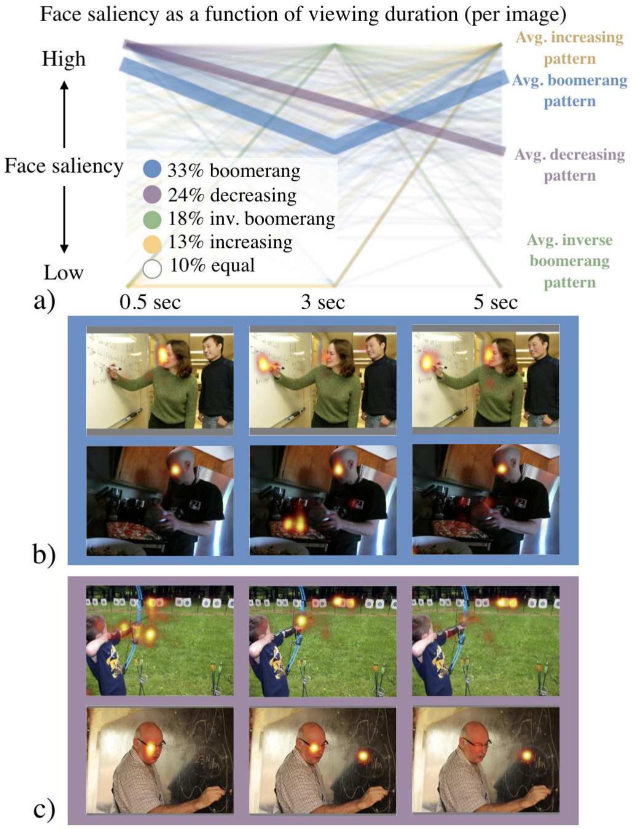


Figure 5: Dominant patterns of human gaze on faces across time. (a) Individual lines plot how the saliency of faces within an image varies across viewing durations. Thicker lines (labeled) are averages over the dominant patterns. We include the percent of images that follow each pattern. (b) Examples where face saliency decreases from 0.5 to 3 sec, increasing again from 3 to 5 sec ("boomerang"). (c) Examples where face saliency decreases from 0.5 to 5 sec.

5. Modeling multi-duration saliency

To efficiently and accurately predict multiple saliency maps for a single image, we introduce the Multi-Duration Saliency Excited Model (MD-SEM), a novel architecture designed for multi-duration saliency (Fig. 6). MD-SEM is the first model that outputs multiple saliency maps corresponding to different viewing durations. The core of our model is a new Temporal Excitation Module (TEM) that applies a time-based re-weighting to saliency feature maps with a minimal increase in parameters. We also design a

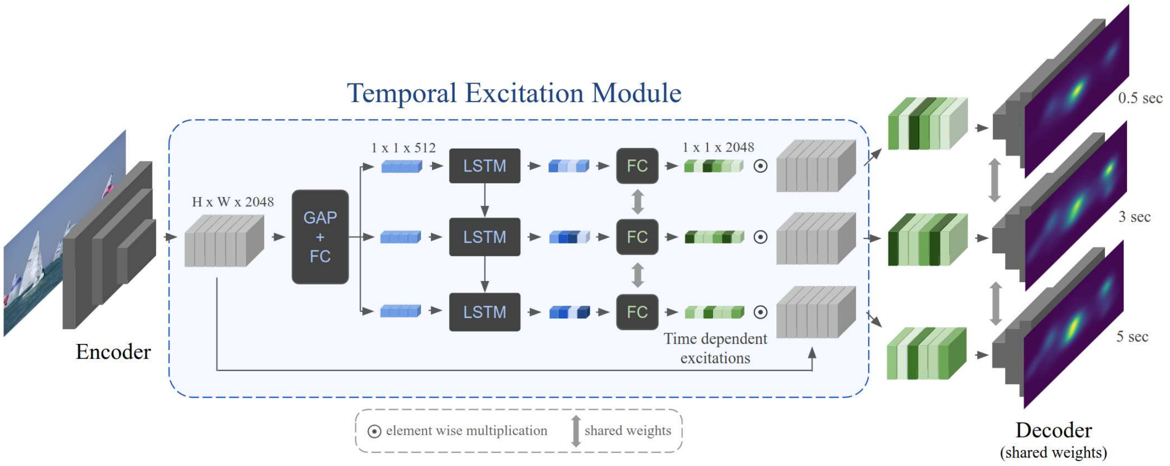


Figure 6: Architecture of our Multi-Duration Saliency Excited Model (MD-SEM). The encoder outputs compressed feature maps that are fed to the Temporal Excitation Module (TEM). In order to predict saliency across durations, TEM uses LSTM cells to generate scaling vectors that re-weight the feature maps differently for each duration. The modified feature maps are then decoded into saliency heatmaps. Reused features and shared weights keep the architecture lightweight.

new loss, the Correlation Coefficient Match (CCM) loss, that encourages the network to capture temporal patterns.

5.1. Architecture motivation

Current saliency architectures tend to be bulky, with large numbers of parameters and specialized modules. For instance, the Saliency Attentive Model (SAM) [22] uses an Attentive Convolutional LSTM and a Learned Prior module with several 5x5 convolutions, bringing the total model size to more than 70M parameters. CASNet [25], which applies channel reweighting over 1024 channels, has 142M parameters. We present an accurate model of reduced size and complexity by distilling the required components to a minimum: (1) a strong Xception-based encoder [20], (2) a temporal processing module that operates on a compressed representation, and (3) a simple regularized decoder.

5.2. Convolutional encoder-decoder

Convolutional encoder-decoder architectures are particularly effective for image-to-image tasks like segmentation [4, 16, 40] and saliency prediction [11, 32, 43]. Encoding the image allows for rich feature extraction and reduces the dimensionality of the input. We use a state-of-the-art backbone as our encoder: the Xception network [20] pretrained on ImageNet. The Xception network is lightweight and accurate (0.790 top-1 accuracy on ImageNet with only 22M parameters) and has shown success in semantic segmentation [17]. For the decoder, our experiments showed that a simple module composed of 3 sets of convolution, up-sampling and dropout layers are sufficient for this task. This choice of module reduces model complexity and implicitly regularizes the network.

5.3. Temporal Excitation Module

To predict multi-duration saliency, we introduce a module that recursively manipulates the feature representation generated by the encoder to adapt it for each duration. Our module uses a Long Short Term Memory (LSTM) network to generate scaling vectors that re-weight the feature maps differently for each of T timesteps (where $T = 3$ in our implementation). Feature map re-weighting has been explored in the form of Squeeze and Excitation Modules [29], but has not been exploited as a temporal modification tool.

The architecture of the Temporal Excitation Module (TEM) is shown in Fig. 6. First, the feature maps generated by the encoder are pooled through global average pooling and passed through a fully connected layer, which reduces the dimensionality of the feature vector and aids in generalization. The output of the dense layer is replicated T times and fed as a sequence to the LSTM. The LSTM then outputs T vectors, which contain information specific to each timestep and will be used to rescale each feature map differently. These vectors are passed through a fully-connected layer that increases their dimensionality to match the channel dimension of the feature maps (C), yielding scaling vectors $s^{(t)}$ of length C. A sigmoid non-linearity ensures that the scaling weights remain within a sensible range. Finally, the block outputs a set of T feature maps, which are obtained by rescaling the original feature maps according to each of the T vectors s. Formally, the module outputs T sets of C feature maps, where each feature map $f_c^{(t)}$ is computed as:

$$f_c^{(t)} = I_c \cdot s_c^{(t)},$$

where I_c is the c-th input feature map and $s_c^{(t)}$ is the scaling weight for duration t and channel c.

Efficiency: TEM is designed to be lightweight. TEM's LSTM operates over a squeezed, low-dimensional vector obtained from pooling input feature maps. By contrast, SAM [22], a top-performing saliency model, uses an LSTM for internal map refinement that operates on full 3D feature maps. Our approach results in an architecture with 30 million parameters, 57% smaller than SAM [22]. In Sec. 6.3, we show that our architecture outperforms SAM.

5.4. Correlation Coefficient Match Loss

To ensure that our network correctly captures differences across viewing durations, we introduce a novel training loss called Correlation Coefficient Match (CCM) loss. This loss forces the network to output saliency maps that reproduce the correlations between ground truth saliency maps at adjacent durations. If ground truth maps at durations t and $t+1$ are dissimilar, we encourage the network to produce equally dissimilar maps at these durations. Given a set of T viewing durations for which we want to predict saliency maps, we calculate the CCM loss by computing Pearson's Correlation Coefficient (CC) on pairs of saliency maps at adjacent durations, then computing the difference between the ground truth and predicted scores. CC is defined as: $CC(y_1, y_2) = \frac{\sigma(y_1,y_2)}{\sigma(y_1)\cdot\sigma(y_2)}$, where $\sigma(y_1, y_2)$ is the covariance of y_1 and y_2. If we let $y^{(t)}$ be the heatmap corresponding to duration t, our CCM loss is:

$$L_{CCM}(y_g, y_p) = \frac{1}{T-1}\sum_{t=0}^{T-1}\left|CC\left(y_g^{(t)}, y_g^{(t+1)}\right) - CC\left(y_p^{(t)}, y_p^{(t+1)}\right)\right|$$

where $y_g^{(t)}$ and $y_p^{(t)}$ are the ground truth and predicted saliency maps for duration t, respectively.

This novel loss boosts performance on multi-duration saliency prediction, increasing the NSS score of MD-SEM by nearly 5% on CodeCharts1K (see the supplement for detailed numbers).

5.5. Implementation details

Architecture: We remove the last fully connected layer from the Xception decoder in order to obtain a feature map of size $H \times W \times 2048$. TEM contains a 512-unit fully connected layer, followed by an LSTM with 512 cells, a ReLU non-linearity, and a sigmoid-activated fully connected layer with 2048 parameters to transform the scaling vector back to its input size. The decoder is composed of 3 sets of convolutional blocks with Dropout. Finally, a 1x1 convolution with 1 filter is used to reduce the final set of feature maps to a single-channel saliency heatmap. Note that the same decoder is applied to each of the T outputs of TEM, thus concentrating time information exclusively in that module and reducing model complexity.

Loss: The network's loss is a weighted combination of our novel CCM loss, Kullback Leibler divergence (KL), Pearson's Correlation Coefficient (CC) and Normalized Scanpath Saliency (NSS) (see [10] for formulations). Since NSS is more robust than other metrics at measuring the quality of saliency predictions [10], we place a higher weight on NSS. We set the weights to 3 for CCM, 10 for KL, -5 for CC and -10 for NSS during SALICON-MD training (Sec. 6.1), but changed the NSS weight to -1 for CodeCharts1K training to account for the reduction in the number of fixations per image.

6. Evaluation

6.1. Datasets

For training, we use the SALICON-MD (Multi-Duration) and CodeCharts1K datasets. We created SALICON-MD from the original SALICON dataset [33] by bucketing a participant's attention locations based on when they occurred. Since no timestamps were provided, we assumed an even distribution across the viewing duration (from 0 to 5 seconds) and split the attention locations into 6 buckets. This time-bucketed data serves as an approximate but large pretraining dataset. For final training and evaluation, we use ground truth multi-duration data from CodeCharts1K (introduced in Sec. 3).

6.2. Training details

Our training scheme takes advantage of both datasets to create a model that is generalizable and accurate at every duration. In order to learn from as much data as possible, we pretrain on SALICON-MD. Pretraining on temporal data that exhibits differences across durations is important so that our model learns at the outset to discriminate between timesteps. We then fine-tune on ground truth CodeCharts1K. For both datasets, we set the batch size to 8 and the initial learning rate to 1e-4, which is reduced by a factor of ten every three epochs. At the beginning of training we freeze the weights of the encoder for one epoch. We found that 10 epochs of training on SALICON-MD and 5 on CodeCharts1K was sufficient. For SALICON-MD, we used the provided test, train, and validation splits. For CodeCharts1K, we trained on 70% of the images, validated on 5%, and tested on 25%.

6.3. Comparison to state-of-the-art

Multi-duration baselines: Our model is first-of-its-kind in its ability to predict saliency at multiple durations. To demonstrate the superiority of our model over existing single-duration models, we compare to a baseline that represents the best alternative for obtaining multiple distinct

		500ms			3000ms			5000ms			All durations		
Model	*Params* ↓	*CC* ↑	*NSS* ↑	*KL* ↓	*CC* ↑	*NSS* ↑	*KL* ↓	*CC* ↑	*NSS* ↑	*KL* ↓	*CC* ↑	*NSS* ↑	*KL* ↓
SAM ×3	210.3M	0.804	3.236	0.366	0.693	2.409	0.545	0.706	2.480	0.537	0.734	2.708	0.483
SAM-MD	70.1M	0.805	3.181	0.370	0.738	2.541	0.469	0.715	2.495	0.535	0.753	2.739	0.458
MD-SEM	**30.9M**	**0.816**	**3.374**	**0.351**	**0.745**	**2.694**	**0.452**	**0.734**	**2.677**	**0.487**	**0.765**	**2.915**	**0.430**

Table 1: Comparison of multi-duration saliency models evaluated on CodeCharts1K. Baselines are SAM×3 (three copies of SAM, each trained exclusively on data for one duration) and SAM-MD (a custom modification of SAM whose LSTM outputs multiple maps). MD-SEM (ours) excels across all three viewing durations, while using substantially fewer parameters.

Model	*NSS* ↑	*CC* ↑	*KL* ↓	*SIM* ↑
SAM-res [22]	1.990	**0.899**	0.610	**0.793**
EML-Net [32]	2.050	0.886	**0.520**	0.780
SalNet [44]	1.859	0.622	-	-
CEDNS	2.045	0.862	1.026	0.753
MD-SEM (Ours)	**2.058**	0.868	0.568	0.774

Table 2: Comparison to state-of-the-art on SALICON test set (LSUN 2017 Challenge).

saliency heatmaps: training multiple copies of a state-of-the-art architecture on the ground truth for 3 different durations. We call this approach *SAM×3*. Next, to demonstrate the advantages of our particular architecture, we benchmark against *SAM-MD*, a modified, multi-duration version of SAM where the LSTM is modified to produce a different saliency map at each timestep. Each output map corresponds to a different viewing duration and the network trains on all three durations simultaneously. The results of these comparisons on all three CodeCharts durations are shown in Table 1. Not only is MD-SEM better at approximating human gaze and differentiating across durations, but it also uses many fewer parameters than the other models. Our model performs particularly well on images from LaMem and CAT2000, but struggles on images with out-of-context objects or complex actions (see the supplement).

Single-duration baselines: We also evaluated our architecture on the conventional single-duration saliency task and obtained a performance competitive with state-of-the-art saliency models. MD-SEM achieves a second-place NSS score on the LSUN 2017 challenge [1] (Table 2).

Qualitatively, our model accurately reproduces many of the dominant human gaze patterns from the CodeCharts1K dataset, such as the tendency of humans to focus on the object of an action at longer viewing durations, and for attention to shift from the center of the image to smaller details and secondary objects (Fig. 10).

7. Applications

Saliency models have proven useful for many image processing applications, including smart cropping, retargeting, and image captioning. Our multi-duration saliency model can contribute additional context by accounting for the expected time that a viewer may have to explore an image. In this section, we discuss how multi-duration saliency can be used to adapt existing saliency-driven applications.

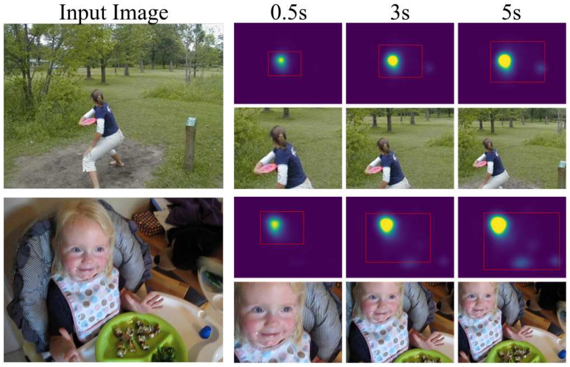


Figure 7: **Cropping.** Images automatically cropped based on cumulative viewing duration by selecting the window with 90% of the most salient image regions as predicted by our model. Image crops for shorter viewing durations contain close-ups of key elements.

Cropping: Automatic image cropping is useful for thumbnailing, view-finding for improved composition, and retargeting for different use cases [23]. Multi-duration saliency allows us to additionally take into account the expected time a viewer will spend on an image (e.g., an image that is part of a passing advertisement should contain fewer elements than if it is the main image on a page). In Fig. 7 we use our multi-duration saliency maps to crop windows that capture 90% of the heatmap density that occurs at or below a particular viewing duration [18]. Our automatically-generated thumbnails contain close-ups of the most important objects at shorter viewing durations.

Compression and rendering: Multi-duration saliency heatmaps can indicate the order in which items in a scene should be rendered to provide a seamless user experience. In Fig. 8 we visualize which elements would be prioritized at different viewing durations. To generate these visualizations, we used Mask R-CNN for instance segmentation [28]. We accumulated saliency heatmap density for each instance across time to determine which instances to pri-

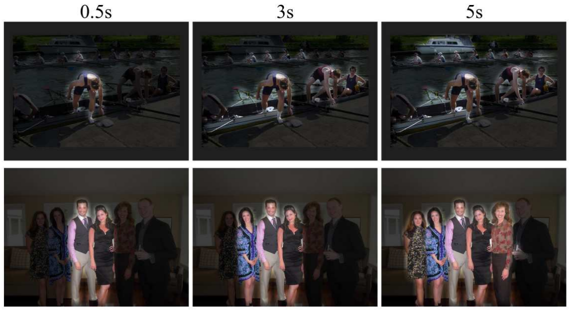


Figure 8: **Compression.** We visualize instance detections that are predicted to attract gaze at different viewing durations (accumulated over time). Content that is salient at short durations could be rendered before content that becomes salient later.

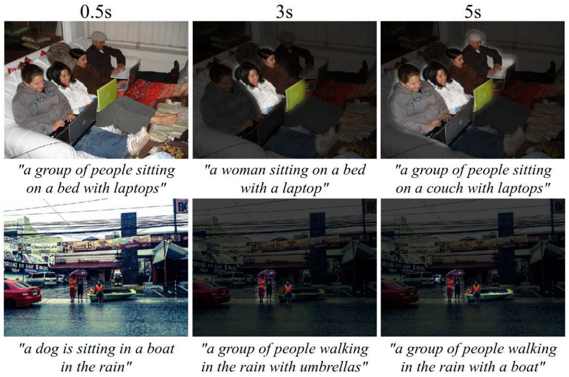

0.5s
3s
5s
"a group of people sitting on a bed with laptops"
"a woman sitting on a bed with a laptop"
"a group of people sitting on a couch with laptops"
"a dog is sitting in a boat in the rain"
"a group of people walking in the rain with umbrellas"
"a group of people walking in the rain with a boat"

Figure 9: **Captioning.** Captions generated by passing saliency-enhanced images to an image captioning model [48], using saliency at different viewing durations to prioritize image content.

oritize. Instances with a mean saliency score in the 90th percentile were kept and the rest of the image was blurred and darkened for visualization purposes.

Captioning: Image captions can facilitate search and improve accessibility. Some recent work attempts to use a saliency map to guide attention for captioning [21]; however the saliency model does not explicitly model the temporal aspect of human attention. In Fig. 9, we used our saliency predictions to focus an image captioning model [48] on regions that should stand out at different viewing durations. Removing the non-salient visual clutter can benefit caption quality.

8. Conclusion

Guided by the insight that where you focus on an image depends on how much time you have to explore it, we tackle the problem of predicting multi-duration saliency: saliency as a function of viewing duration. We propose a scalable, crowdsourceable technique for gathering ground truth multi-duration saliency data and use it to collect the CodeCharts1K dataset. Our LSTM-based saliency model is a top performer at predicting conventional saliency while also providing predictions at multiple durations. We provided initial hints of how multi-duration saliency could be used in applications which require prioritizing visual content, but unfortunately, *your viewing time is up*.

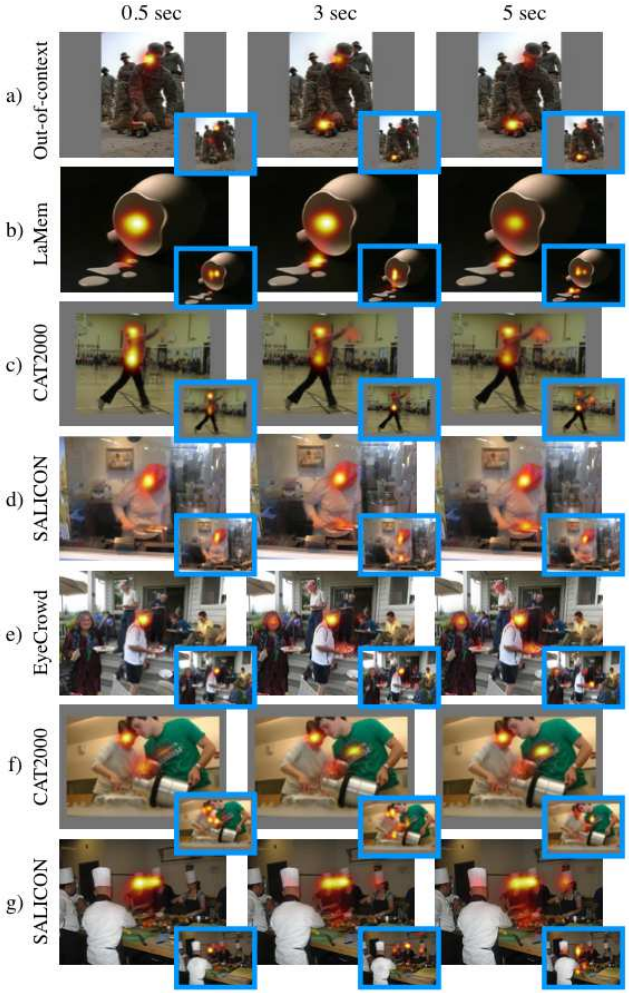


Figure 10: MD-SEM predictions on various datasets. Insets contain human ground truth from CodeCharts1K. Our model approximates human attention by shifting saliency from faces to objects of action across time (a,c,d) and shifting the center of focus from the center of the image to secondary image regions at longer viewer durations (b,e). Difficult cases for our model include cluttered scenes with many objects, people, or complex actions (f,g).

Acknowledgements: We thank our funding sources: the Vannevar Bush Faculty Fellowship program by the ONR (N00014-16-1-3116) and the SystemsThatLearn@CSAIL / Ignite Grant (to A.O). Work supported in part by cloud credits from the MIT Quest for Intelligence.

References

[1] SALICON Saliency Prediction Challenge (LSUN 2017). https://competitions.codalab.org/competitions/17136#results.

[2] Marc Assens, Xavier Giró i Nieto, Kevin McGuinness, and Noel E. O'Connor. Pathgan: Visual scanpath prediction with generative adversarial networks. *CoRR*, abs/1809.00567, 2018.

[3] Marc Assens Reina, Xavier Giró-i Nieto, Kevin McGuinness, and Noel E. O'Connor. Saltinet: Scan-path prediction on 360 degree images using saliency volumes. In *ICCV Workshop on Egocentric Perception, Interaction and Computing*, Oct 2017.

[4] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[5] Wilma A Bainbridge, Phillip Isola, and Aude Oliva. The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General*, 142(4):1323, 2013.

[6] Ali Borji. Saliency prediction in the deep learning era: An empirical investigation. *arXiv preprint arXiv:1810.03716*, 2018.

[7] Ali Borji and Laurent Itti. CAT2000: A large scale fixation dataset for boosting saliency research. *CVPR'15 Workshop on the Future of Datasets*, 2015.

[8] Guy Thomas Buswell. How people look at pictures: a study of the psychology and perception in art. 1935.

[9] Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba. Mit saliency benchmark.

[10] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *IEEE transactions on pattern analysis and machine intelligence*, 41(3):740–757, 2019.

[11] Zoya Bylinskii, Nam Wook Kim, Peter O'Donovan, Sami Alsheikh, Spandan Madan, Hanspeter Pfister, Fredo Durand, Bryan Russell, and Aaron Hertzmann. Learning visual importance for graphic designs and data visualizations. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*, pages 57–69. ACM, 2017.

[12] Zoya Bylinskii, Adrià Recasens, Ali Borji, Aude Oliva, Antonio Torralba, and Frédo Durand. Where should saliency models look next? In *European Conference on Computer Vision*, pages 809–824. Springer, 2016.

[13] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Computer vision and pattern recognition (CVPR), 2018 IEEE conference on*. IEEE, 2018.

[14] Moran Cerf, E Paxon Frady, and Christof Koch. Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of vision*, 9(12):10–10, 2009.

[15] Jiansheng Chen, Gaocheng Bai, Shaoheng Liang, and Zhengqin Li. Automatic image cropping: A computational complexity study. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 507–515, 2016.

[16] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2018.

[17] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 801–818, 2018.

[18] Yi-Ling Chen, Tzu-Wei Huang, Kai-Han Chang, Yu-Chen Tsai, Hwann-Tzong Chen, and Bing-Yu Chen. Quantitative analysis of automatic image cropping algorithms:a dataset and comparative study. In *IEEE WACV 2017*, 2017.

[19] Myung Jin Choi, Antonio Torralba, and Alan S Willsky. Context models and out-of-context objects. *Pattern Recognition Letters*, 33(7):853–862, 2012.

[20] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

[21] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. Paying more attention to saliency: Image captioning with saliency and context attention. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(2):48, 2018.

[22] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. Predicting human eye fixations via an lstm-based saliency attentive model. *IEEE Transactions on Image Processing*, 27(10):5142–5154, 2018.

[23] Rachel England. Twitter uses smart cropping to make image previews more interesting. https//engadget.com/2018/01/25/twitter-uses-smart-cropping-to-make-image-previews-m

[24] Seyed A Esmaeili, Bharat Singh, and Larry S Davis. Fast-at: Fast automatic thumbnail generation using deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4622–4630, 2017.

[25] Shaojing Fan, Zhiqi Shen, Ming Jiang, Bryan Koenig, Juan Xu, Mohan Kankanhalli, and Qi Zhao. Emotional attention: A study of image sentiment and visual attention. pages 7521–7531, 06 2018.

[26] Adam Geitgey. Face recognition. http://github.com/ageitgey/face_recognition.

[27] Rui Han and Shuangjiu Xiao. Human visual scanpath prediction based on rgb-d saliency. In *Proceedings of the 2018 International Conference on Image and Graphics Processing*, ICIGP 2018, pages 180–184, New York, NY, USA, 2018. ACM.

[28] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[29] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[30] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. Salicon: Reducing the semantic gap in saliency prediction by

adapting deep neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 262–270, 2015.

[31] Laurent Itti and Christof Koch. Computational modeling of visual attention. *Nature reviews. Neuroscience*, 2:194–203, 04 2001.

[32] Sen Jia. Eml-net: An expandable multi-layer network for saliency prediction. *arXiv preprint arXiv:1805.01047*, 2018.

[33] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1072–1080, 2015.

[34] Ming Jiang, Juan Xu, and Qi Zhao. Saliency in crowd. In *European Conference on Computer Vision*, pages 17–32. Springer, 2014.

[35] Tilke Judd, Fredo Durand, and Antonio Torralba. Fixations on low-resolution images. *Journal of Vision*, 11(4):14–14, 2011.

[36] Aditya Khosla, Akhil S. Raju, Antonio Torralba, and Aude Oliva. Understanding and predicting image memorability at a large scale. In *International Conference on Computer Vision (ICCV)*, 2015.

[37] Nam Wook Kim, Zoya Bylinskii, Michelle A Borkin, Krzysztof Z Gajos, Aude Oliva, Fredo Durand, and Hanspeter Pfister. Bubbleview: an interface for crowdsourcing image importance maps and tracking visual attention. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 24(5):36, 2017.

[38] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. Eye tracking for everyone. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2176–2184, 2016.

[39] Olivier Le Meur and Thierry Baccino. Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior research methods*, 45(1):251–266, 2013.

[40] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[41] Anelise Newman, Barry McNamara, Camilo Fosco, Yun Bin Zhang, Pat Sukhum, Matthew Tancik, Nam Wook Kim, and Zoya Bylinskii. TurkEyes: A web-based toolbox for crowdsourcing attention data. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020.

[42] Thuyen Ngo and B.S. Manjunath. Saccade gaze prediction using a recurrent neural network. In *IEEE International Conference on Image Processing*, Sep 2017.

[43] Junting Pan, Cristian Canton, Kevin McGuinness, Noel E. O'Connor, Jordi Torres, Elisa Sayrol, and Xavier and Giro-i Nieto. Salgan: Visual saliency prediction with generative adversarial networks. In *arXiv*, January 2017.

[44] Junting Pan, Elisa Sayrol, Xavier Giro-i Nieto, Kevin McGuinness, and Noel E O'Connor. Shallow and deep convolutional networks for saliency prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 598–606, 2016.

[45] Alexandra Papoutsaki, Patsorn Sangkloy, James Laskey, Nediyana Daskalova, Jeff Huang, and James Hays. Webgazer: Scalable webcam eye tracking using user interactions. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence-IJCAI 2016*, 2016.

[46] Sangdon Park, Wonsik Kim, and Kyoung Mu Lee. Abnormal object detection by canonical scene-based contextual model. In *European Conference on Computer Vision (ECCV)*, 2012.

[47] Michael Posner and Yoav Cohen. Components of visual orienting. *Attention and performance X: Control of language processes*, 32:531–, 01 1984.

[48] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7008–7024, 2017.

[49] Dmitry Rudoy, Dan B Goldman, Eli Shechtman, and Lihi Zelnik-Manor. Crowdsourcing gaze data collection. *Proceedings of ACM Collective Intelligence Conference*, 2012.

[50] Arthur Samuel and Donna Kat. Inhibition of return: A graphical meta-analysis of its time course and an empirical test of its temporal and spatial properties. *Psychonomic bulletin review*, 10:897–906, 01 2004.

[51] Benjamin W Tatler, Mary M Hayhoe, Michael F Land, and Dana H Ballard. Eye guidance in natural vision: Reinterpreting salience. *Journal of vision*, 11(5):5–5, 2011.

[52] Hamed R Tavakoli, Fawad Ahmed, Ali Borji, and Jorma Laaksonen. Saliency revisited: Analysis of mouse movements versus fixations. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 1774–1782, 2017.

[53] Yixiu Wang, Bin Wang, Xiaofeng Wu, and Liming Zhang. Scanpath estimation based on foveated image saliency. *Cognitive Processing*, 18, 10 2016.

[54] Zijun Wei, Jianming Zhang, Xiaohui Shen, Zhe Lin, Radomír Mech, Minh Hoai, and Dimitris Samaras. Good view hunting: learning photo composition from dense view pairs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5437–5446, 2018.

[55] Juan Xu, Ming Jiang, Shuo Wang, Mohan S Kankanhalli, and Qi Zhao. Predicting human gaze beyond pixels. *Journal of vision*, 14(1):28–28, 2014.

[56] Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas Guibas, and Li Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *2011 International Conference on Computer Vision*, pages 1331–1338. IEEE, 2011.